Digital Libraries, Intelligent Data Analytics, and Augmented Description:

A Demonstration Project

Final Report

Prepared by Elizabeth Lorang, Leen-Kiat Soh, Yi Liu, and Chulwoo Pack

University Libraries & Department of Computer Science and Engineering

University of Nebraska–Lincoln

Submitted to the Library of Congress

January 10, 2020
Rev. March 26, 2020[1]

[1] This revision includes updated information about the code repository in section 4 (Code & Data), a new footnote about Beyond Words on page 10, and two additional tables (Table 9 and Table 10), and it corrects several typos.

SUMMARY

From July 16-to November 8, 2019, the Aida digital libraries research team at the University of Nebraska-Lincoln collaborated with the Library of Congress on "Digital Libraries, Intelligent Data Analytics, and Augmented Description: A Demonstration Project." This demonstration project sought to (1) develop and investigate the viability and feasibility of textual and image-based data analytics approaches to support and facilitate discovery; (2) understand technical tools and requirements for the Library of Congress to improve access and discovery of its digital collections; and (3) enable the Library of Congress to plan for future possibilities. In pursuit of these goals, we focused our work around two areas: extracting and foregrounding visual content from Chronicling America (chroniclingamerica.loc.gov) and applying a series of image processing and machine learning methods to minimally processed manuscript collections featured in By the People (crowd.loc.gov). We undertook a series of explorations and investigated a range of issues and challenges related to machine learning and the Library's collections.

This final report details the explorations, addresses social and technical challenges with regard to the explorations and that are critical context for the development of machine learning in the cultural heritage sector, and makes several recommendations to the Library of Congress as it plans for future possibilities. We propose two top-level recommendations. First, the Library should focus the weight of its machine learning efforts and energies on social and technical infrastructures for the development of machine learning in cultural heritage organizations, research libraries, and digital libraries. Second, we recommend that the Library invest in continued, ongoing, intentional explorations and investigations of particular machine learning applications to its collections. Both of these top-level recommendations map to the three goals of the Library's 2019 digital strategy.

Within each top-level recommendation, we offer three more concrete, short- and medium-term recommendations. They include, under social and technical infrastructures: (1) Develop a statement of values or principles that will guide how the Library of Congress pursues the use, application, and development of machine learning for cultural heritage. (2) Create and scope a machine learning roadmap for the Library that looks both internally to the Library of Congress and its needs and goals and externally to the larger cultural heritage and other research communities. (3) Focus efforts on developing ground truth sets and benchmarking data and making these easily available. Nested under the recommendation to support ongoing explorations and investigations, we recommend that the Library: (4) Join the Library of Congress's emergent efforts in machine learning with its existing expertise and leadership in crowdsourcing. Combine these areas as "informed crowdsourcing" as appropriate. (5) Sponsor challenges for teams to create additional metadata for digital collections in the Library of Congress. As part of these challenges, require teams to engage across a range of social and technical questions and problem areas. (6) Continue to create and support opportunities for researchers to partner in substantive ways with the Library of Congress on machine learning explorations. Each of these recommendations speak to the investigation and challenge areas identified by Thomas Padilla in *Responsible Operations: Data Science, Machine Learning, and AI in Libraries.*

This demonstration project—via its explorations, discussion, and recommendations—shows the potential of machine learning toward a variety of goals and use cases, and it argues that the technology itself will not be the hardest part of this work. The hardest part will be the myriad challenges to undertaking this work in ways that are socially and culturally responsible, while also upholding responsibility to make the Library of Congress's materials available in timely and accessible ways. Fortunately, the Library of Congress is in a remarkable position to advance machine learning for cultural heritage organizations, through its size, the diversity of its collections, and its commitment to digital strategy.

TABLE OF CONTENTS

Tables, Figures & Charts…… i

1 Introduction…… 1

2 Participants & Roles…… 1

3 Timeline…… 1

4 Code & Data…… 2

5 Demonstration Project Approach & Design…… 2

6 The Explorations…… 3
- 6.1 Exploration: Document Segmentation…… 5
- 6.2 Exploration: Graphic Element Classification and Text Extraction…… 9
- 6.3 Exploration: Document Type Classification…… 14
- 6.4 Exploration: Digitization Type Differentiation…… 16
- 6.5 Exploration: Document Image Quality Assessment (DIQA) and Advanced DIQA…… 19
- 6.6 Exploration: Document Clustering…… 25

7 Discussion…… 27
- 7.1 Social…… 27
- 7.2 Technical…… 28
- 7.3 Social-Technical…… 30

8 Recommendations…… 30
- 8.1 Social and Technical Infrastructures…… 31
  - 8.1.1 Develop a statement of values or principles 32
  - 8.1.2 Create and scope a machine learning roadmap…… 32
  - 8.1.3 Focus efforts on developing ground truth…… 33
- 8.2 Explorations and Investigations…… 34
  - 8.2.1 Join efforts…… 35
  - 8.2.2 Sponsor challenges…… 35
  - 8.2.3 Continue to support research partnerships. 36

9 Conclusion…… 39

Bibliography…… 40

Appendix 1: Work-in-Progress Reports …… 42

Appendix 2: Work-in-Progress Presentations…… 119

TABLES, FIGURES & CHARTS

TABLES

Table 1. The explorations pursued as part of the demonstration project and their selected potential applications. 4

Table 2. Results of page segmentation when training and evaluating the dhSegment model on two sets of historical newspapers. 6

Table 3. Average performance of pre-training investigation and fine-tuning approaches. 12

Table 4. Precision, recall, and f1-score of VGG-16 as trained on RVL_CDIP dataset. 14

Table 5. Configuration of suffrage_1002 dataset..... 15

Table 6. Precision, recall, and f1-scores of VGG-16 on suffrage_1002 testing set. 15

Table 7. Breakdown of projects and actual classifications of content as digitized from microform or digitized from an original item. 17

Table 8. Comparison of accuracy of compactness of two algorithms. 24

Table 9. Infrastructure and application recommendations mapped to elements of the Library of Congress's digital strategy. 37

Table 10. Recommendations Mapped to areas of investigation and challenge areas outlined in Padilla's Responsible Operations. 38

FIGURES

Figure 1. Visual representation of the explorations and their relationships to one another. 3

Figure 2. Segmentation result of ENP_500_v4 on a Chronicling America image (sn92053240-19190805.jpg). . 6

Figure 3. Segmentation result of ENP_500_v4 on a Chronicling America image (sn84026820_00271765095_1917050501_0153.jpg). 7

Figure 4. Segmentation result of ENP_500_v4 on a Chronicling America image (sn82014086_00295866135_1917091301_0116.jpg). 8

Figure 5. Segmentation result of ENP_500_v4 on Chronicling America image (sn86063952-19190805.jpg). 9

Figure 6. The Beyond Words data, which we treated as ground truth 12

Figure 7. Each of the three columns with graphical content have much larger 13

Figure 8. In this case, the "Buy Liberty Bonds" advertisement is not represented in the ground truth 13

Figure 9. Prediction failure cases. 15

Figure 10. Facing pages of a document, digitized from the original 17

Figure 11. A handwritten manuscript page image, digitized from the original 18

Figure 12. A digital image of a coin, digitized from the original item 18

Figure 13. A digital image of a photograph, digitized from the original item 19

Figure 14. An image with a low-contrast score from the Civil War years. 22

Figure 15. Images from three different clusters. 26

Figure 16. Images from three different clusters following intensity value normalization. 27

CHARTS

Chart 1. Skewness measures of 35,900 images from minimally processed Civil War collections. 20

Chart 2. Average contrast scores of materials within decade-ranges. 21

Chart 3. Average contrast scores of materials from the decade 1860-1869 by year. 21

Chart 4. Average range effect of the Civil War collection over time. 22

Chart 5. Average bleed-through/noise in materials from the Civil War collection, by decade. 23

Chart 6. The compactness of the Europeana Newspapers dataset. 24

1 Introduction

In response to notice ID 030ADV19Q0274, "The Library of Congress – Pre-processing Pilot," the Aida digital libraries research team at the University of Nebraska-Lincoln (UNL) proposed "Digital Libraries, Intelligent Data Analytics, and Augmented Description: A Demonstration Project." The proposal was awarded a research services contract from the Library of Congress. From July 16-to November 8, 2019, members of the Aida research team conducted a series of explorations and analyses to assist the Library of Congress in assessing possible applications of machine learning within the Library. Three broad goals framed this work: (1) develop and investigate the viability and feasibility of textual and image-based data analytics approaches to support and facilitate discovery; (2) understand technical tools and requirements for the Library of Congress to improve access and discovery of its digital collections; and (3) enable the Library of Congress to plan for future possibilities.

This report summarizes the design of the demonstration project and the activities of the Aida research team; presents key findings resulting from these activities; makes recommendations for possible paths forward; and provides documentation for the major activities, including code, data, and reports-in-progress completed during the project. This report serves two main purposes: to document the work completed and to extrapolate from that work to broader implications for machine learning endeavors at the Library of Congress.

2 Participants & Roles

University of Nebraska-Lincoln

- Elizabeth Lorang, senior adviser
- Leen-Kiat Soh, senior adviser
- Yi Liu, research associate and developer
- Chulwoo (Mike) Pack, research associate and developer
- Ashlyn Stewart, research assistant

Library of Congress[†]

- Meghan Ferriter, Chief (Acting) LC Labs/Senior Innovation Specialist
- Abbey Potter, Senior Innovation Specialist
- Jaime Mears, Senior Innovation Specialist
- Eileen Jakeway, Innovation Specialist
- Tong Wang, Senior IT Specialist, OCIO
- Lauren Algee, Senior Innovation Specialist
- Victoria Van Hyning, Senior Innovation Specialist

3 Timeline

July 16, 2019

Project kick-off meeting held at the Library of Congress

July 19-August 23, 2019

First-round of iterative development, onsite at the Library of Congress

[†] In addition to these key contributors, many others at the Library of Congress supported this demonstration project in a variety of ways, including through their hospitality, encouragement, brainstorming, and interest in this project. We are grateful for and indebted to their efforts.

August 26-November 8, 2019

Second round of iterative development, offsite at the University of Nebraska-Lincoln

November 6, 2019

Delivery of preliminary results via virtual meeting

November 7 – January 9

Development of open repository of code, data, and documentation; development of final report

January 10, 2020

Delivery of final results via in-person meeting at Library of Congress

4 Code & Data

Code and descriptions of data are available via the Library of Congress's GitHub organization page at the "Exploring ML with Project Aida" repository, https://github.com/LibraryOfCongress/Exploring-ML-with-Project-Aida. Following submission to the Library of Congress, code, data, and this report will also be available via https://projectaida.org.

5 Demonstration Project Design & Approach

With the size of the Library of Congress's digital collections and the many potential areas of impact, we might have pursued any number of questions in this demonstration project. Scoping our work, both with regard to the questions we pursued and the number and type of explorations, was critical. We anchored our work around two areas: (1) extracting and foregrounding visual content from Chronicling America (chroniclingamerica.loc.gov) through a variety of techniques and approaches and (2) applying a series of image processing and machine learning methods and techniques to minimally processed manuscript collections featured in By the People (crowd.loc.gov). We identified these areas of focus because they drew on collections already deemed significant by the Library of Congress and because they had a degree of ground-truthing work already completed. In addition, they offered the opportunity to explore the advantages and disadvantages and the strengths and weaknesses of computational/machine learning approaches as compared to data and information generated by experts, casual users, and researchers. Working with these collections had the further benefit of significant opportunity to create new, rich, and varied metadata about them, so that the Library might explore the ways in which more robust metadata might allow for alternative points of entry into the materials and the opportunity for Library staff and researchers to pursue questions of varying nature.

Ultimately, we designed a series of explorations that allowed us to investigate a range of issues and challenges related to machine learning and the Library's collections. The explorations were developed through an iterative process and in regular consultation with members of the Library of Congress staff, both to learn from their expertise and to make sure the questions we were pursuing were of value and interest to the Library. Through that process, some explorations merged, others concluded more quickly than others, and areas of inquiry seeded in one exploration began to sprout in others as well. Individually, the explorations pursued particular technical and collections-oriented questions. We also used the explorations as points of entry into—and paths to reflection about—larger issues, questions, and challenges for machine learning and cultural heritage.

6 THE EXPLORATIONS

This section presents an overview and details of six explorations: Document Segmentation; Graphic Element Classification and Text Extraction; Document Type Classification; Digitization Type Differentiation; Document Image Quality Assessment and Advanced Document Image Quality Assessment; and Document Clustering. Figure 1 and Table 1 identify and show relationships among the explorations and summarize them. In our look at each exploration, we identify guiding questions; outline and describe our approaches, techniques, and methods; present high-level results and analysis; and offer ideas toward future development and/or potential applications.

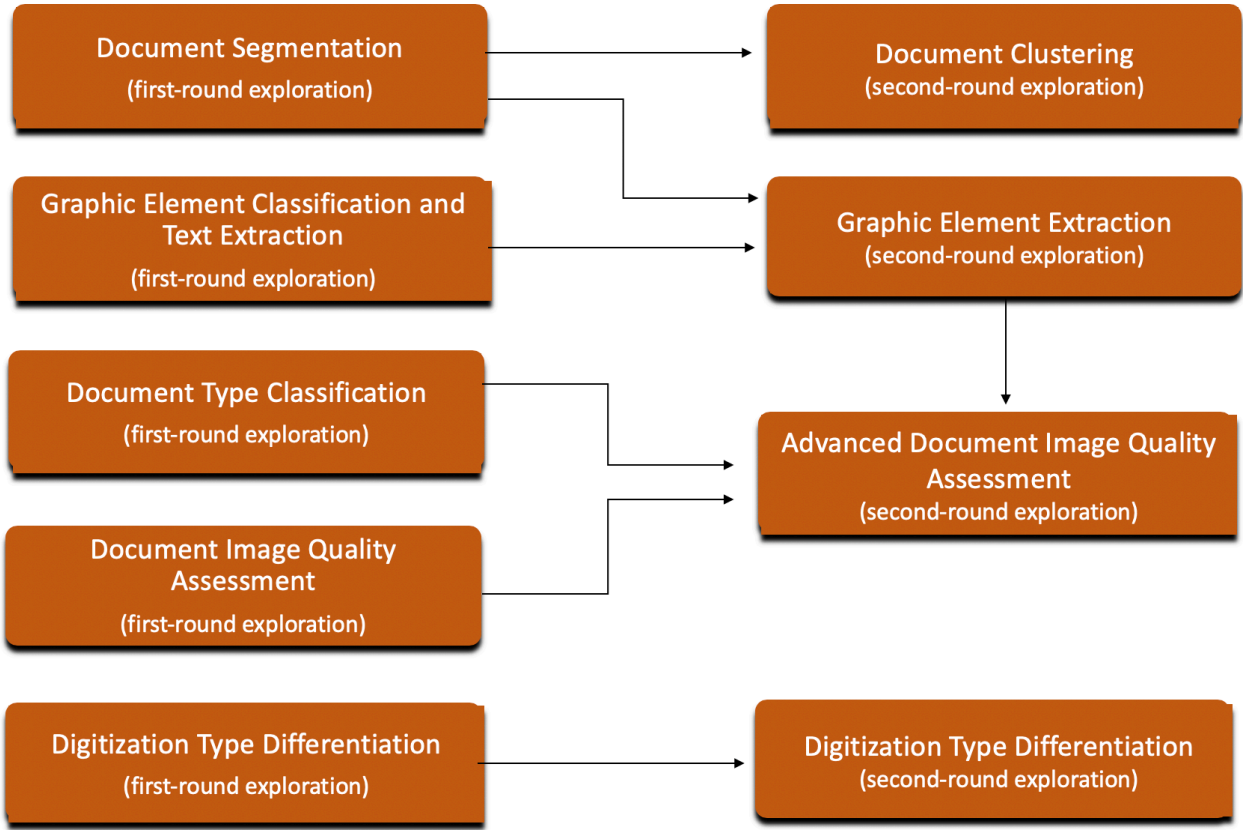


FIGURE 1. VISUAL REPRESENTATION OF THE EXPLORATIONS AND THEIR RELATIONSHIPS TO ONE ANOTHER.

TABLE 1. THE EXPLORATIONS PURSUED AS PART OF THE DEMONSTRATION PROJECT AND THEIR SELECTED POTENTIAL APPLICATIONS.

		Selected Potential Applications					
	Technical Objective	**Metadata generation (structural, descriptive, etc.)**	**Graphical content extraction**	**Influence decision-making for human and/or machine processing**	**Faceted data for end-users/researchers in search/discovery interface**	**Ground truth and benchmark sets for machine learning and image analysis projects competitions**	**Understanding collections**
First-Round Explorations							
Document Segmentation	Find and localize image-like components in newspaper pages	✓	✓		✓	✓	
Graphic Element Classification and Text Extraction	Find and localize graphical content, extract text from this content in newspapers	✓	✓		✓	✓	
Document Type Classification	Classify manuscript collection pages as handwritten, printed, mixed	✓		✓	✓	✓	✓
Document Image Quality Assessment	Analyze quality of manuscript collection page images	✓		✓	✓	✓	✓
Digitization Type Differentiation	Classify manuscript collection images as digitized from original or microform	✓		✓	✓	✓	✓
Second-Round Explorations							
Document Clustering	Extract high-level features, cluster, investigate similarity	✓		✓	✓	✓	✓
Figure/Graph Extraction	Find and localize image-like components in newspaper pages	✓	✓		✓	✓	
Advanced Document Image Quality Assessment	Analyze quality of manuscript collection page images for compactness	✓		✓	✓	✓	✓
Digitization Type Differentiation	Fine-tune classification of manuscript collection images as digitized from original or microform	✓		✓	✓	✓	✓

6.1 Exploration: Document Segmentation

The goal of this exploration was to see if we could localize textual zones, figures, layout borders, and tables and then identify image-like components in historic newspaper pages. Currently, newspaper page images presented through Chronicling America are not zoned or segmented below the page level. In addition, content within a newspaper page is not identified or classified by genre, type, or other features. This exploration, then, was guided by the questions: how might we use image zoning and segmentation to generate additional information about newspaper pages in the Chronicling America corpus? Could image zoning and segmentation be used to pull out graphical content from Chronicling America newspapers? How might machine learning projects draw on ground truth or benchmark data already generated through crowdsourcing efforts?

This exploration applied the dhSegment tool for historical document image processing to historical newspapers.[3] ResNet, a feature extractor in dhSegment, is capable of encoding an image down to a set of high-level visual features effectively and efficiently. We applied dhSegment to two sets of newspaper images, one set from the Library of Congress's Beyond Words project (http://beyondwords.labs.loc.gov/#/), which is based on Chronicling America newspapers, and one from the Europeana Newspapers Project (https://www.primaresearch.org/datasets/ENP).

In the Beyond Words project, members of the public drew rectangular zones around illustrations, photographs, comics, and cartoons in World War I-era newspapers, and users also transcribed captions for this content. In a subsequent stage of work, users could apply a typology to the graphical content, choosing among editorial cartoon, comics/cartoon, illustration, photograph, or map. Our expectation was that Beyond Words would provide ground truth data against which we might verify machine learning-based approaches to the same challenges (graphical content zoning).

We obtained a subset of 1,532 newspaper page images from Beyond Words and corresponding data for graphical-content zones. We used 1,226 images for training and 306 images for evaluation. In two different test scenarios (BW_1500_v1 and BW_1500_v2), when trained and evaluated on Beyond Words newspaper pages, we achieved best accuracy scores of 87% and 88%. Table 2 outlines results. Unfortunately, the relatively high accuracy scores are misleading upon further examination, since the model's behavior of predicting most pixels to be background pixels is guaranteed to achieve high accuracy. The low values for the best mean intersection over union (mIoU) scores verify this problematic behavior in the model, as we observe only a 26% and 24% overlap between the target class and the model's prediction.

For further exploration of the approach, we also trained and evaluated the model on a set of 481 pages from the Europeana Newspapers corpus. These newspaper page images are already zoned and segmented, with the segments classed as background, text, figure, separator or table. These classes are different from the classes in the Beyond Words dataset, which were all classes of graphical content or background. When we trained and evaluated the model on the Europeana Newspapers, we were verifying against the respective classes of the set.

In deploying on the Europeana Newspapers dataset in four scenarios (ENP_500_v1, ENP_500_v2, ENP_500_v3, ENP_500_v4), we achieved best accuracy scores of 88%, 89%, 91%, and 91%, and best mIoU scores of 64%, 64%, 69%, and 69%. In these scenarios, text regions are included in the ground-truth, and thus the model's simple guessing that everything is background is penalized. The high accuracy scores are more trustworthy in these scenarios, as further corroborated by the higher scores for mIoU.

[3] Seguin and Ares Oliveira, dhSegment; Ares Oliveira, Seguin, and Kaplan, "DhSegment."

TABLE 2. RESULTS OF PAGE SEGMENTATION WHEN TRAINING AND EVALUATING THE DHSEGMENT MODEL ON TWO SETS OF HISTORICAL NEWSPAPERS, A ROUGHLY 1500-PAGE SET FROM BEYOND WORDS/CHRONICLING AMERICA AND AN APPROXIMATELY 500-PAGE SET FROM THE EUROPEANA NEWSPAPERS COLLECTION.

Model	Train/Eval Size	Classes	Weighted Training	Pre-processing (Normalization)	Best Score	
					Accuracy	mIoU
BW_1500_v1	1226/306	0: Background 1: Editorial cartoon 2: Comics/cartoon 3: Illustration 4: Photograph 5: Map	No	No	0.87	0.24
BW_1500_v2			Yes [10; 22; 20; 18; 8; 22]		0.88	0.26
ENP_500_v1	385/96	0: Background 1: Text 2: Figure 3: Separator 4: Table	Yes [5; 10; 40; 10; 35]	No	0.88	0.64
ENP_500_v2				Yes	0.89	0.64
ENP_500_v3			No	No	0.91	0.69
ENP_500_v4				Yes	0.91	0.69

We did not conduct a broad deployment of the Europeana Newspapers model on Beyond Words/Chronicling America pages, because we did not have verifiable, commensurate ground truth across the sets. We did, however, conduct a limited test of the Europeana Newspapers-trained model (specifically ENP_500_v3) on Chronicling America page images, and the visual results are encouraging. See Figure 2, Figure 3, Figure 4, and Figure 5 for true-positive correlations, according to visual inspection, as well as for examples of false-positives and false-negatives, respectively.

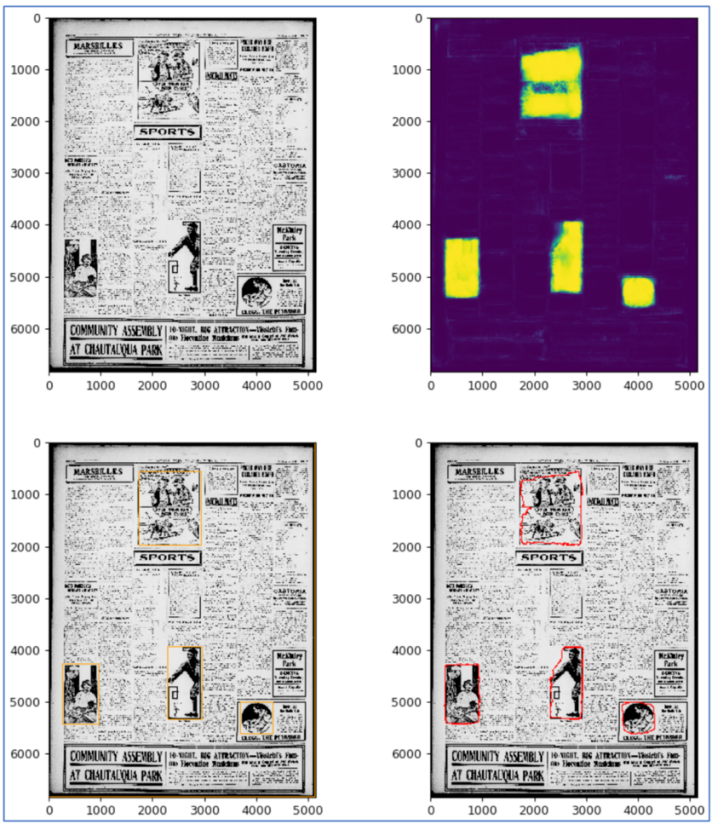


FIGURE 2. SEGMENTATION RESULT OF ENP_500_V4 ON A CHRONICLING AMERICA IMAGE (SN92053240-19190805.JPG). CLOCKWISE FROM TOP- LEFT: (1) INPUT, (2) PROBABILITY MAP FOR FIGURE CLASS, (3) DETECTED FIGURES IN POLYGON, AND (4) DETECTED FIGURES IN BOUNDING-BOX. IN THE PROBABILITY MAP, PIXELS WITH A HIGHER PROBABILITY OF BELONGING TO THE FIGURE CLASS ARE SHOWN WITH A BRIGHTER COLOR.

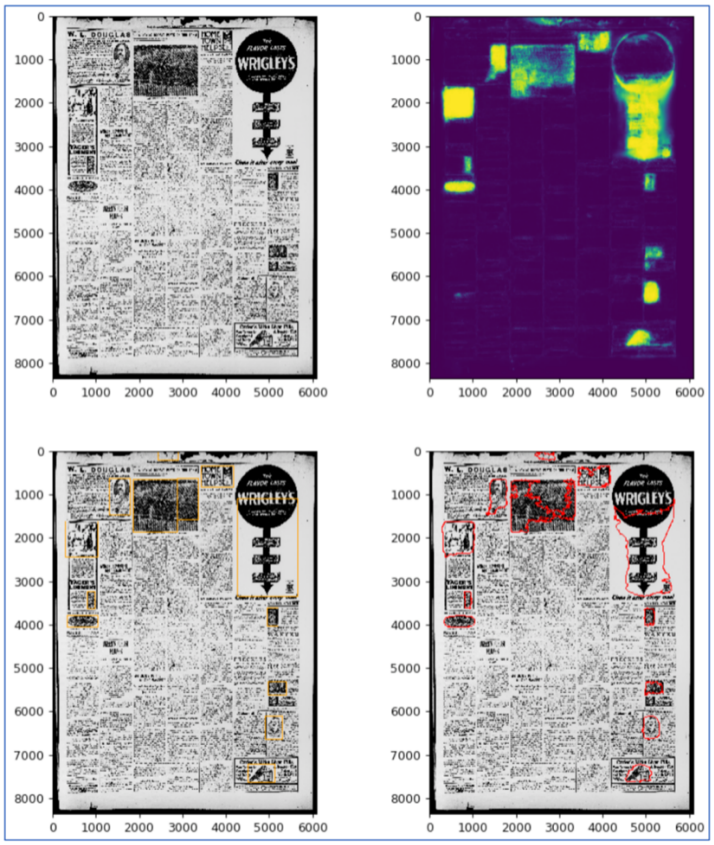


FIGURE 3. SEGMENTATION RESULT OF ENP_500_V4 ON A CHRONICLING AMERICA IMAGE (SN84026820_00271765095_1917050501_0153.JPG). CLOCKWISE FROM TOP-LEFT: (1) INPUT, (2) PROBABILITY MAP FOR FIGURE CLASS, (3) DETECTED FIGURES IN POLYGON, AND (4) DETECTED FIGURES IN BOUNDING-BOX. IN THE PROBABILITY MAP, PIXELS WITH A HIGHER PROBABILITY OF BELONGING TO THE FIGURE CLASS ARE SHOWN WITH A BRIGHTER COLOR.

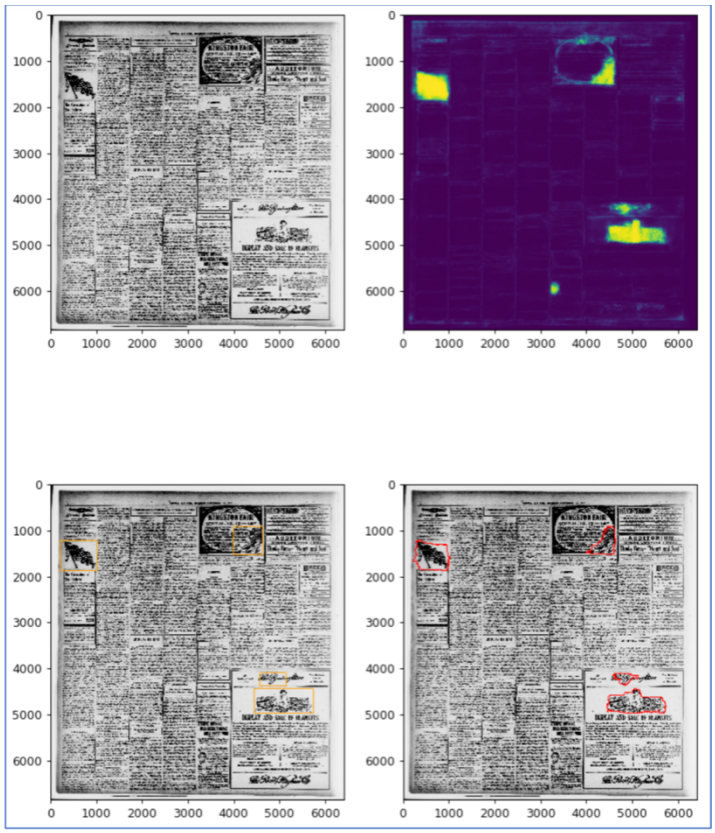


FIGURE 4. SEGMENTATION RESULT OF ENP_500_V4 ON A CHRONICLING AMERICA IMAGE (SN82014086_00295866135_1917091301_0116.JPG). CLOCKWISE FROM TOP-LEFT: (1) INPUT, (2) PROBABILITY MAP FOR FIGURE CLASS, (3) DETECTED FIGURES IN POLYGON, AND (4) DETECTED FIGURES IN BOUNDING-BOX. IN THE PROBABILITY MAP, PIXELS WITH A HIGHER PROBABILITY OF BELONGING TO THE FIGURE CLASS ARE SHOWN WITH A BRIGHTER COLOR.

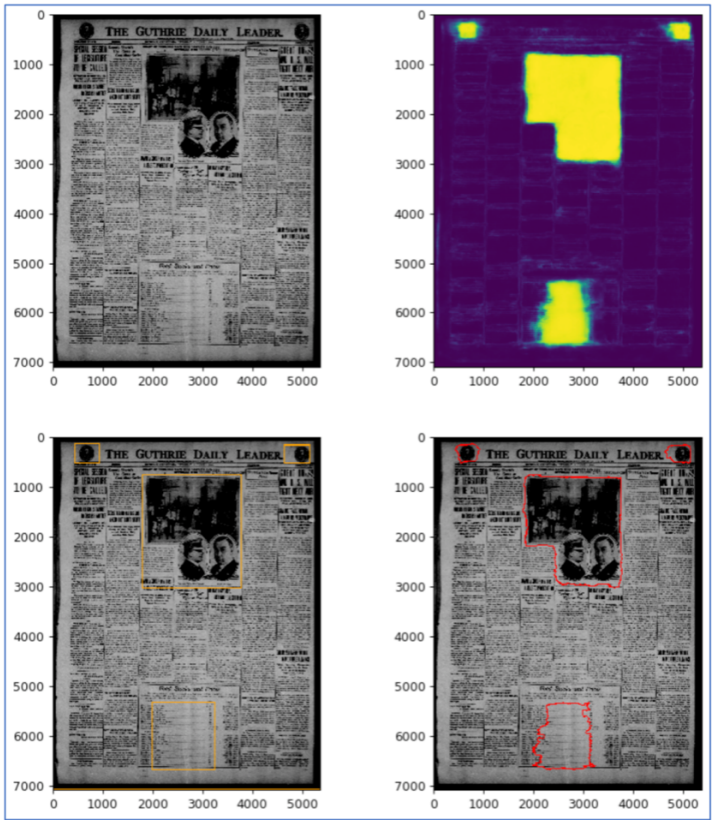


FIGURE 5. SEGMENTATION RESULT OF ENP_500_V4 ON CHRONICLING AMERICA IMAGE (SN86063952-19190805.JPG). CLOCKWISE FROM TOP- LEFT: (1) INPUT, (2) PROBABILITY MAP FOR FIGURE CLASS, (3) DETECTED FIGURES IN POLYGON, AND (4) DETECTED FIGURES IN BOUNDING-BOX. IN THE PROBABILITY MAP, PIXELS WITH A HIGHER PROBABILITY OF BELONGING TO THE FIGURE CLASS ARE SHOWN WITH A BRIGHTER COLOR.

6.2 Exploration: Graphic Element Classification and Text Extraction

This exploration was similar to the document segmentation exploration and became closer in goal and scope to that exploration over its iterations. Initially, the goal of this exploration was to find and localize figures, illustrations, and cartoons present in historical newspaper page images; classify the graphical content; and extract any text from the graphical content in order to generate a transcription of the textual content. By its second iteration, this exploration focused on fine-tuning the identification of graphical content in historic newspaper page images and the distinction of graphical content regions from textual content regions. The questions that guided this exploration throughout its development included: how might we use image zoning and segmentation, and text extraction from graphical regions, to generate additional information about newspaper pages in the Chronicling America corpus? Could image zoning and segmentation be used to pull out graphical content from Chronicling America newspapers? What benefits do different types or approaches to zoning and segmentation have for various information tasks? What strategies might be necessary to deal with rare content types in the training and evaluation of machine learning systems?

This exploration proceeded in two phases. The first phase established a conceptual model and workflow for a two-stepped approach that would result in segmented and classified graphical regions from historic newspaper pages and the segmentation and recognition of textual content in the graphical regions. Conceptually, our approach was

based on dhSegment, but instead of combining U-Net[4] and ResNet-50[5] as we did in the document segmentation exploration, we used the ResNeXt[6] classification model, a next-generation model from the ResNet that was employed in dhSegment. Such network combination belongs to the family of fully convolutional neural networks (FCN). We call our FCN that uses ResNeXt "U-NeXt." The goal was to see whether we could further enhance the results obtained with dhSegment by using this newer method. The model training was based on the pre-trained ResNeXt model for ImageNet. Finally, in the conceptual workflow, we planned to use EAST[7] text detection to find textual regions in the graphical images and use an optical character recognition process to recognize the textual strings within the graphical zones.

Our goal was to apply this conceptual model to newspaper page images from the Beyond Words project. In the above document segmentation exploration, the mIoU score was only 24%-26% on the Beyond Words dataset. We considered possible reasons that for the low mIoU scores. One possibility was that the feature extractor, ResNet, was not powerful enough to extract high-level features from the dataset for identification and classification. Notably, the ResNet model was reported by He et al. in 2015. However, in 2017, they reported a second-generation, ResNeXt, which beat the previous record on an ImageNet challenge.[8] Another possibility was that the rareness of some types of regions, such as maps, which comprise 1% of the regions, might skew the training process. As a result, we decided to test the ResNeXt model.

In addition, the data from Beyond Words were not sufficiently reliable for training purposes for this exploration.[9] One challenge with the data is that it does not include graphical content in advertisements; our model does not distinguish between graphical content in advertisements and graphical content in other types of content zones—graphical content is graphical content, at the stage of graphical content recognition and segmenting. In addition, not all graphical content has necessarily been marked on a page in the Beyond Words dataset. Since machine learning models will try to find all graphical content within the input page, such missing graphical regions can confuse the model during the training process. Another challenge, which we explore more fully below, is that ground-truth regions are not necessarily tightly mapped to the actual shape of the graphical region.

In its second phase of development, then, this exploration become a refinement of the original document segmentation exploration. We deconstructed the conceptual model described above and focused only on implementing the U-NeXt fully convolutional neural network for the purposes of graphical content extraction and classification. Our goal in doing so was to see if we could further improve upon the results reported by the dhSegment authors and the results we achieved in our document segmentation exploration that implemented dhSegment.

In the second phase of this exploration, we first conducted a pre-training investigation, which involved training and testing on the Europeana Newspapers dataset, since it is more comprehensively labeled than the Beyond Words

---

[4] Ronneberger, Fischer, and Brox, "U-Net."
[5] He et al., "Deep Residual Learning for Image Recognition."
[6] Xie et al., "Aggregated Residual Transformations for Deep Neural Networks." Compared to ResNet, ResNeXt uses grouped convolution (i.e. side-by-side convolution layers) in each residual block. The usage of grouped convolution was first mentioned in AlexNet. See Krizhevsky, Sutskever, and Hinton, "ImageNet Classification with Deep Convolutional Neural Networks."
[7] Zhou et al., "EAST."
[8] Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge."
[9] For more on Beyond Words, see the Library of Congress Labs' Experiments page, https://labs.loc.gov/experiments/?st=gallery Note as well that Beyond Words was not implemented with the purpose of creating training data or being used as training data.

data for our purposes. This pre-training investigation reached 91.30% pixel-wise accuracy and 57.19% for mIoU, with a testing performance of 81.90% pixel-wise accuracy and 48.18% mIoU. Note that the investigation did not reach the score of dhSegment on the ENP dataset in the document segmentation exploration. However, considering the ENP dataset and the Beyond Word dataset only share partial features, so it is not necessary to train the investigation to its best state. In fact, the observed convergence indicated the parameters were getting trained to fit the task and the model was ready for fine-tuning.

Then, we conducted a series of fine-tuning investigations, which involved four different approaches:

1. The first approach trained and tested U-NeXT on the Beyond Words dataset without using the Europeana Newspapers-trained classifier. This approach was meant to serve as a baseline design. We observed convergence in both training and testing curves, but the testing curve showed instability with rapid high and low variation during the investigation. Statistics showed that the classifier failed to recognize classes of editorial cartoons, illustrations, and maps. These three classes were the three rarest classes in the ground truth set, and the misrecognition issue is likely caused by the rareness of corresponding classes.
2. The second approach used the Europeana Newspapers-trained classifier as the beginning classifier. We then trained and tested it on the Beyond Words dataset. We added this design because using a pre-trained classifier for a similar task could help the overall fine-tuning investigations address challenges with reliable ground truth when working only with the Beyond Words dataset. Though performance indicators appeared promising, upon further investigation, the classifier trained during the fine-tuning experiment attempted to classify many pixels as background pixels after training convergence. Therefore, while the performance statistics are better than the first fine-tuning experiment numerically, the actual performance is worse, since none of the object classes (specific types of graphical content) were recognized.
3. The third approach replaced a deconvolutional layer with a resizing layer in the deep learning model. For this approach, we trained and tested on the Beyond Words dataset. This approach is designed to address a problem with the deconvolutional layer[10]; the resizing layer is perceived as an improvement on the overall technique. The pixel-wise testing accuracy is higher than in fine-tuning approach #1, but the mIoU is lower than in that fine-tuning approach. As with fine-tuning approach #2, we also found that pixel-wise accuracy and mIoU of the editorial cartoon, illustration, and map classes are zeros. However, the testing curve did not show the same instability as in fine-tuning approach #1. This result suggests that the resizing layer helped to address the challenge with the deconvolutional layer, so it is more stable than fine-tuning approach #1, though less accurate overall.
4. The fourth approach performed a two-class segmentation and classification, instead of six-class processes on the Beyond Words dataset for both training and testing. We reduced the number of classes to two because the training dataset is biased where there is a predominantly large number of background pixels compared to other classes of pixels.[11] Pixels in non-background classes comprise only 11.79% of the entire training dataset in total. By collapsing all the object pixels into one class, we can reduce the imbalance in the number of pixels in each class during training. The results indicate that training a classifier to learn information from rare classes is very hard. Combining five non-background classes into one class decreases the complexity of the task. The combined class segmentation outperformed the other fine-tuning experiments.

[10] Odena, Dumoulin, and Olah, "Deconvolution and Checkerboard Artifacts."

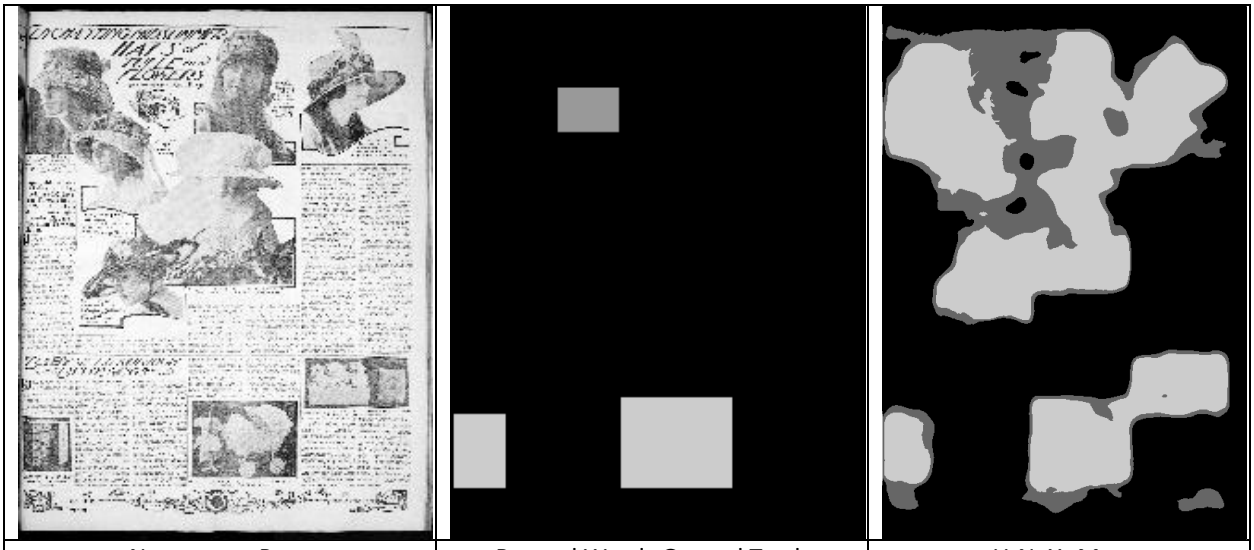
[11] There are 88.21% pixels in background class, but for the rest of classes, only 0.71% in editorial cartoon class, 2.89% in comics/cartoon class, 1.38% in illustration class, 6.64% in photograph class, and 0.18% in map class.

Approaches two through four are variants of the first approach. See Table 3 for a summary of results for the pre-training investigation and the four fine-tuning investigations.

TABLE 3. AVERAGE PERFORMANCE OF PRE-TRAINING INVESTIGATION AND FINE-TUNING APPROACHES.

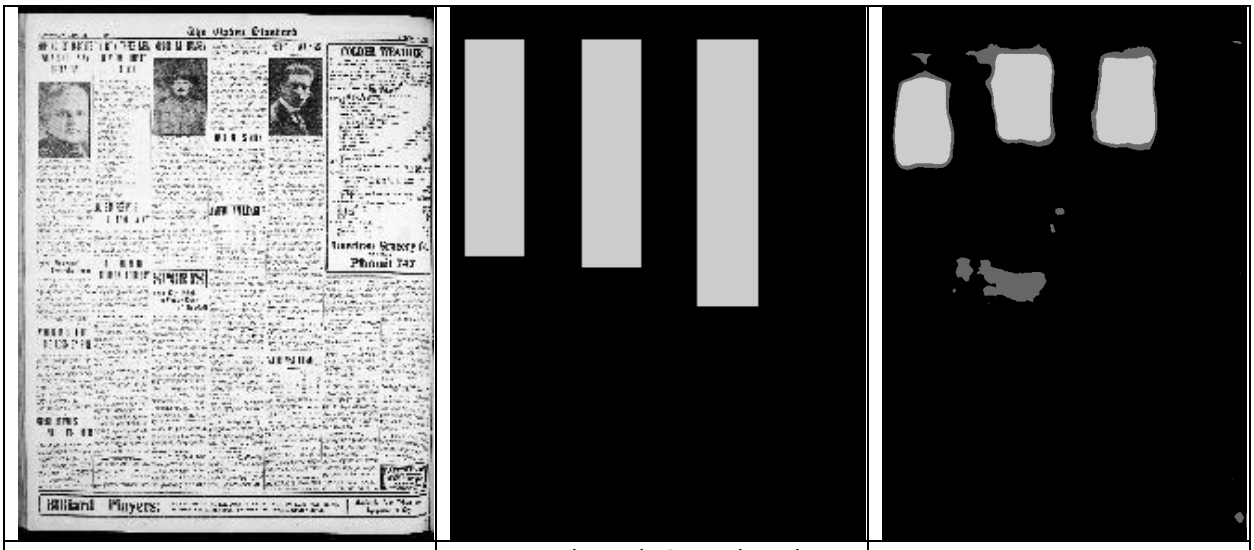
	Pre-Training Investigation		**Without Pre-Trained Europeana Newspapers Classifier**		**Using Pre-Trained Europeana Newspapers Classifier**		**Using Resizing Layers**		**Combined Two-Class Segmentation**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Accuracy	91.30%	81.90%	89.08%	80.11%	89.42%	85.53%	88.90%	86.69%	91.76%	88.89%
mIoU	57.19%	48.18%	50.43%	38.00%	41.21%	38.57%	51.31%	37.84%	71.44%	64.97%

From these investigations, we conclude that U-NeXt—especially the combined two-class segmentation—is promising for segmentation and zoning. At the same time, the fine-tuning approaches offered evidence that the Beyond Words dataset was not sufficient ground truth for our purposes. We found two issues. First, non-identified or incompletely identified graphical images in the Beyond Words dataset appear to be widespread to an extent that is problematic for training. For example, as shown in Figure 6, a large portion of a photograph in the document is missing from the ground truth but is captured by our U-NeXt classifier. Second, the rectangular regions do not necessarily match to the actual graphical content. For instance, as shown in Figure 7, the ground truth region includes a large portion of the textual content.



Newspaper Page	Beyond Words Ground Truth	U-NeXt Map

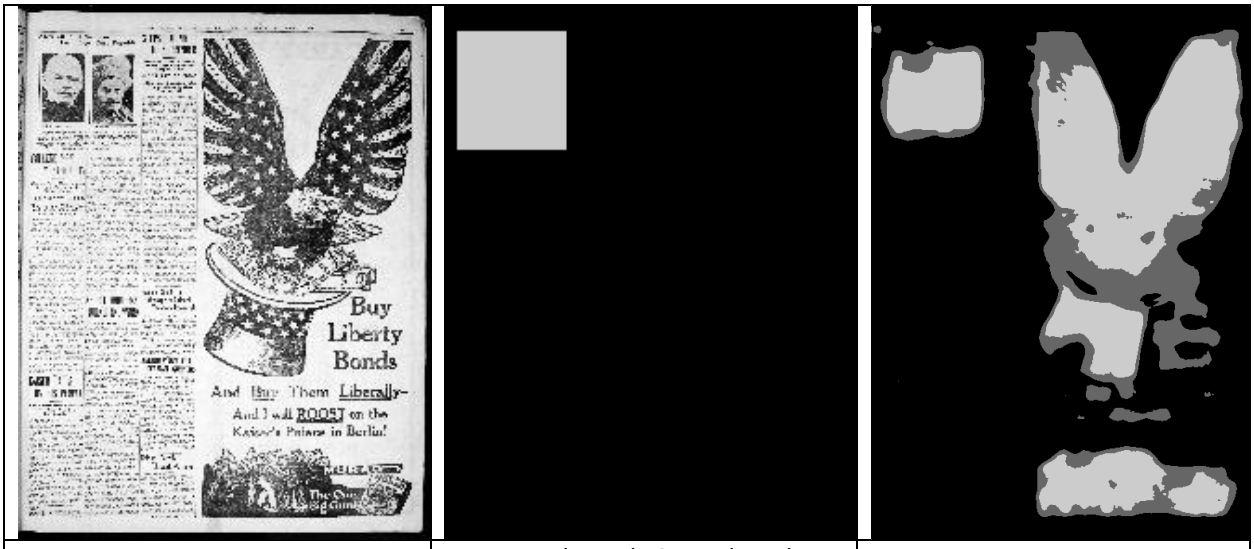
FIGURE 6. THE BEYOND WORDS DATA, WHICH WE TREATED AS GROUND TRUTH, IS MISSING MUCH OF THE GRAPHICAL CONTENT ON THE PAGE, WHILE THE U-NEXT MAP APPEARS MORE REPRESENTATIVE OF THE ORIGINAL PAGE.



Newspaper Page	Beyond Words Ground Truth	U-NeXt Map

FIGURE 7. EACH OF THE THREE COLUMNS WITH GRAPHICAL CONTENT HAVE MUCH LARGER BOUNDING BOXES IN THE GROUND TRUTH THAN WHAT CORRESPONDS TO THE ACTUAL GRAPHICAL CONTENT. THE U-NEXT MAP APPEARS MORE REPRESENTATIVE OF THE ORIGINAL.

These challenges with the Beyond Words dataset as ground truth for this exploration also lead us to believe that our classifier may be more accurate than the current statistical results would suggest. The U-NeXt model tries to fit the exact shape of the figure and graph region. Figure 8 shows, for example, that the model tried to fit the exact shape of the eagle on the right-hand side of the newspaper page. Since the ground truth included rectangular bounding boxes, we are not comparing like to like in our pixel-wise and mIoU comparisons.



Newspaper Page	Beyond Words Ground Truth	U-NeXt Map

FIGURE 8. IN THIS CASE, THE "BUY LIBERTY BONDS" ADVERTISEMENT IS NOT REPRESENTED IN THE GROUND TRUTH, LIKELY BECAUSE IT IS AN ADVERTISEMENT AND THEREFORE OUT OF SCOPE FOR THE BEYOND WORDS PROJECT. THE U-NEXT MAP, HOWEVER, RECOGNIZES THE GRAPHICAL CONTENT AND CLOSELY FITS TO IT.

6.3 Exploration: Document Type Classification

This exploration pursued whether we could effectively distinguish among handwritten, printed, and mixed (both handwritten and printed) documents within a collection of minimally processed manuscript materials at the Library of Congress. This exploration was guided by the questions: what features might be useful for influencing processing pipelines, for generating additional metadata, or for distinguishing among materials? How viable might large-scale indexing of documents be, for certain types of criteria? To what level of performance could we meta-tag document images? Would a deep learning model that had shown remarkable performance for natural scene images also show promising performance for document images? Or, to be more precise, would a feature extractor trained with millions of natural scene images also capably extract useful features for document images?

This exploration drew on current state-of-the-art methods in natural image and document classification. In particular, we extended the use of convolutional neural networks for classifying natural images to the task of classifying document images. Based on the findings of Harley et al. and Afzal et al., we used the VGG method with 16 categories (VGG-16) pre-trained on the Ryerson Vision Lab Complex Document Information Processing (RVL_CDIP) dataset.[12] The *RVL_CDIP* dataset, which is publicly available, consists of 400,000 document images that are divided into 16 evenly distributed classes. The dataset is provided in three different sets: training, validation, and test set. The training set contains 320,000 images of 16 different evenly distributed classes (i.e., about 20,000 images per class). Both validation and test sets together contain 40,000 images of 16 different evenly distributed classes (2,500 images per class).

We first set out to reproduce the results reported in the work of Harley et al. and assessed classification performances of VGG-16, pre-trained on ImageNet, and trained and tested with *RVL_CDIP* dataset. The advantage of doing so is that once we created a model trained on this large-scale document image dataset, we can reuse the rich features that this model has learned for many document analysis tasks, such as for our current ask of document type classification. The entire training process took only three epochs to converge with promising classification results. This indicates that features obtained from natural scene images (i.e., ImageNet) are general enough to be applied to documents. The resultant classification performance metrics—precision, recall, and f1-score—are shown in Table 4. On average, each metric shows around 87%, which aligns well with the result reported by Harley et al.

TABLE 4. PRECISION, RECALL, AND F1-SCORE OF VGG-16 AS TRAINED ON RVL_CDIP DATASET. THE ALPHABETIC LABELS CORRESPOND TO THE FOLLOWING LABELS: LETTER, FORM, EMAIL, HANDWRITTEN, ADVERTISEMENT, SCIENTIFIC REPORT, SCIENTIFIC PUBLICATION, SPECIFICATION, FILE FOLDER, *NEWS ARTICLE*, *BUDGET*, *INVOICE*, *PRESENTATION*, *QUESTIONNAIRE*, *RESUME*, AND *MEMO*. OUR CLASS OF INTEREST, HANDWRITTEN, IS BOLDED.

	A	B	C	**D**	E	F	G	H	I	J	K	L	M	N	O	P	Avg
Precision	86	74	98	**89**	89	73	90	99	89	92	87	91	78	91	92	88	87
Recall	94	79	97	**96**	91	73	93	91	97	86	83	86	79	73	94	91	87
F1	86	77	97	**92**	90	73	91	90	93	89	85	88	79	81	93	90	87

Next, we generated our own model for the specific task of classifying documents in one of three types, handwritten, typed/typeset, or mixed (both handwritten and typed/typeset). For this task, we retrained the model obtained from

[12] Harley, Ufkes, and Derpanis, "Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval"; Afzal et al., "Cutting the Error by Half"; Simonyan and Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition."

the above with a dataset derived from the Suffrage: Women Fight for the Vote campaign from By the People. The dataset for this exploration is hereafter referred to as suffrage_1002.

The suffrage_1002 dataset had 1,002 manually classified images and was balanced across handwritten, typed/typeset, and mixed. This ground truth set was created by members of the project team. The entire dataset was split into three sets—training, validation, and test—with a ratio of 8:1:1. In order to keep the class balanced during this split, we dropped three datapoints, one of each class. The final size of the dataset was therefore 999 images. See Table 5 for the breakdown by sets and class.

TABLE 5. CONFIGURATION OF SUFFRAGE_1002 DATASET.

	Handwritten	Typed/Typescript	Mixed	Total
Train	267	267	267	801
Validation	33	33	33	99
Test	33	33	33	99
Total	333	333	333	999

We use the same *VGG-16* architecture as above, but the output tensor was adjusted to have a shape of 3, the number of classes specified in suffrage_1002. Overall, our model's classification performance on the testing set shows about 90% for precision, recall, and f1-score, as shown in Table 6. We believe that these scores, which are a bit lower than those reported above in our attempts to reproduce Harley et al., are due to challenging characteristics of *mixed* type document images; for example, mixed materials may have negligible or statistically challenging amounts of handwriting in typed document and vice versa. See Figure 9 for examples.

TABLE 6. PRECISION, RECALL, AND F1-SCORES OF VGG-16 ON SUFFRAGE_1002 TESTING SET.

	Handwritten	Typed/Typescript	Mixed	Avg
Precision	89	91	90	90
Recall	97	94	79	90
F1	93	93	84	90

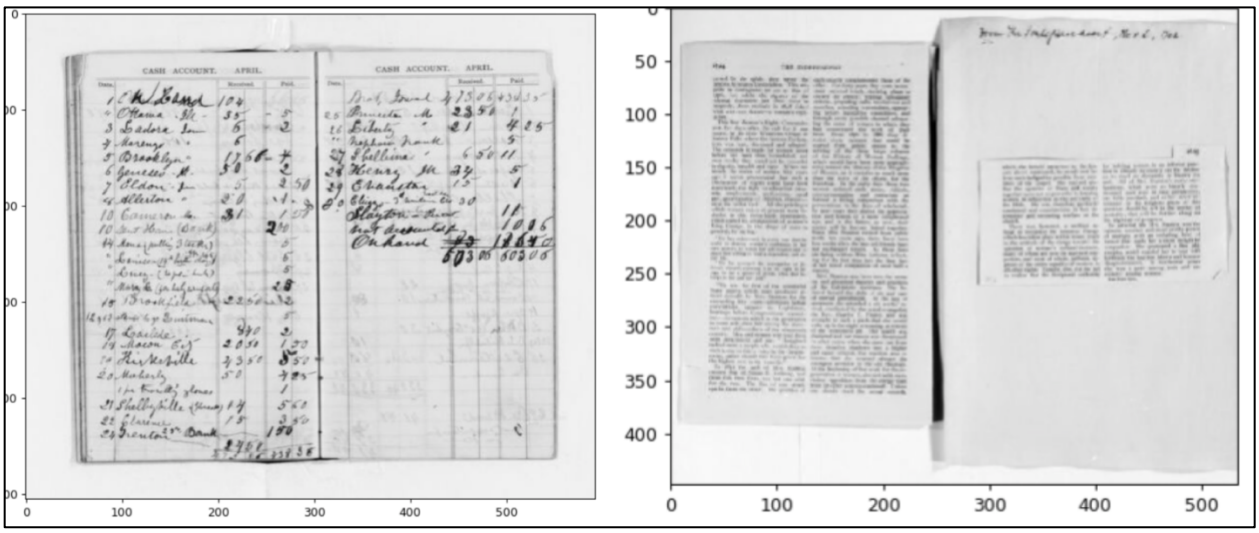


FIGURE 9. PREDICTION FAILURE CASES. IN THE LEFT EXAMPLE, THE MODEL CLASSIFIED THE DOCUMENT AS HANDWRITTEN RATHER THAN MIXED. NOTE THAT THE PRINTED REGIONS ARE VERY SMALL COMPARED TO THE HANDWRITTEN CONTENT IN THE IMAGE. IN THE RIGHT

EXAMPLE, THE MODEL CLASSIFIED THE DOCUMENT AS PRINTED RATHER THAN MIXED. HERE, THE HANDWRITTEN REGION IS VERY SMALL COMPARED TO THE PRINTED REGION IN THE IMAGE.

In the case of the examples in Figure 9, both images have technically been classified incorrectly, according to our current model and set of definitions. Both images depict documents that feature printed and handwritten content, and therefore both technically fit the definition of mixed. However, this example provides an opportunity to consider whether the actual mixed nature of these materials matters for processing purposes. If the Library of Congress were interested in this question for the benefit of helping to make decisions about how to handle particular types of materials—for example, that materials with significant handwritten content get passed to human experts, whether within the Library or outside of it—then materials with some, but limited print content are usefully grouped with handwritten materials. Likewise, if there was a strong mix of content types, that might also signal materials for human processing, whereas in the example of materials with minimal handwritten content, it may be fine to pass those materials off to more automated processes.

6.4 Exploration: Digitization Type Differentiation

The purpose of this exploration was to distinguish among digital images created through digitization from different source types. In particular, we sought to distinguish between items digitized from an original document item and those digitized from a microform reproduction of an original item. We expected that digitization source should be a relatively easy feature to distinguish and could have a variety of potential use cases for both internal processes and decision-making at the Library and for end users and researchers. A variety of questions sat behind this exploration. As with the document classification exploration, we wondered: what features might be useful for influencing processing pipelines, for generating additional metadata, or for distinguishing among materials? How viable might large-scale indexing of documents be, for certain types of criteria? To what level of performance could we meta-tag document images? We also wondered who might benefit from the ability to facet or search according to this particular criterion—digitization source—and how that might information might be made available.

This exploration proceeded in two phases. In both, we used ResNeXt, a deep learning method, to differentiate among images digitized from an original and those digitized from a microform reproduction. All images for this exploration came from the minimally processed manuscript collections included in the By the People Civil War campaign.

We first retrieved 36,103 images from the minimally processed Civil War materials and manually inspected 10,508 images, or slightly less than 30% of the total images. We determined digitization source ground truth for each of these 10,508 images. We then sampled a subset of 1,200 images from the 10,508 in a balanced set (600 images of each type, digitized from original and digitized from microform). In a 10% test of 120 sample images, the classifier was 100% accurate in classifying images as digitized from an original item or from a microform. We had concern, however, that this 100% accuracy was likely too good to be true when deployed over a larger set of images. We therefore proposed to compare the ratio of items digitized from microform to items digitized from original items to more comprehensively evaluate our approach. Based on the ground truth classification of the 10,508 images, we would expect a 1:16 ratio of images digitized from microform to images digitized from original items across the Civil War dataset.

In the second phase of this exploration, we fine-tuned the classifier, classified 36,103 images retrieved from the Civil War manuscript collections, and compared the number and ratio of expected classification to real classifications. In fine-tuning the classifier, we achieved a training accuracy of 98.52%, and a validation accuracy of 100%. In order to determine an ideal point between underfitting and overfitting the classifier, we calculated the harmonic mean of training performance and validation performance, to avoid both underfitting and overfitting.

We then used the fine-tuned classifier to classify the 36,103 images. Based on the first phase of this exploration, we expected that the image set would include 2,256 document images digitized from microfilm and 33,847 digitized from their original source. In reality, the classifier identified 2,834 images as digitized from microform and 33,269 images as digitized from an original item. See Table 7. Therefore, while we expected a classification ratio of 1:16 (microform to original), the achieved classification ratio was 1:11.74. The classifier was more aggressive in identifying images as having been digitized from a microform reproduction than we would have anticipated based on our initial tests.

TABLE 7. BREAKDOWN OF PROJECTIONS AND ACTUAL CLASSIFICATIONS OF CONTENT AS DIGITIZED FROM MICROFORM OR DIGITIZED FROM AN ORIGINAL ITEM.

Total Images	Expected Microform Source	Classified Microform Source	Expected Original Source	Classified Original Source
36,103	2,256	2,834	33,847	33,269

Without identifying the ground truth of each of the 36,103 items, we cannot be sure if the 1:16 ratio is entirely accurate. However, we do know that the classifier was more aggressive in identifying items as digitized from microform reproductions than in classifying them as digitized from an original item. For example, each of the items in Figure 10, Figure 11, Figure 12, and Figure 13, while actually digitized from original items, were classified as being digitized from microform reproductions.

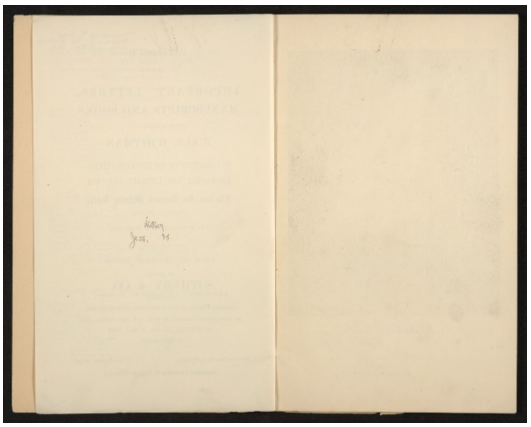


FIGURE 10. FACING PAGES OF A DOCUMENT, DIGITIZED FROM THE ORIGINAL, THAT THE CLASSIFIER CLASSIFIED AS HAVING BEEN DIGITIZED FROM MICROFORM. THERE IS MINIMAL CONTENT ON THE PAGES.

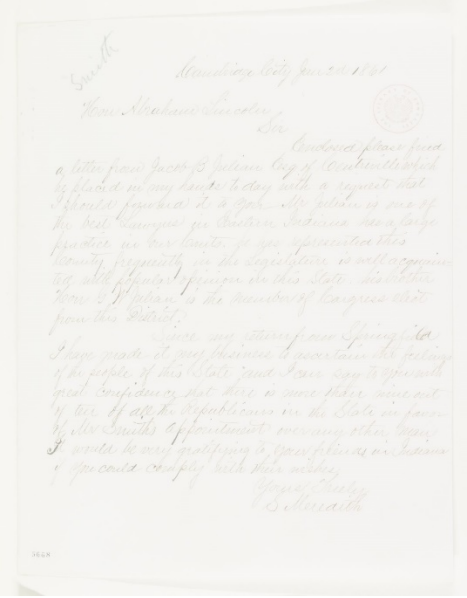


FIGURE 11. A HANDWRITTEN MANUSCRIPT PAGE IMAGE, DIGITIZED FROM THE ORIGINAL, AND CLASSIFIED BY OUR CLASSIFIER AS HAVING BEEN DIGITIZED FROM MICROFORM. THE CONTRAST OF THE IMAGE IS LOW.

FIGURE 12. A DIGITAL IMAGE OF A COIN, DIGITIZED FROM THE ORIGINAL ITEM, AND CLASSIFIED BY OUR CLASSIFIER AS HAVING BEEN DIGITIZED FROM A MICROFORM REPRODUCTION.

FIGURE 13. A DIGITAL IMAGE OF A PHOTOGRAPH, DIGITIZED FROM THE ORIGINAL ITEM, AND CLASSIFIED BY OUR CLASSIFIER HAS HAVING BEEN DIGITIZED FROM A MICROFORM SOURCE.

We believe these misclassifications were due to limitations of our training set, which did not include blank pages digitized from original items, photographs (that is, photographs of people that are included in the manuscript collections, for example), images with poor document contrast, or 3-d objects represented in the collections.

In the future, two options could effectively improve the performance further. First, we can expand the training database to include more examples and types of materials. Second, we can apply a pre-processing step to normalize the document image quality for the collection before the prediction stage. Overall, however, results were promising and suggest that automated type differentiation is viable and computationally cheap.

6.5 Exploration: Document Image Quality Assessment (DIQA) & Advanced DIQA

This exploration set out to analyze the quality of document images in minimally processed manuscript collections based on a variety of criteria with the goal of using information about image quality to inform future processes and toward making this information available for researchers looking for particular kinds of images (or images of particular quality). This exploration was guided by the questions: how might we distinguish among materials that most need human intervention, whether by Library of Congress staff or via crowdsourcing and the public, and those materials that might be well-suited to machine approaches? And when might materials be best suited to a combined approach? Could image quality assessments be useful in compiling ground truth and benchmarking sets in some capacity? Likewise, might such features be useful further downstream for users, to be able to facet for difficulty, for example? How might metadata about image quality of document images enrich understanding of individual items and of collections and corpora? To what extent can quality be computationally assessed, and might it help to better understand overall visual attributes of a dataset?

This exploration proceeded in two phases. In the first phase, we measured a set of image properties for each of 35,990 images retrieved from minimally processed manuscript collections included in the By the People Civil War

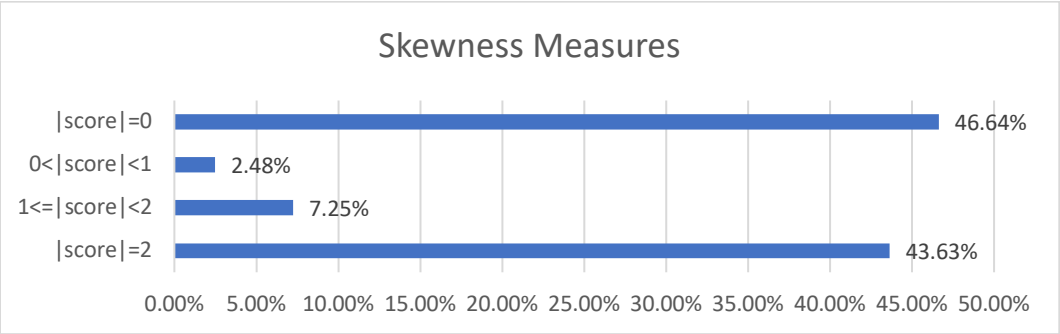
campaign. These image properties included skewness, contrast, range-effect, and bleed-through (background noise).[13]

6.5.1 Document Image Quality Assessment (DIQA)

6.5.1.1 Skewness

The skewness measure ranges from a score of -2 to 2, with any score other than 0 indicating skew is present. For example, a skew of -2 indicates significant counterclockwise skew, while a score of 2 indicates significant clockwise skew. Of the 35,990 images, nearly 50% of the images show no or negligible skew (a score of 0, or between 0 and |1|). Nearly 44% of the images (43.63%), are significantly skewed. See Chart 1.

CHART 1. SKEWNESS MEASURES OF 35,990 IMAGES FROM MINIMALLY PROCESSED CIVIL WAR COLLECTIONS. 43.63% OF IMAGES ARE SIGNIFICANTLY SKEWED (SCORE |2|).



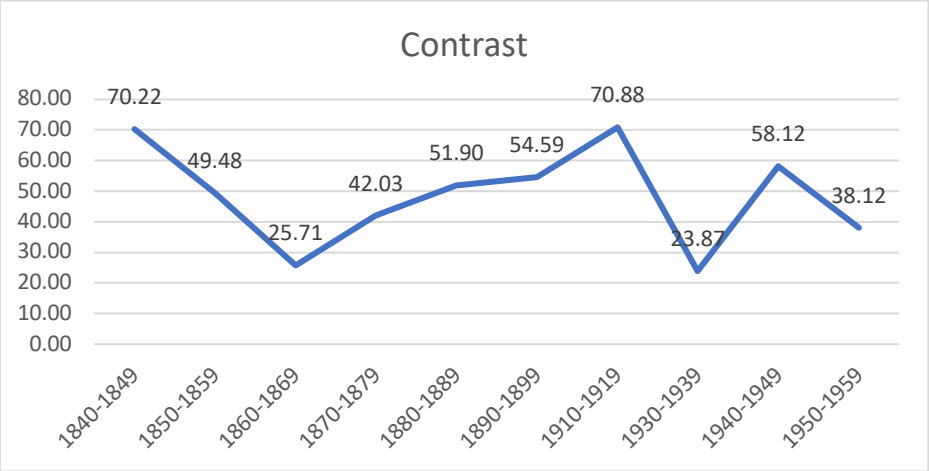
6.5.1.2 Contrast

Based on earlier work, the Aida team has determined that a contrast score of 30 or above indicates a good quality contrast in a digital image of a historic document;[14] the higher the contrast score, the better the visual quality. We plotted contrast over time (original date of document page, based on existing metadata) and determined that the two decades of materials represented in the Civil War collection fell below the threshold for a good contrast score: 1860–1869 and 1930–1939. See Chart 2.

[13] The document image quality assessment algorithms used in this exploration were developed as part of the Aida team's earlier efforts to assess qualities of newspaper page images from 1834 to 1922. See Lorang, Soh, Liu, Pack, and Rahimi, "Using Chronicling America's Images to Explore Digitized Historic Newspapers & Imagine Alternative Futures."

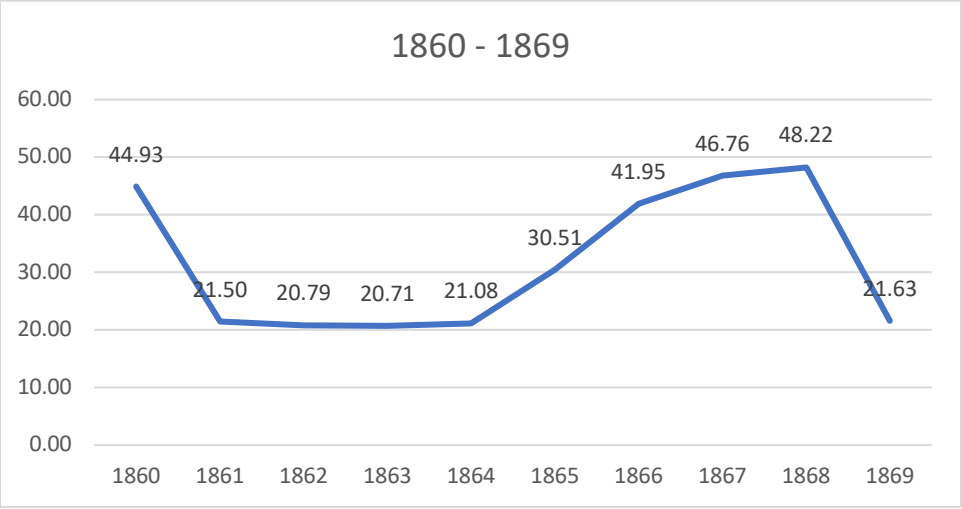
[14] Lorang, Soh, Liu, Pack, and Rahimi, "Using Chronicling America's Images to Explore Digitized Historic Newspapers & Imagine Alternative Futures."

CHART 2. AVERAGE CONTRAST SCORES OF MATERIALS WITHIN DECADE-RANGES. FOR EXAMPLE, MATERIALS FROM THE PERIOD 1840-1849 HAVE AN AVERAGE CONTRAST SCORE OF 70.22, WHILE MATERIALS FROM THE PERIOD 1930-1939 HAVE AN AVERAGE CONTRAST SCORE OF 23.87.



While images from 1930 to 1939 result in lowest contrast score, the significant majority of images in the collection—roughly 90%—date to between 1860-1869. These dates also overlap with the actual years of the Civil War, making the images from that decade most critical for further analysis. When we look more closely at the contrast scores within this decade, 1861, 1862, 1863, and 1864 all show average contrast scores below 22. See Chart 3. These data suggest that materials from most of the actual Civil War years have the lowest contrast in the collection and also that their contrast is below the threshold for good visual contrast. The low contrast can make these materials challenging for computational processing and also for human readers.

CHART 3. AVERAGE CONTRAST SCORES OF MATERIALS FROM THE DECADE 1860-1869 BY YEAR. MATERIALS FROM 1861 HAVE AN AVERAGE CONTRAST SCORE OF 21.50, FOR EXAMPLE, WHILE MATERIALS FROM 1868 HAVE AN AVERAGE CONTRAST SCORE OF 48.22.



We suspect that the low score could be document images that are digitized from handwritten letters, shown in Figure 14. There are two persistent features among these letters that could lower the contrast score. First, the original paper is often yellowish in hue. Second, the ink or pencil is often light—again, whether due to original inscription or time, or a combination of elements.

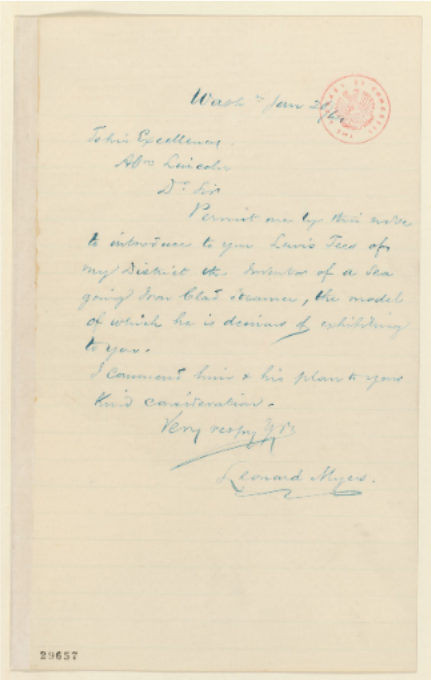
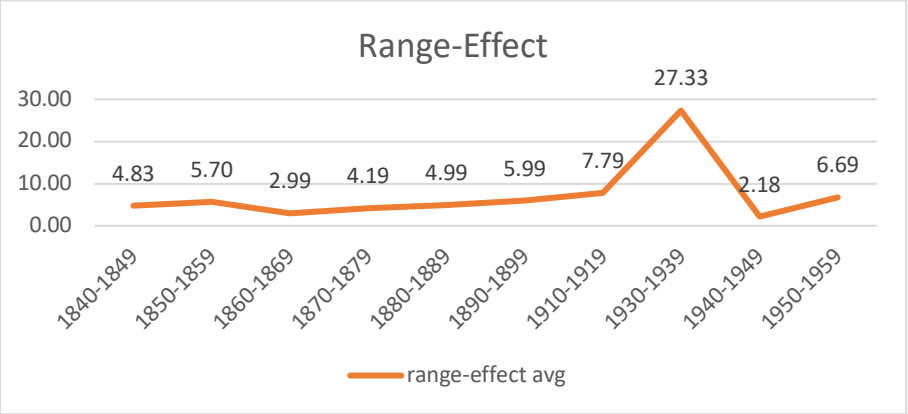


FIGURE 14. AN IMAGE WITH A LOW-CONTRAST SCORE FROM THE CIVIL WAR YEARS. FEATURES OF THIS IMAGES, SUCH AS THE COLOR OF THE PAPER AND THE LIGHT INK INSCRIPTION, MAY BE COMMON ACROSS THE MATERIALS, LEADING TO OVERALL LOW CONTRAST SCORES.

6.5.1.3 RANGE-EFFECT

The lower the range effect score, the better the quality of the image with regard to this feature. An ideal score is zero, and our team's earlier work with historic newspapers suggests that a range-effect score lower than three is indicative of a good-quality image. The materials from the decade 1860-1869 have an average range-effect score of 2.99. With materials in this decade comprising 90% of the images in the set, we believe range effect is not a substantial challenge for images in the Civil War collections. There are some noticeable outliers, as show in Chart 4, though in the case of the images from the period 1930-1939, the average score was significantly affected by two images with very high range-effect scores.

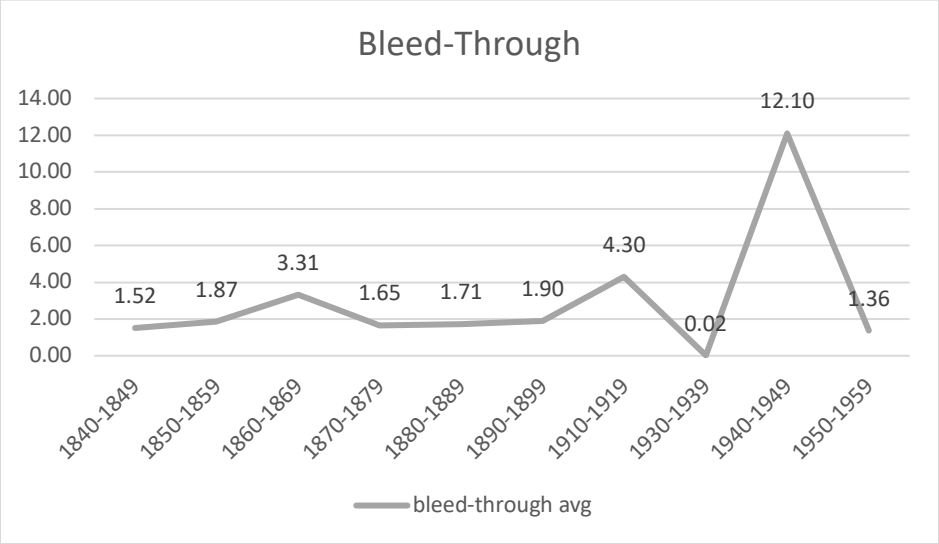
CHART 4. AVERAGE RANGE EFFECT OF THE CIVIL WAR COLLECTION OVER TIME. OVERALL, RANGE EFFECT IS LOW, WITH A SPIKE IN 1930-1939, LIKELY ATTRIBUTABLE TO A COUPLE OF DOCUMENTS IN A SMALL SET OF MATERIALS FROM THAT DECADE.



6.5.1.4 BLEED-THROUGH

Our analysis indicates that very few of the images studied suffer from significant bleed-through, which is a measure of noise in the overall image. While we do not have an objective measure for a good bleed-through score (meaning minimal bleed-through is present), an ideal score is zero. The majority of the 35,990 images return low bleed-through scores. See Chart 5. In this test, there are 76 images from the decade 1940-1949 that have high bleed-through scores and cause the average to spike in that decade.

CHART 5. AVERAGE BLEED-THROUGH/NOISE IN MATERIALS FROM THE CIVIL WAR COLLECTION, BY DECADE.



One caveat, however, is that in our processing, a document image is first converted into a grayscale image by the evaluation algorithm. Many of the pages in the collection are a yellow-hued paper that results in a dark background after the conversion. The presence of a dark background affects bleed-through evaluation and may result in a faulty evaluation.

6.5.2 ADVANCED DIQA

In the second iteration of DIQA, we combined elements of the original DIQA, document type differentiation, and document image segmentation explorations. Specifically, we measured a document image's compactness. In this case, compactness represents the number of zones (i.e., text blocks, figure) in a document image and may be considered as an indicator of document complexity.

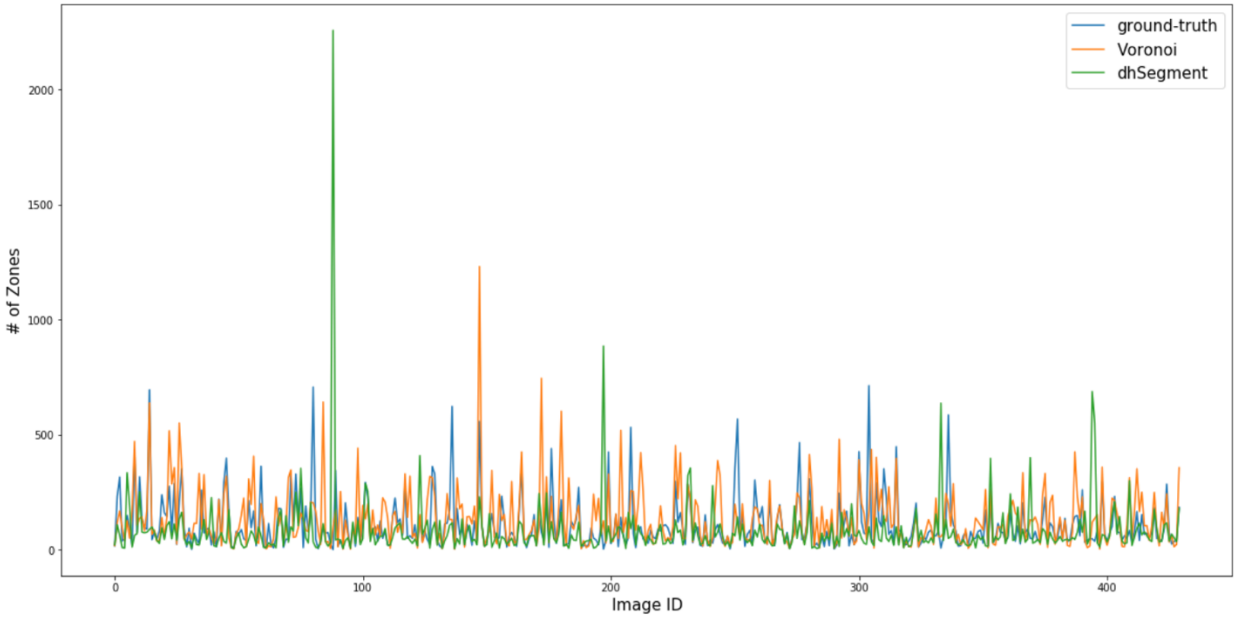
To proceed, we first assessed the compactness measures obtained by two segmentation algorithms on a dataset with known, reliable ground-truth. We then applied the compactness measure to minimally processed manuscript collections from the Civil War. Finally, we compared our measure of complexity with difficulty scores ascribed to materials through the Library of Congress's By the People site, to see if we could determine a correlation between our measure of complexity and document difficulty scores.

The two document segmentation algorithms used were (1) non-machine-learning, Voronoi diagram-based algorithm; and (2) machine-learning-based algorithm, dhSegment, used in other explorations reported here. We tested the two segmentation algorithms on 430 document images collected from the Europeana Newspapers dataset and counted the number of regions segmented by each algorithm. Then, we compared the result with the number of actual regions stored in the ground truth. See Table 8 and Chart 6.

TABLE 8. COMPARISON OF ACCURACY OF COMPACTNESS OF TWO ALGORITHMS.

	Voronoi-based Segmentation	Deep learning-based Segmentation
Mean Differences	75	77
STD Differences	97	154

CHART 6. THE COMPACTNESS OF THE EUROPEANA NEWSPAPERS DATASET.



Based on the mean and standard deviation of the difference between the number of zones detected by the algorithms and the ground truth, the compactness obtained by the non-machine-learning-based segmentation algorithm is slightly better than the machine-learning-based algorithm. In addition, when we looked at the number of zones and the compactness of the newspapers dataset, even though the number of zones detected by the two algorithms does not perfectly match with the number of actual zones, we observed a certain degree of similarity between compactness and the actual busyness of document images.

With this understanding, we then turned to several minimally processed manuscript collections from the Civil War, which are featured on By the People and are part of the Civil War campaign. Specifically, we used the Veroni-based approach to assess the compactness of four digitized document image collections, identified as “Civil War,” “Clara Barton,” “Letters to Lincoln,” and “Walt Whitman.” We analyzed within collections and by year/across time.

Three collections—Civil War, Letters to Lincoln, and Walt Whitman—show similar compactness distributions. The Clara Barton collection shows a thicker tail, which indicates that images in this collection tend to have a busier layout. Likewise, the Clara Barton collection shows notable changes in compactness across time represented in the collection. Our results suggest that items in the Clara Barton collections from 1862 to 1869 have more complex and busier layouts than those from 1850 to 1861. In the other three collections, we did not find notable compactness differences by time period.

Finally, we wondered if we might find a correlation between our compactness score and the difficulty score applied to an image through the By the People platform. We wondered: did the compactness, or busy-ness, of an image, which can be understood as a marker of complexity, correspond to images with higher difficulty scores? Ultimately, we could not correlate our measure of compactness with the difficulty score from By the People.

The difficulty score itself may not match human perception of difficulty as complex, non-linear relationships exist among visual features. In general, image quality assessment includes both machine and human perceptions of quality of an image. For machine perception, quality assessment evaluates difficulties to predict or categorize an image for a machine. And for human perception, quality assessment evaluates difficulties in understanding and interpreting an image based on the visual appearance.

6.5.3 Potential Applications

Much of this this exploration did not apply machine learning and instead was purely an image processing and image analysis exploration. We pursued it as part of this machine learning project in order grapple with which types of investigations require machine learning and when might other computational approaches be helpful to doing more with digital collections. In addition, such an exploration can facilitate future machine learning applications and endeavors. In cultural heritage digital libraries, administrative and descriptive metadata are common, even if the descriptive metadata are often limited. Various of our other explorations throughout this demonstration project, such as approaches to segmentation and classification, are toward enriching descriptive metadata and also have implications for enhanced structural metadata. As researchers begin to process large quantities of document images to develop robust classifiers or to develop generalizable automated systems, there is an increasing need for metadata about the image quality of the digitized document images, such as average intensity of an image, contrast, range effects, layout structure, etc., such that researchers might query and retrieve specific subsets of document images based on these qualities for testing.

6.6 Exploration: Document Clustering

This exploration extended from the initial documentation segmentation exploration and applied clustering to document images. Drawing on our other work with ResNet and dhSegment, we wondered whether document images clustered together share similar visual features recognizable to human observers. For example, would page images with graphical content cluster? Could we discern other clustering features? Could such clusters be useful in decision-making, for metadata generation, or other processes?

Two assumptions shaped this exploration. The first was that the deep visual representation of each datapoint contains enough feature information to be clustered. Second, in the clustered manifold, datapoints residing in the same neighborhood will share similar visual metadata with one another.

This exploration proceeded in two parts. In both parts, we used dhSegment to extract high-level visual features and then clustered the features using t-SNE, a state-of-the-art clustering method.[15] The dataset was a set of 96 page images from the Europeana Newspapers collection. From each of the 96 page images, we extracted a set of feature maps—so-called latent space—learned by a deep model, the ResNet-50 + U-Net that we trained for the document segmentation exploration. In this approach, the size of the latent space is calculated by the formula image width/32 x image height/32 x 2049.

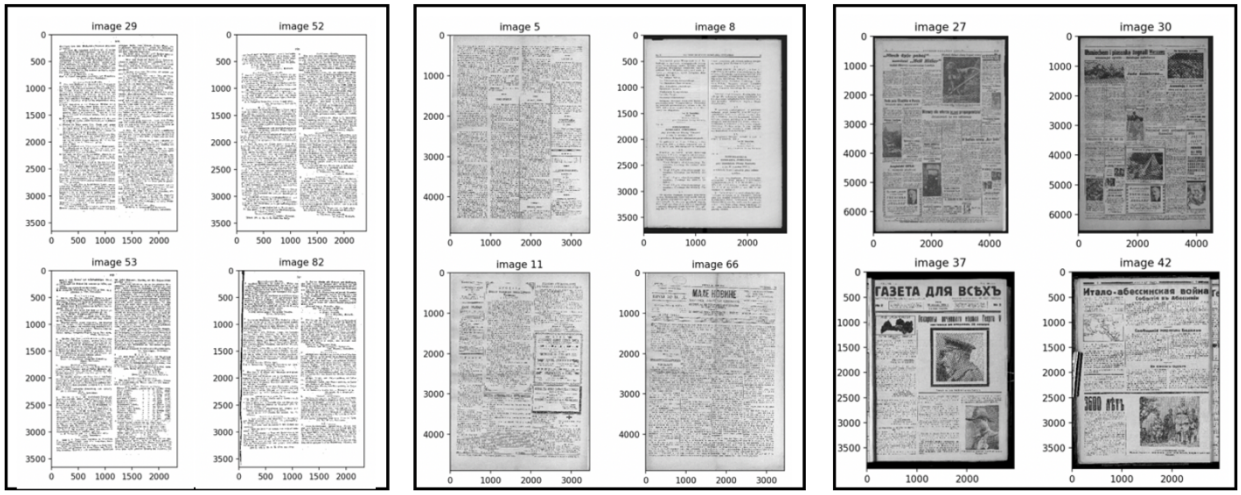
This exploration faced two challenges from the outset. First, the sizes of the input images varied, and they could not be reduced to the same proportions without distorting the images and thus their visual features. The size differences

[15] Maaten and Hinton, "Visualizing Data Using T-SNE."

were due to variation in the original dimensions of the newspapers as well as to variations introduced during the microform process (how much non-newspaper space was captured in the duplication process, for example). These differences meant that the size of our latent space was inconsistent. A second challenge for this exploration was that the latent space was too large. The resolution of an input image might be 1800 x 2400, meaning that the corresponding latent space became 1800 x 2400 x 2408. This scenario would contain redundant information and degrade clustering performance in both quality and computation time. To address these challenges, we performed dimensionality reduction recorded the intensity of features but not their spatial location.

These reduced latent space feature maps were then clustered using t-SNE. Ninety-six (96) datapoints in 2048-dimensional space grouped into roughly three clusters in two-dimensional space. Once the images were clustered in this low-dimensional space, we visually inspected and analyzed for (1) *intraclass correlation*, or whether datapoints in the same cluster share similar visual features, and (2) *interclass correlation*, or whether different clusters show dissimilarity to each other.

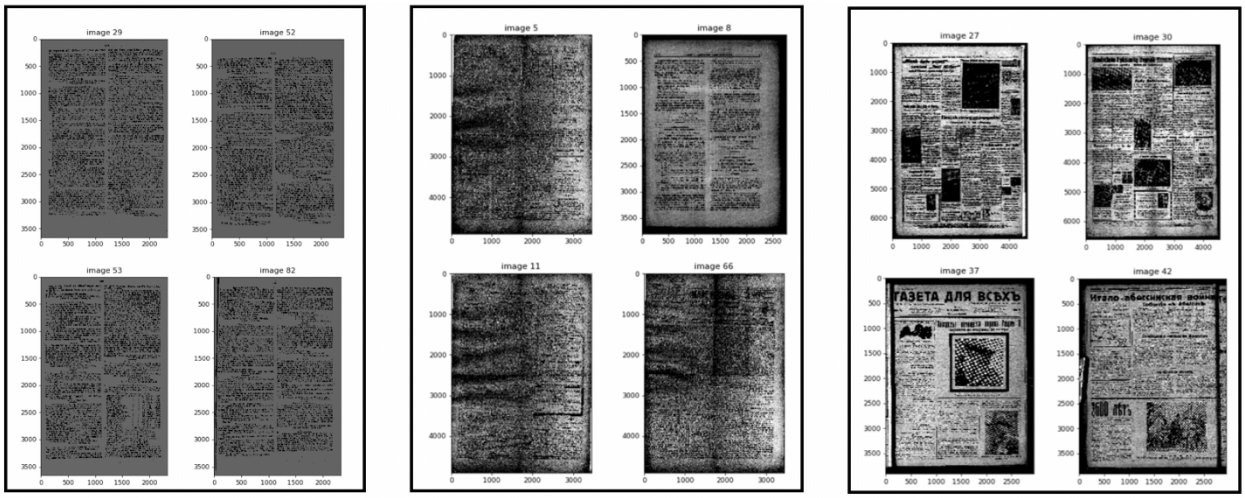
In the first part of this work, a visual inspection of sampled images from the same clusters does suggest shared visual features; for example, all four images in each box in Figure 15 show similar degrees of brightness and contrast. This result implies that there is a certain amount of intraclass correlation; images in the same cluster somewhat resemble each other.



(A) (B) (C)

FIGURE 15. IMAGES FROM THREE DIFFERENT CLUSTERS. IMAGES IN THE SAME CLUSTER SHARE SIMILAR CHARACTERISTICS, WHEREAS OTHER CLUSTERS SHOW DIFFERENT CHARACTERISTICS. FOR EXAMPLE, IMAGES IN (A) SHOW HIGH CONTRAST AND SIMPLE LAYOUT STRUCTURE. THE IMAGES IN (B) SHOW A RELATIVELY GRAY APPEARANCE WITHOUT FIGURE COMPONENTS. THE IMAGES IN (C) SHOW A RELATIVELY DARKER APPEARANCE WITH FIGURE COMPONENTS.

Following the first part of this exploration, we questioned whether the clustering results were simply based on the intensity value of the images. Thus, in the second part of this exploration, we clustered deep visual representations extracted from images that have been normalized to have zero mean and a unit standard deviation of intensity value. See Figure 16. From this second phase of the exploration, we observe two things. First, the clustering result using the deep visual representation excluding intensity shows a similar pattern to that of using the deep visual representation including intensity. This outcome suggests that the performance of our clustering approach is not based primarily on intensity features. Second, based on the observation that some datapoints sharing similar layout structures are slightly separated from each other compared to the first experiment clustering result, intensity does have an effect on the clustering process, even if it is not a primary basis of clustering.



(A) (B) (C)

FIGURE 16. IMAGES FROM THREE DIFFERENT CLUSTERS FOLLOWING INTENSITY VALUE NORMALIZATION. THE RESULTS SHOW SIMILAR CLUSTERING PATTERNS AS IN THE PREVIOUS FIGURE.

This exploration suggests that a set of deep visual representations of document images can be mapped into a low-dimensional space efficiently and effectively and neighboring datapoints show considerable visual similarity. This visual similarity is not based primarily on simple intensity features but rather on high-level visual features, such as layout density. We see potential in document clustering for enriched metadata as well as in using visually similar images for launching further study of materials within the Library of Congress and for allowing researchers/users to see visually similar images to those they are currently exploring.

For future development, we recommend looking at unsupervised approaches in order to build a more generic clustering solution not limited to a particular document domain or corpus. We also recommend exploring more sophisticated approaches to reducing dimensionality than what we have adopted here, in order to retain spatial information. And, we imagine combining this clustering approach with the results of document image quality assessment and the notion of a document complexity score, in order to see if there is a correlation between image quality assessment, complexity, and clustering.

7 Discussion

The explorations framed above only scratch the surface of the types of investigations to be pursued with machine learning and the information that can be gleaned from and about digitized materials, the collections in which they sit, and about organizational and institutional practices and beliefs. We knew from the outset of this demonstration project that scoping our work would be crucial; we were already aware of the magnitude of possibilities and potential investigations, whether for a short-term project such as this one or more sustained research and development. Nonetheless, through the above explorations, we developed a heightened awareness of the number of possibilities and challenges, both those social and technical, as well as of their scale. In this section, we move from the more specific questions and narrower areas of focus pursued in the explorations to a set of themes, ideas, and questions that served as a backdrop to or emerged over the course of the larger project.

7.1.1 Social

Processing image and textual data with existing machine learning platforms and programs is increasingly accessible. That is not to say that doing so is exactly plug-and-play, but the technical, conceptual, and domain knowledge needed

to simply pass data in to a machine learning pipeline and obtain some results appears lower each year. This perceived simplicity, however, hides significant complexity, nuance, assumptions and decision-making, and labor. Furthermore, this perceived simplicity has the potential to mask the implications of machine learning-generated knowledge, implications which range from the humorous and mundane to the profound and life-changing.

Domains considering implementing machine learning must engage deeply and critically with the technology, what it does, and what it means. For cultural heritage digital libraries, now is a critical moment to grapple with epistemologies of machine learning and the knowledge it structures, shapes, and appears to codify. Some elements of these epistemological conversations may transcend domains and applications, but these conversations also must be rooted in the specificities of the cultural heritage sector. In particular, libraries must grapple with their historical foundations and practices and with the potential consequences of these practices for machine learning. Previous and ongoing collecting and description practices, for example, were and are colonialist, racist, hetero- and gender-normative, and supremacist in other structural and systemic ways. These understandings are the foundation on which training and validation data will be created and assembled; they will become reinscribed as statements of truth, even as we elsewhere champion the potential of computational approaches to uncover hidden histories, identities, and perspectives in collections. To engage machine learning in cultural heritage must mean confronting these histories, committing to the hard work of acknowledgment and rectification, and not simply reproducing them and giving them a whole new scale of power. There should not be a future for machine learning in digital libraries that is not first and foremost committed to, in the words of Thomas Padilla, "responsible operations" and to all of the ongoing, cross-cutting work that responsible operations entail.[16]

Early in this demonstration project, Meghan Ferriter framed a range of different types of machine learning explorations and their outcomes. These included machine learning in the Library of Congress for description, discovery, and delight.[17] Ferriter's framing highlights another important feature in considering machine learning for digital libraries. Each of these endeavors--machine learning for description, discovery, and delight—has the potential to help people see materials from new angles, to peruse them in alternative ways, and to begin to frame additional questions and ways of thinking. At the same time, each of these purposes foregrounds different values and carries with it a different set of requirements and responsibilities. Naming and framing such purposes can help us think about the requirements and responsibilities of projects with these different ends. Building on Ferriter's "three Ds," we add as well "deployment" and "debate/dialogue." These categories need not be mutually exclusive, nor do we imagine a prescription for how machine learning in any of these realms should proceed. Instead, as a community of practice and as communities of researchers, what do we expect from projects and applications that proceed with these—and other—purposes in mind? Perhaps most critically, for any project that is about large-scale deployment, or a deployment of machine learning that may have significant implications for reasons beyond scale, what expectations do we hold as to what such projects must do, consider, make transparent? What contexts must we be able to see and understand?

7.1.2 Technical

Across the above explorations, several high-level themes and questions routinely surfaced related to what we might consider technical aspects of supporting and pursuing machine learning in the Library of Congress.

The first is that on a basic level, computational access to the Library of Congress's digital objects is relatively straightforward. We were able to retrieve significant data—image, textual—via the Library's application

[16] Padilla, *Responsible Operations*.
[17] Ferriter, et al., Kick-off meeting.

programming interface and other bulk download options. This collections as data approach is an important layer for machine learning.

Even with this relatively straightforward baseline access to the digital materials, however, we depended on our inside access to people at the Library, made possible through this demonstration project, in order to make sense of some of the data. For example, what does the difficulty score encoded in the Beyond Words JSON data mean? How was it determined and by whom? How do the coordinates and size information in Chronicling America OCR.xml files correspond with size and coordinate information in Beyond Words? Are there existing values for digitization source type for digital collections and items? *Such questions and challenges may suggest the need for additional levels of documentation and/or to new types of reference support needed in the Library of Congress as it facilitates emergent areas of research with its digital collections.* We anticipate that the Library's Mellon-funded project, Computing Cultural Heritage in the Cloud, will further advance thinking and conversations on these topics.

Basic, computational access to digital collections ticks one box in the roadmap toward machine learning. Machine learning approaches also require accurate ground truth data from which to learn and validate. In the case of the explorations framed above, even when it seemed we could utilize existing Library of Congress data—generated by in-house experts or members of the public through crowdsourcing—as ground truth, *ground truth data proved challenging*. Ultimately, we had to create ground truth sets ourselves or turn to externally available datasets that provided the type/nature of ground truth information needed. Sometimes, we had to create these ground truth sets because the data did not otherwise exist as verifiable data, as in the case of the handwritten-typed-mixed project, for example. In other cases, the nature of the ground truth did not fit with our proposed approach, as in the difference between the rectangular bounding boxes of the Beyond Words project and the shape-fitting segmentation of our efforts in document segmentation and graphical content extraction. This reality about ground truth data was not wholly unexpected and is not a criticism of the Library's efforts or of individuals' labor and effort over time. What it may suggest, however, is that the *bibliographic information and collections-centered metadata previously pursued in libraries is a limited vision of what will be needed for machine learning applications and new areas of research*.

The lack of robust, varied, sizable, well-documented ground truth is a significant technical challenge to the development of machine learning approaches for cultural heritage. Cultural heritage digital libraries need ground truth data particular to their types of materials and also relevant to the type and variety of questions information professionals, researchers of various domains, and other users wish to pursue about these materials. In broad strokes, machine learning models developed and trained on other types of ground truth sets skew toward the contemporary and born-digital and are transferable to digitized historical materials only to a point. Maringanti, Samarakoon, and Zhu report, for example, that the current learning models for photograph description have been developed on photographs of the modern world and do not fit well with historical photographs in a research library's collections.[18] Likewise, historical materials may introduce additional challenges of noise and quality, whether due to material conditions, legacies of care, intervening technological processes, and more. Furthermore, datasets for competitions that focus on historical documents are relatively small, they are not comprehensive of the range of materials in collections as large and diverse as those in cultural heritage institutions, the statements about ground truth represented in them are typically narrow in application, and such ground truth sets are often siloed.

The challenges around ground truth connect with other questions that surfaced across many of our explorations. These questions included, how might data created by users via the Library of Congress's crowdsourcing projects be used as ground truth data? What size of ground truth and training sets are necessary for different purposes? Are

[18] Maringanti, Samarakoon, and Zhu, "Machine Learning Meets Library Archives."

ground truth data created for one purpose transferrable for other purposes? What happens when we attempt to extrapolate from ground truth created for one purpose to another? Or when there isn't a direct match between ground truth data and output data?

Likewise, as a backdrop to a number of the explorations, we wondered about *the interplay of human expertise and processes and machine knowledge and processes.* What human-computer processes might be viably and validly adopted and operationalized as, say, part of a daily routine? What human-computer approaches are viable and valid in terms of effectiveness and efficiency in order to address issues of scalability? What value might there be in cross-learning, loop-learning, and cross-processing, where machines learn from humans, humans respond to and adapt understanding based on machine learning, and this looped learning informs processes and decision-making? Rather than seeing machine learning as an end, how can the Library of Congress embed and value critique across such a system, so that both human and machine assumptions are routinely tested? What are the foundational data and metadata needed and required to facilitate cross-learning and cross-processing? What is the place for data-science paradigms, where problems or issues are derived bottom-up—are surfaced through the collections and feature analysis—rather than top-down? We would be premature and ill-equipped to answer such questions based on only the explorations above, but we highlight them here as recurrent questions and provocations over the course of the explorations.

7.1.3 Social-Technical

In separating the technical from the social above, our intention is not to suggest a binary, a neat division, or a siloing of responsibility.[19] There is a critical interplay between the social and the technical in machine learning, not least because machine learning has significant social consequences. At best, machine learning will be incomplete without the social and cannot proceed without the technical, and technology cannot be divorced from the societies and individuals that develop it or the social realities it constructs. Our distinction above is largely one of convenience, and the lines between the social and technical blur quickly in our recommendations.

8 Recommendations

As the largest library in the world and with the ambitious, forward-looking digital strategy announced in 2019, the Library of Congress is uniquely situated to play a leadership role in advancing the theory and practice of machine learning in the cultural heritage sector. With this leadership role in mind, we propose two top-level recommendations for the Library of Congress's efforts around machine learning and as it moves forward in its work to "throw open the treasure chest," "connect," and "invest in our future."[20] The first is that the Library should focus the weight of its machine learning efforts and energies on social and technical infrastructures for the development of machine learning in cultural heritage organizations, research libraries, and digital libraries. Second, we recommend that the Library invest in continued, ongoing, intentional explorations and investigations of particular machine learning applications to its collections.

What we do *not* recommend at this particular moment in time is the broad application of machine learning to the Library's digital collections with the purpose of broadly making claims about the materials or restructuring access to them. On a very practical level, such broad application would be premature due to the challenges with ground truth data and validation and considering the many critical conversations yet to take place around whose, and what, human knowledge becomes the basis of machine learning. We advise against a "more product, less process"

[19] Padilla refers to the "technical, organizational, and social challenges" as multiple, integral facets of a machine learning agenda for libraries. See *Responsible Operations*, p. 6.
[20] Library of Congress, "Digital Strategy for the Library of Congress."

approach to machine learning applications. The ways in which machine learning-generated knowledge stands to influence decision-making and codify particular understanding are too profound and too powerful to adopt such an approach, or make such a commitment, at this nascent stage.

Below, we provide more detail about the two top-level recommendations and offer several short- and medium-term recommendations in support of these top-level recommendations. Each of these top-level recommendations directly map to the three goals of the "Digital Strategy for the Library of Congress." In addition, while we hope the role and importance of people will be clear in everything that follows, we want to say directly here: people are central to all of the recommendations that follow. None of the recommendations imagine a library without information professionals and experts. Any future for machine learning in libraries will require an investment in people with many types of expertise, and a best-case future for machine learning in cultural heritage organizations is that the people who work in them are able to bring even more of their experience and expertise to bear.

8.1 FOCUS THE WEIGHT OF THE LIBRARY'S MACHINE LEARNING EFFORTS AND ENERGIES ON SOCIAL AND TECHNICAL INFRASTRUCTURES FOR THE DEVELOPMENT OF MACHINE LEARNING IN CULTURAL HERITAGE ORGANIZATIONS, RESEARCH LIBRARIES, AND DIGITAL LIBRARIES.

We recommend that the Library dedicate itself to a range of infrastructure projects that will create a strong foundation for machine learning in the profession and field, particularly as applied to historical cultural heritage materials. The paramount machine learning need within the cultural heritage sector at this time is the development of infrastructure.[21] These infrastructures include educative infrastructures, through which cultural heritage professionals develop further literacy in computational thinking and methods, particularly through the lens of critical information studies. The needed infrastructures also include platforms for conversations—and the pursuant conversations themselves—about the language of description and the corresponding social and cultural values signaled in that language, as well as about who we engage in these processes and applications. Likewise, the needed infrastructures include pathways for gathering and delivering machine learning models and verifiable learning data that extend beyond individual projects, as well as for bringing together cross-domain researchers toward the purpose of machine learning for cultural heritage.

This top-level recommendation, that the Library focus the weight of its energies on social and technical infrastructures, connects with many of the goals in the Library's digital strategy. This recommendation aligns particularly well with the Library's interests in *maximizing use of content*; *supporting emerging styles of research*; *welcoming other voices*; *driving momentum in our communities*; *cultivating an innovation culture*; *ensuring enduring access to content*; and *building toward the horizon*. (See Table 9, which maps our recommendations to the Library of Congress digital strategy.)

Under this broad umbrella, we propose several more specific recommendations, through which we believe the Library of Congress could have the most immediate and most significant impact, drawing on both its status and position as well as existing areas of expertise. We have not attempted to be exhaustive in these recommendations; instead we emphasize activities that we consider critical or high priority given community needs, the Library's existing expertise/leadership, and opportunity. To that end, we recommend that the Library of Congress should:

- Develop a statement of values or principles that will guide how the Library of Congress pursues the use, application, and development of machine learning for cultural heritage.

[21] Padilla has framed this need as a research agenda for libraries, including the areas of machine learning, data science, and artificial intelligence. See Padilla for an even more extensive list of needs.

- Create and scope a machine learning roadmap for the Library that looks both internally to the Library of Congress and its needs and goals and externally to the larger cultural heritage and other research communities.
- Focus efforts on developing ground truth sets and benchmarking data and making these easily available.

In the following subsections, we describe these recommendations in more detail. We also map each recommendation to areas of investigation and challenge areas outlined in *Responsible Operations*. (See Table 10, which maps each recommendation to areas of investigation and challenge areas outlined in *Responsible Operations*.)

8.1.1 Develop a statement of values or principles that will guide how the Library of Congress pursues the use, application, and development of machine learning for cultural heritage.

The Library of Congress should articulate a statement of values or principles with regard to the adoption, use, and development of machine learning. Such a statement can address machine learning and cultural heritage broadly—what is the vision of machine learning for cultural heritage that Library of Congress aspires to—and also frame the values and principles under which the Library of Congress will pursue the development and application of machine learning. If units within the Library seek to apply machine learning to collections, under what principles and values should that work proceed? What are the expectations around transparency and explainability, both for internal and external audiences, for example? Or around confronting problematic historical knowledge and knowledge structures in training data? The crafting of such a statement of values or principles is an opportunity for developing increased literacy and fluency around machine learning, if the Library engages its staff broadly in the developing and education around such a statement.

This recommendation maps to the following investigation areas and challenges in *Responsible Operations*:

- Committing to Responsible Operations
  - Managing Bias
  - Transparency, Explainability, Accountability
  - Distributed Data Science Fluency

8.1.2 Create and scope a machine learning roadmap for the Library that looks both internally to the Library of Congress and its needs and goals and externally to cultural heritage and other communities of research and practice.

For this demonstration project, we scoped our explorations above as a response to the seemingly endless array of opportunities for applying machine learning to the Library of Congress's digital collections. As the Library of Congress continues to explore the intersection of machine learning and digital cultural heritage, the Library likewise will need to focus and scope its efforts. Such scoping and an overall roadmap are necessary for the Library to influence and have a strong impact on the development of machine learning in cultural heritage organizations.

The roadmap should be informed by the statement of values recommended in 4.1.1. In addition, other recommendations in this report may be points on that roadmap. The investigation areas in *Responsible Operations* may provide a useful framework for such a roadmap. What are the Library's goals and objectives in each of the investigation areas? Will it pursue all of the areas or prioritize particular areas? With regard to the Library's goals and objectives, are there investigations areas that the Library would add? Do recommendations from that report offer ways of thinking about and structuring a longer-term roadmap?

This recommendation maps to the following investigation areas and challenges in *Responsible Operations*:

- Committing to Responsible Operations
    - Transparency, Explainability, Accountability
    - Distributed Data Science Fluency
- Workforce Development
    - Investigating Core Competencies
    - Committing to Internal Talent
- Any/all of the investigation and challenge areas that the Library of Congress would choose to prioritize in its own roadmap and plan

8.1.3 Focus efforts on developing ground truth sets and benchmarking data and making these easily available.

One key way for the Library of Congress to advance machine learning for cultural heritage is creating and distributing ground truth sets drawn from its diverse digital collections and making available benchmarking data for computational approaches on those sets. Ground truth data and benchmarks will allow researchers—including cultural heritage professionals, computer scientists, and developers—to focus their energies and research, development, and analysis, rather than on creating one-off, niche datasets. The availability of ground truth and benchmarks also create the possibility of more rapid development around particular problem domains.

Creating and distributing ground truth sets will foreground the significance of metadata, including technical, structural, and descriptive. For descriptive metadata, we recommend distinguishing between at least two types of descriptive metadata, one that is descriptive of the content of the historical materials, including metadata about what is depicted and represented as well as how, and another that is descriptive of the properties of the image, including features such as digitization source, contrast, skew, noise, range effect, complexity (or a difficulty measure of some sort). Underscoring this idea is that the ground truth sets will have interest to researchers of many disciplines and research interests (those interested in the materials themselves as cultural objects and those interested in them for their value to computer science development, for example).

Within this recommendation, we offer two sub-recommendations:

8.1.3.1 Development of DocuNet

We recommend the Library of Congress develop, or partner in developing, DocuNet, an image database of historical documents with accompanying taxonomic and typological metadata. DocuNet would be valuable to researchers in library and related sectors and also to information science and computer science researchers. We see it as one effective way to encourage additional machine learning researchers and those interested in computer vision, among other domains, to delve into historical document analysis. Features or characteristics important to a DocuNet are ground-truth (e.g., document types, coordinates of article regions, etc.); openness; diversity and balance (e.g., different document types should be comprehensively covered and equally distributed); and clear objectives (e.g., segmentation, classification, clustering, etc.).

8.1.3.2 Pursuit of Low-cost Ground-truthing

We also recommend that the Library explore options for, and contribute to efforts to advance, low-cost ground-truthing. Having subject matter experts hand-label data is expensive and is a barrier to machine learning, whether in the Library of Congress or on external research teams. Importantly, by low-cost ground-truthing, we do not mean exploitative labor such as through Mechanical Turk. Instead, the Library could pursue heuristics-based models. In this form of ground-truthing, computers learn the heuristics that are created by humans, and then computers label data using the heuristic rules, constraints, distributions,

and/or variances of the dataset. Such an approach may be less accurate than item-by-item expert-labeled ground truth, but it may still be able to produce effective machine learning systems. While the potential for low-cost ground-truthing of this sort remains to be seen, we believe it is a worthwhile area of inquiry as part of a larger commitment to support for machine learning infrastructure.

Recommendation 4.2.3 and its corresponding sub-recommendations map to the following investigation areas and challenges in *Responsible Operations*:

- Committing to Responsible Operations
  - Managing Bias
- Description and Discovery
  - Enhancing Description at Scale
- Shared Methods and Data
  - Shared Development and Distribution of Training Data

8.2 INVEST IN CONTINUED, ONGOING, INTENTIONAL EXPLORATIONS AND INVESTIGATIONS OF PARTICULAR MACHINE LEARNING APPLICATIONS TO ITS COLLECTIONS.

Much as it did through this demonstration project and other activities (its Innovator-in-Residence program, for example), the Library should continue to invest in explorations and investigations of particular applications of machine learning on its collections, with an eye toward both internal operations and impacts on external users. Continued explorations, tests, and experiments will prove crucial to the ongoing inquiry needed to more fully evaluate the potential of machine learning for digital libraries. We recommend that such explorations are framed and understood as *intellectual endeavors* rather than being large output-driven and are collaborations among computer scientists, developers, and information professionals, drawing in other participants and stakeholders as appropriate to the project. We also encourage the Library of Congress to be careful in the presentation of machine learning generated data, particularly when that data might be read or experienced by others as uncontested knowledge or fact about cultural heritage materials, and also with care and concern about what is absent as well as what is present.

This recommendation, that the Library invest in continued, ongoing, intentional explorations and investigations of particular machine learning applications to its collections, connects with many of the goals in the Library's digital strategy. These include, in particular, *supporting emerging styles of research*, *welcoming other voices*; *driving toward momentum in our communities*; *cultivating an innovation culture*; and *building toward the horizon*.

Again, we propose several more specific recommendations, through which we believe the Library of Congress could have the most immediate and most significant impact, drawing on both its status and position as well as existing areas of expertise. We have not attempted to be exhaustive in these recommendations; instead we emphasize activities that we consider critical or high priority given community needs, the Library's existing expertise/leadership, and opportunity. To that end, we recommend that the Library of Congress should:

- Join the Library of Congress's emergent efforts in machine learning with its existing expertise and leadership in crowdsourcing. Combine these areas as "informed crowdsourcing" as appropriate.
- Sponsor challenges for teams to create additional metadata for digital collections in the Library of Congress. As part of these challenges, require teams to engage across a range of social and technical questions and problem areas.

- Continue to create and support opportunities for researchers to partner in substantive ways with the Library of Congress on machine learning explorations.

In the following subsections, we describe these recommendations in more detail. We also map each recommendation to areas of investigation and challenge areas outlined in *Responsible Operations*.

8.2.1 Join the Library of Congress's Emergent efforts in machine learning with its existing expertise and leadership in crowdsourcing. Combine these areas as "informed crowdsourcing" as appropriate.

Through its By the People application and campaigns, as well as through earlier efforts, the Library of Congress has established a strong portfolio of crowdsourcing experience. Through the Library of Congress Labs, the Library also has strong leadership in crowdsourcing, including technical development for crowdsourcing, and in designing and developing challenges. We see significant potential in bringing together machine learning and crowdsourcing efforts, as an effort to combine an existing strength of the Library with an emergent area of interest and impact. Such a joining of efforts would allow the Library of Congress to make even greater use of crowdsourced information, toward challenges of scalability. Doing so also creates the opportunity for greater conceptual understanding and practical development. For example, joining these areas, even in a limited way, would allow the Library to research cross-learning and looped learning. Such a combined approach has the potential to improve machine learning models, particularly in applications that require a higher-level of understanding. In a hypothetical project, members of the crowd might receive labeled data from a model; users then revise the labels, and the model improves its predictions based on those revisions. With each successive iteration, the model improves further.

This recommendation maps to the following investigation areas and challenges in *Responsible Operations*:

- Description and Discovery
  - Enhancing Description at Scale
- Workforce Development
  - Committing to Internal Talent
- Shared Methods and Data
  - Shared Development and Distribution of Training Data

8.2.2 Sponsor challenges for teams to create additional metadata for digital collections in the Library of Congress through Machine Learning. As part of these challenges, require teams to engage across a range of social and technical questions and problem areas.

The Library has a history of creating and sponsoring challenges, such as the Congressional Data Challenge and challenges focused on Chronicling America data. We recommend that the Library build on this prior experience to organize and offer new sponsored challenge opportunities about machine-learning generated metadata. Such explorations have the potential to move forward a range of critical conversations and needs. The purpose of this recommendation is multipart: (1) To see what types of metadata researchers/teams might produce. What metadata is of interest to them? (2) To encourage the creation of particular types of metadata, including through an expanded sense of what descriptive metadata might include and what is of descriptive value (e.g., metadata that are representational of the content and metadata that are descriptive of features such as noisiness, quality, and so on); (3) To anchor critical engagement with core problems, such as of bias in the data and in what may be produced, as inseparable from technical development; and (4) To emphasize, underscore, and champion that cross-disciplinary, community-centered and community-engaged development is required for responsible machine learning.

This recommendation maps to the following investigation areas and challenges in *Responsible Operations*:

- Committing to Responsible Operations
  - Managing Bias
  - Transparency, Explainability, and Accountability
- Description and Discovery
  - Enhancing Description at Scale
- Shared Methods and Data
  - Shared Development and Distribution of Methods
- Sustaining Interprofessional and Interdisciplinary Collaboration

8.2.3 Continue to create and support opportunities for researchers to partner in substantive ways with the Library of Congress on machine learning explorations.

Even if the Library were able to dedicate many staff members to a full-time focus on machine learning, the challenges of machine learning for cultural heritage are large and significant enough that the Library will need to continue its collaborations with external researchers. Such opportunities need not be sponsored by the Library itself, though they could be. However they are facilitated, we recommend that the Library see formal collaborations as central to taking this machine learning work forward. As researchers who have worked with Library of Congress data for many years—and over which time we have had many positive and helpful interactions with Library staff, who probably went well beyond the call of duty in their help to us—we benefitted in significant ways from the additional levels of access to Library staff this this particular demonstration project and the formal collaboration afforded. Understandably, the Library cannot support every machine learning endeavor at this level or engage with every research team in this way, and there will be challenges of scale. Nonetheless, we recommend that some measure and shape of formal collaboration opportunities be part of the Library's support for both machine learning explorations and larger social and technical infrastructures.

This recommendation maps to the following investigation areas and challenges in *Responsible Operations*:

- Shared Methods and Data
  - Shared Development and Distribution of Methods
- Sustaining Interprofessional and Interdisciplinary Collaboration

TABLE 9. INFRASTRUCTURE AND APPLICATION RECOMMENDATIONS MAPPED TO ELEMENTS OF THE LIBRARY OF CONGRESS'S DIGITAL STRATEGY.

Digital Strategies	Recommendations on Infrastructure	Recommendations on ML Applications
maximizing use of content	✓	
supporting emerging styles of research	✓	✓
welcoming other voices	✓	✓
driving momentum in our communities	✓	✓
cultivating an innovation culture	✓	✓
ensuring enduring access to content	✓	
building toward the horizon	✓	✓

TABLE 10. RECOMMENDATIONS MAPPED TO AREAS OF INVESTIGATION AND CHALLENGE AREAS OUTLINED IN PADILLA'S RESPONSIBLE OPERATIONS.

Strategies	Sub-Strategies	Statement of Vision	Roadmap of ML	Ground-Truthing & Benchmarking	ML + Crowd-sourcing Efforts	Sponsoring Challenges	Research Partnerships
Committing to Responsible Operations	Managing Bias	✓	✓	✓		✓	
	Transparency, Explainability, Accountability	✓	✓			✓	
	Distributed Data Science Fluency	✓	✓				
Workforce Development	Investigating Core Competencies		✓		✓		
	Committing to Internal Talent		✓				
Description & Discovery	Enhancing Description at Scale			✓	✓	✓	
Shared Methods and Data	Shared Development and Distribution of Training Data			✓	✓		
	Shared Development and Distribution of Methods					✓	✓
Sustaining Interprofessional & Interdisciplinary Collaboration						✓	✓

9 CONCLUSION

There is rich potential for machine learning to augment the description and accessibility of materials in the Library of Congress, to inform understanding of collections and choices about how materials are processed and by whom, and to address issues of scale. The Library of Congress is in a remarkable position to advance machine learning for cultural heritage organizations, through its size, the diversity of its collections, and its commitment to digital strategy. This demonstration project—via its explorations, discussion, and recommendations—has shown the potential of machine learning toward a variety of goals and use cases, and it has argued that the technology itself will not be the hardest part of this work. The hardest part will be the myriad challenges to undertaking this work in ways that are socially and culturally responsible, while also upholding responsibility to make the Library's materials available in timely and accessible ways.

BIBLIOGRAPHY

Afzal, Muhammad Zeshan, Andreas Kölsch, Sheraz Ahmed, and Marcus Liwicki. "Cutting the Error by Half: Investigation of Very Deep CNN and Advanced Training Strategies for Document Image Classification." In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 01:883–88, 2017. https://doi.org/10.1109/ICDAR.2017.149.

Ares Oliveira, Sofia, Benoit Seguin, and Frederic Kaplan. "dhSegment: A Generic Deep-Learning Approach for Document Segmentation." In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 7–12, 2018. https://doi.org/10.1109/ICFHR-2018.2018.00011.

Ferriter, Meghan, et al. Kick-off meeting for "Digital Libraries, Intelligent Data Analytics, and Augmented Description." Washington, DC, July 16, 2019.

Harley, Adam W., Alex Ufkes, and Konstantinos G. Derpanis. "Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval." In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 991–95, 2015. https://doi.org/10.1109/ICDAR.2015.7333910.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition," 770–78, 2016. http://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Kong, Tao, Anbang Yao, Yurong Chen, and Fuchun Sun. "HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection," 845–53, 2016. https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Kong_HyperNet_Towards_Accurate_CVPR_2016_paper.html.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems 25*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, 1097–1105. Curran Associates, Inc., 2012. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

Library of Congress. "Digital Strategy for the Library of Congress." 2019. https://www.loc.gov/digital-strategy/

Lorang, Elizabeth, Leen-Kiat Soh, Yi Liu, Chulwoo Pack, and Delaram Rahimi, "Using Chronicling America's Images to Explore Digitized Historic Newspapers & Imagine Alternative Futures," National Digital Newspaper Program Annual Meeting, Washington, D.C., September 26, 2018. https://digitalcommons.unl.edu/library_talks/143/

Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing Data Using T-SNE." *Journal of Machine Learning Research* 9, no. Nov (2008): 2579–2605.

Maringanti, Harish, Dhanushka Samarakoon, and Bohan Zhu, "Machine Learning Meets Library Archives: Image Analysis to Generate Descriptive Metadata." 2019. https://www.lyrasis.org/Leadership/Documents/Catalyst%20Fund/UU-version2-MachineLearning-CatalystFund-WhitePaper.pdf

Odena, Augustus, Vincent Dumoulin, and Chris Olah. "Deconvolution and Checkerboard Artifacts." *Distill* 1, no. 10 (October 17, 2016): e3. https://doi.org/10.23915/distill.00003.

Padilla, Thomas. *Responsible Operations: Data Science, Machine Learning, and AI in Libraries*. Dublin, OH: OCLC Research. 2019. https://doi.org/10.25333/xk7z-9g97.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation." In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, edited by Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, 234–41. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015. https://doi.org/10.1007/978-3-319-24574-4_28.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge." International Journal of Computer Vision, 2015.

Seguin, Benoit and Sofia Ares Oliveira. dhSegment [computer program]. https://github.com/dhlab-epfl/dhSegment.

Simonyan, Karen, and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *ArXiv:1409.1556 [Cs]*, April 10, 2015. http://arxiv.org/abs/1409.1556.

Xie, Saining, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. "Aggregated Residual Transformations for Deep Neural Networks," 1492–1500, 2017. http://openaccess.thecvf.com/content_cvpr_2017/html/Xie_Aggregated_Residual_Transformations_CVPR_2017_paper.html.

Zhou, Xinyu, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. "EAST: An Efficient and Accurate Scene Text Detector," 5551–60, 2017. http://openaccess.thecvf.com/content_cvpr_2017/html/Zhou_EAST_An_Efficient_CVPR_2017_paper.html

APPENDIX 1: WORK-IN-PROGRESS REPORTS

Progress report

Yi Liu

Current Stage

Beyond Words collection encourages users to fix segmentation issue, classify categories of snippets, and transcribe caption of graphic images on newspaper pages. This process fits our first proposed project. Hence, in the first stage, we want to extract graphic content from newspaper pages based on classified data on Beyond Words. And metadata can be generated based on the retrieved corpus. According to downloaded data on Beyond Words, there are approximately 1,500 pages can be used as ground truth for training. However, there are several issues. First, there are missed graphic snippets on newspaper pages. For example, there are pages of which only one out of three graphic snippets are classified in the downloaded ground truth. The category and transcription information of rest two graphic snippets are missing. Second, the segmentation of the snippet uses a simple rectangle, which causes inaccuracy of segment information. For example, two-segment regions are overlapping because the shape of the graphic snippet is not rectangle-shaped. At this stage, we plan to ignore these issues for now. Further attempts will be applied after observation of the reaction of the model to the ground truth extracted from Beyond Words.

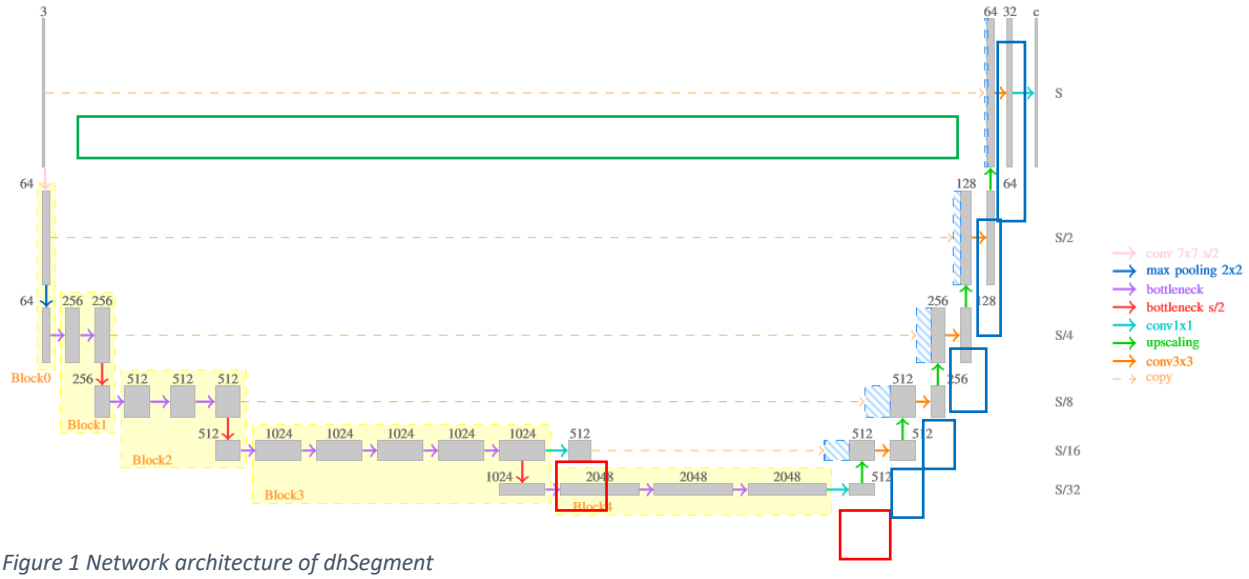


Figure 1 Network architecture of dhSegment

State-of-Art

dhSegment [Sofia et al. 2018] showed a promising result on a segmentation task for European newspapers using a Fully Convolution Network (FCN). dhSegment builds the FCN, shown in Figure 1, by combining ResNet-50 [He et al. 2015] and U-Net [Ronneberger et al. 2015] models. In addition, the dhSegment was not trained from scratch. The encoder part (ResNet-50) of dhSegment classifier was transfer learned from the pre-trained Resnet-50 model for ImageNet. In the implementation detail of dhSegment, there were three differences compare to original ResNet, shown in Figure 2, and U-Net, shown in Figure 3. First, comparing to original ResNet, dhSegment added one convolutional layer after the third residual block and the fourth residual block, shown in red rectangles in Figure 1. The purpose of the change was to decrease

the number of parameters and reduce memory usage. Second, comparing to original U-Net, dhSegment used only one 3x3 convolutional layer in each deconvolution stage, shown in blue rectangles in Figure 1 and 3, while the original U-Net used two 3x3 convolutional layers in each deconvolution stage. This change could result in a faster training speed since numbers of parameters were reduced. However, there was no detailed justification in [Sofia et al. 2018]. Third, ResNet had one more convolution stage than U-Net. Hence, there was an additional bridged deconvolution stage in dhSegment, shown in a green rectangle in Figure 1.

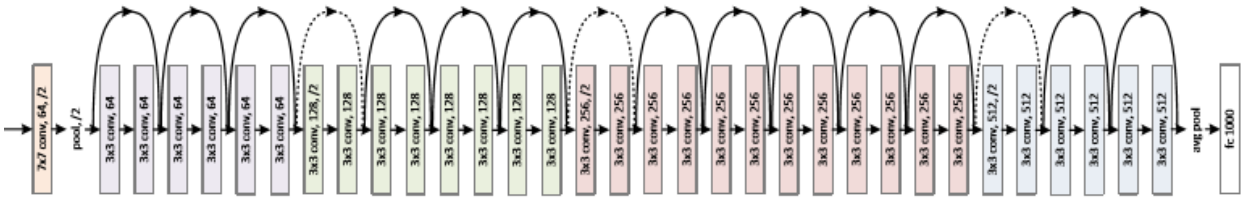


Figure 2 Network architecture of ResNet

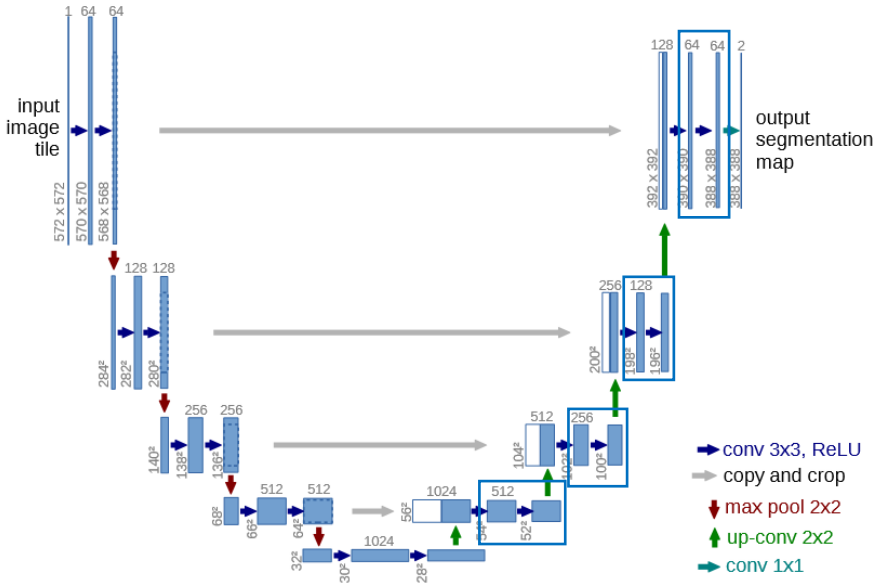


Figure 3 Network architecture of U-Net

ResNeXt [Xie et al. 2017] is the current state-of-art in ImageNet competition. Comparing to ResNet, ResNext used grouped convolution (i.e. side-by-side convolution layers) in each residual block, shown in Figure 4. The usage of grouped convolution was first mentioned in AlexNet [Krizhevsky et al. 2012]. The creation of the grouped convolution was for training models on multiple processor cores. By applying grouped convolution in residual blocks, ResNext showed there were improvements on ImageNet dataset, shown in Figure, 5.

EAST [Zhou et al. 2017] is a text detection approach for scene images. EAST combined HyperNet [Kong 2016] and U-Net to detect accurate text region in scene images. In addition, EAST is a text orientation agnostic approach, meaning East can detect tilted text regions. Further, scene images such as the photograph, are considered graphic images. In Beyond Words collection, figures/illustrations are snippets of a graphic region. Hence, EAST text detection applies to Beyond Words collection to extract texts in the

figure/illustration. An example, in Figure 6, showed the performance of EAST on one image from Beyond Words collection.

HyperNet is originally proposed for object detection. First, it inherited pre-trained AlexNet to extract feature maps. Second, a region-of-interest (ROI) pooling was applied to localize object. Third, a region refinement was applied to refine ROI. And, finally, two consecutive fully connected layers were applied to classify ROI found previously.

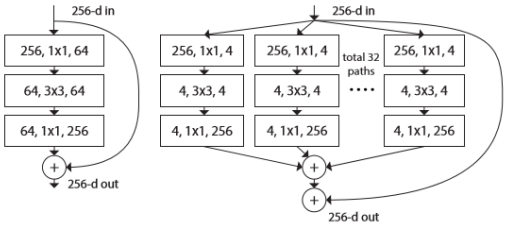


Figure 4 ***(Left)*** *Residual blocks of ResNet.* ***(Right)*** *Residual blocks of ResNext.*

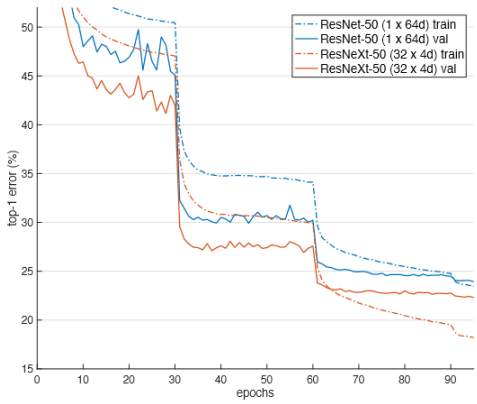


Figure 5 Comparative results between ResNet and ResNeXt on ImageNet-1K dataset.

Figure 6 EAST text detection on Beyond Words snippet. Blue rectangles indicate detected text regions.

Proposing Approach

A two-step approach is proposed at this stage. The first step, an FCN (U-NeXt) combining ResNeXt and U-Net will be built and trained to segment and classify graphic snippets on newspaper pages based on ground truth extracted from Beyond Words. Besides, the training of the FCN will be based on pre-trained ResNeXt model for ImageNet to reduce training parameters. Based on dhSegment, using transfer learning is able to boost training effectiveness, and preserve a good performance. The second step, a text segmentation, and recognition model will be built to retrieve textual content in the graphic snippets (i.e. extracted graphic snippets from the first step). Hence, EAST text detection will be applied to find text regions for an OCR process to retrieve words within graphic snippets. Finally, retrieved words will be encoded into metadata for further usages, such as search queries.

Current Progress

The implementation of the U-NeXt uses MXNet framework has been finished and tested. Currently, a transfer learning process is constructing for further test. The model architecture graph is shown in Appendix I.

The model is training on HCC (UNL resource) server for now. If the AWS in the Library of Congress became a preferred process location, we can move on to the AWS later.

Potential Problem

The major concern is the quality of the ground truth from Beyond Words. We noticed some graphic snippets appeared on the page are missing in the ground truth. Since machine learning models will try to find all graphic content within the input page. Such missing graphic snippets can confuse the model during the training process. Hence, data from Beyond Words may not be able to use directly as training data before fixing of the quality issue. We may try to use an existing European newspaper collection to train the model, then use Beyond Words data for fine-tuning.

Reference

[1] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[2] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

[3] Kong, T., Yao, A., Chen, Y., & Sun, F. (2016). Hypernet: Towards accurate region proposal generation and joint object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 845-853).

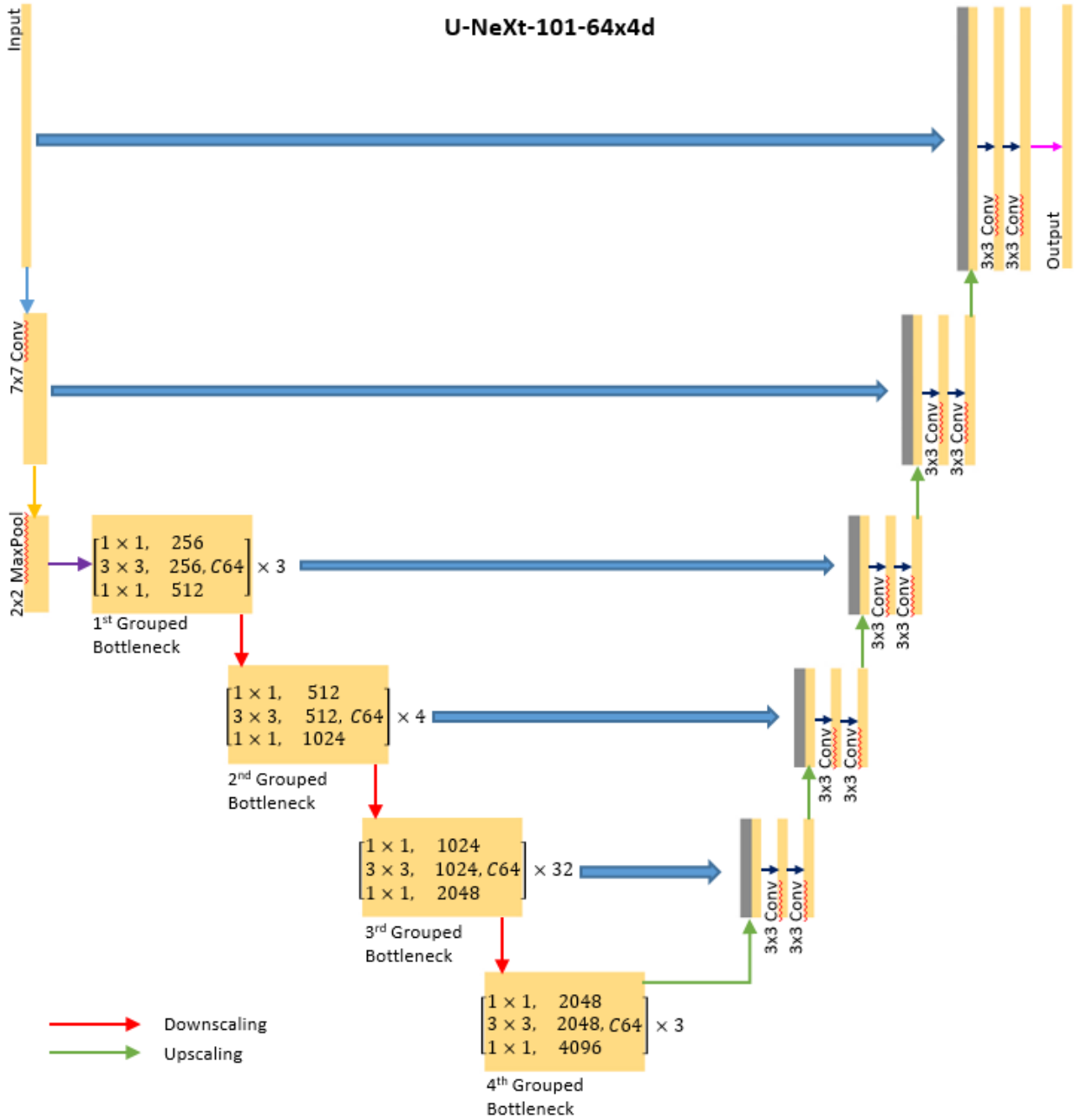
[4] Oliveira, S. A., Seguin, B., & Kaplan, F. (2018, August). dhSegment: A generic deep-learning approach for document segmentation. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)* (pp. 7-12). IEEE.

[5] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham.

[6] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1492-1500).

[7] Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., & Liang, J. (2017). EAST: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 5551-5560).

Appendix I


U-NeXt-101-64x4d
Input
7x7 Conv
2x2 MaxPool
1 × 1, 256
3 × 3, 256, C64
1 × 1, 512
× 3
1st Grouped Bottleneck
1 × 1, 512
3 × 3, 512, C64
1 × 1, 1024
× 4
2nd Grouped Bottleneck
1 × 1, 1024
3 × 3, 1024, C64
1 × 1, 2048
× 32
3rd Grouped Bottleneck
1 × 1, 2048
3 × 3, 2048, C64
1 × 1, 4096
× 3
4th Grouped Bottleneck
3x3 Conv
Output
Downscaling
Upscaling

Progress Report

07/31/2019
Mike Pack

Background

Based on the discussion in the kick-off meeting, there are two main tasks I am currently working on:

1. Page segmentation: Aims to identify image-like components—such as cartoons, illustrations, photographs, and maps—from Chronicling America corpus.
    a. *dhSegment* is known to be the state-of-the-art page segmentation algorithm in literature (https://arxiv.org/abs/1804.10371)
    b. The concept of this model is to combine two deep learning models—*ResNet-50* and *U-net*—which are known to be the best model for image classification and pixelwise-classification problem.
    c. Open-source code is provided by the author in GitHub (https://github.com/dhlab-epfl/dhSegment)
2. Metadata generation: Explore various approaches to find what will be the best way to build a well-structured metadata for image-like components.

Discussion of achievements

1. Page segmentation
   1.1. As a pilot experiment, we were able to train *dhSegment* using a small subset (30 images) of **European Historical Newspaper Dataset** (ENP); and obtained a **promising result** (Please see Figure 1).
      1.1.1. We have also confirmed that the model trained on ENP dataset is also capable of separating images from **Chronicling America corpus** into background, text, and figure sub-regions (Please see Figure 2).
   1.2. We have explored the **Beyond Words** JSON data to analyze and construct a well-formed ground-truth set for training the model.
      1.2.1. A script is implemented. This script converts a single JSON file into a number of XML files equal to the number of actual newspaper pages presented in the JSON file. Note that we are using XML format the **PAGE XML** format is known to be a standard format for newspaper segmentation competition (https://www.primaresearch.org/tools/PAGELibraries).
   1.3. We have confirmed that the training result of the model trained on the Beyond Words dataset is **not promising enough** compared to the model trained on ENP dataset.
      1.3.1. More detailed discussion about this result is described in the following "Discussion of problems" section.

2. Metadata-generation
   2.1. I have not started on this task yet, however, I have shared one idea in our shared folder (https://docs.google.com/document/d/1H0oIUh76_QXslCs_PPvf0lV56zJUot3tza0LjfKdG9U/edit).

Discussion of problems

There are three following main concerns in Beyond Words ground-truth dataset that might cause a model to be hindrance during the training:

1. Inconsistency
   1.1. Not all figure entities presented in a page are annotated. More detailed information is described in Figure 3.
2. Imprecision
   2.1. Most of the time, a simple rectangle annotation contains regions that are not relevant to the corresponding class. More detailed information is described in Figure 4.
3. Data imbalance
   3.1. In the JSON file, the class of most of the figure entities is "Photograph." With the imbalanced dataset, a model can be biased to learn a set of features relevant to the majority class during the training. More detailed information is described in Figure 5.

Discussion of work that lies ahead

1. Segmentation
   1.1. Training model with Beyond Words dataset to address data imbalance problem
   1.2. Training model with enlarged ENP dataset
2. Meta-data generation
   2.1. Explore techniques to generate meta-data relevant to image quality
   2.2. Explore techniques to generate meta-data relevant to image context

Figures

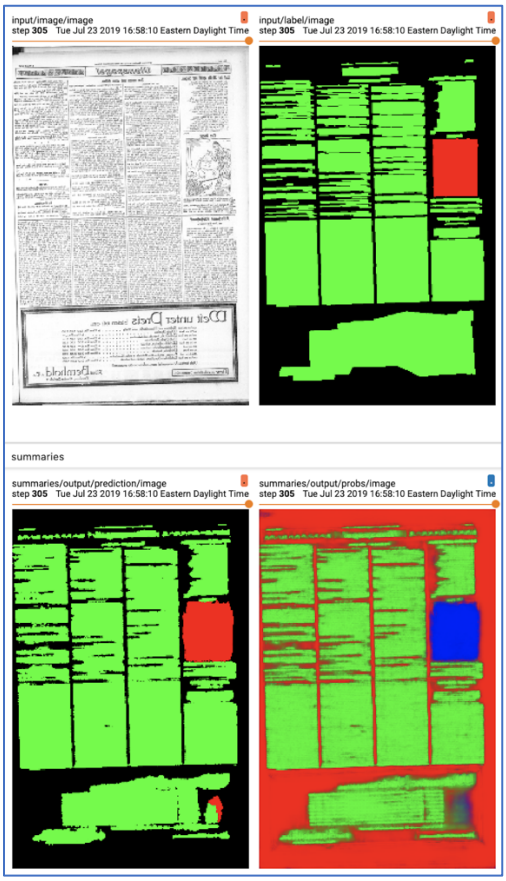


Figure 1. Visual inspection on the segmentation result of model trained on ENP dataset. Clockwise from top-left: (1) Input, (2) ground-truth, (3) probability map, and (4) prediction. In ground-truth, each pixel is labeled as following: (1) black=background, (2) green=text, and (3) red=figure. The probability map here shows the model's pixel-wise prediction value, for example each pixel will have a list of probability values, such as [background:0.2, text:0.7, figure:0.1]. The prediction map is a thresholded result from the probability map, using the arguments of the maxima (i.e., argmax), for example, argmax[background:0.2, text:0.7, figure:0.1]= text:0.7. The color representation of the probability map is the same as the ground-truth.

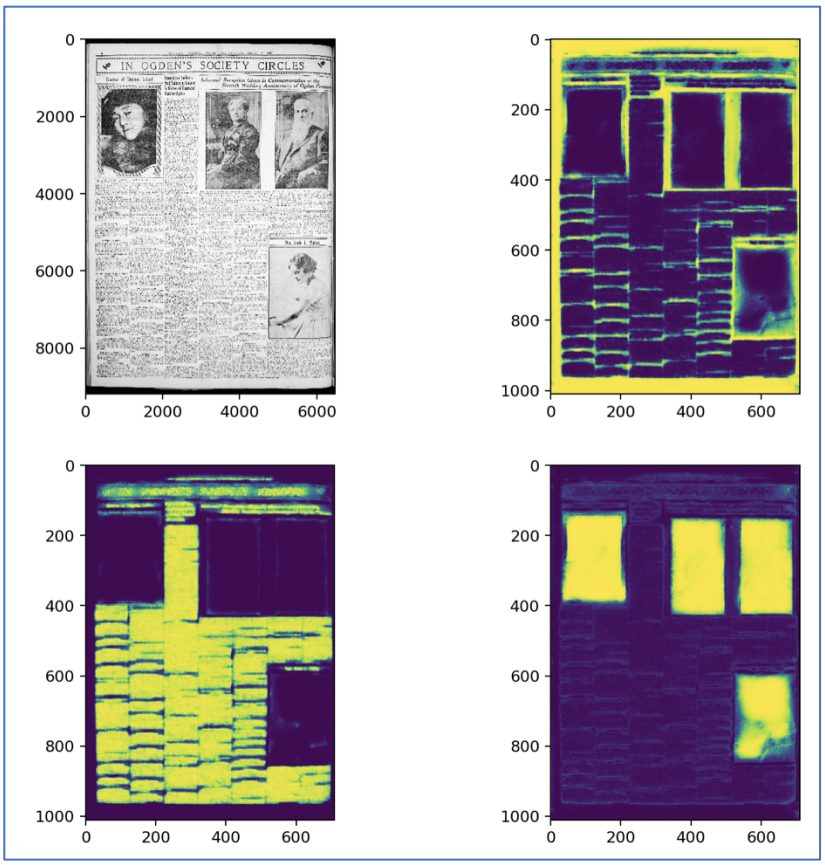


Figure 2. Visual inspection on the segmentation result of model trained on ENP dataset. Note the image shown here is from the Chronicling America corpus, which is never shown to the model during the training. Clockwise from top-left: (1) Input, (2) background-map, (3) image-map, and (4) text-map. In each map, brighter (yellow-ish) region indicates the region of interest with high probability.

Figure 3. Visual inspection on the segmentation result of model trained on Beyond Words dataset. Clockwise from top-left: (1) Input, (2) ground-truth, and (3) prediction. Note here that model makes a reasonable guess that there are multiple figure-like regions in a given page, but the inaccurate ground-truth missing some figure-like regions penalize the model's prediction, which is problematic since it will confuse the model to learn a set of useful features.

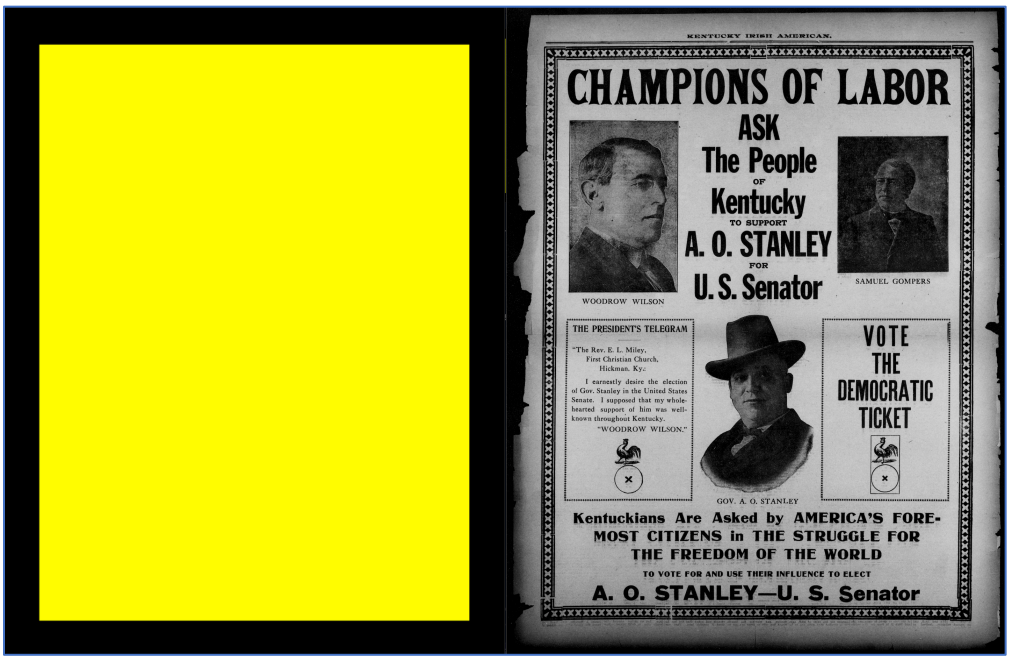


Figure 4. From left to right: (1) ground-truth (yellow: Photograph and black: background) and (2) original image. Note here that in the ground-truth, non-photograph-like (e.g., texts) components are included within the yellow rectangle region. The best-case scenario is to have a more accurate annotation with polygon so that each ground-truth entity can contain only photograph-like pixels.

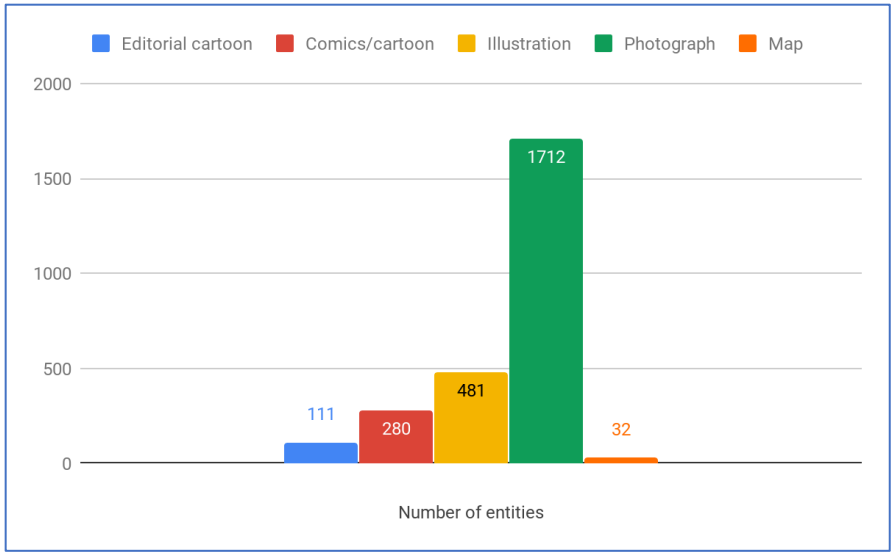


Figure 5. Number of entities in the Beyond Words JSON file. Note here that the dataset is overwhelmed with photograph class (4%, 11%, 18%, 65%, and 1%).

Progress Report

08/05/2019
Mike Pack

Background

1. Page segmentation: Aims to identify image-like components—such as cartoons, illustrations, photographs, and maps—from Chronicling America corpus by using state-of-the-art deep learning model.
2. Metadata generation: Aims to build a metadata generator for image-like components.

Discussion of achievements

1. Page segmentation
   1.1. Using different training configurations, different models are trained using two datasets: (1) Beyond Words and (2) European historical NewsPapers.

Model	train/eval size	Classes	Weighted training	Pre-processing (Normalization)	Best Score	
					Accuracy	mIoU
BW_1500_v1	1226/306	0: Background 1: Editorial cartoon 2: Comics/cartoon 3: Illustration 4: Photograph 5: Map	No	No	0.87	0.24
BW_1500_v2			Yes [10;22;20;18;8;22]		0.88	0.26
ENP_500_v1	385/96	0: Background 1: Text 2: Figure 3: Separator 4: Table	Yes [5;10;40;10;35]	No	0.88	0.64
ENP_500_v2				Yes	0.89	0.64
ENP_500_v3			**No**	**No**	**0.91**	**0.69**
ENP_500_v4				**Yes**	**0.91**	**0.69**

*Accuracy: Pixel-wise accuracy.
*mIoU: Average intersection over union.
*Normalization: Zero mean unit variance

   1.2. Note that **models trained with ENP dataset** show better segmentation performance than with BW dataset in both accuracy and mIoU.
      1.2.1. High accuracy of BW_1500_v1 and v2 is not a surprising result. Since most of each image has only a few regions of interest, so from a model's point of view, assuming and predicting most of pixels to be a background is guaranteed to obtain high accuracy. The models' this sort of behavior can be verified based on their poor performance on mIoU metric, which quantifies the *percent overlap* between the target class and model's prediction.
      1.2.2. On contrast, high accuracy of from ENP_500_v1 to v5 is a meaningful result. Since text regions are included in the ground-truth, and thus a model's simple guessing that everything is background will get penalized. Also, we can see relatively high mIoU measures.
   1.3. Note that pre-processing does not play a significant role in improving segmentation performance.
   1.4. Note that weighted training causes a performance degradation.
   1.5. Post-processing has been implemented
      1.5.1. Eliminate small regions
      1.5.2. Draw bounding-box or polygon
   1.6. Actual testing on some Chronicling America images are shown in Figure 1 to 4.

2. Metadata-generation
    2.1. Approach 1: Google Cloud Platform (GCP) Vision API
        2.1.1. GCP Vision API provides a set of pre-trained machine learning models that can assign labels to images and quickly classify them into a number of predefined categories. For example, we can utilize their (1) object detection, (2) face recognition, (3) read printed and handwritten text, (4) similar image recommendation, or (5) basic image property generation (e.g., color space).
        2.1.2. For the visual demonstration, see Figure 5.
    2.2. Approach 2: Explore Automatic Image Annotation (AIA) research field and find the best model that fits our dataset.
        2.2.1. Since my work has mainly focused on page segmentation and GCP Vision API last week, I need more time to work on this.

Discussion of problems

1. Page segmentation
    1.1. As can be seen in Figure 2 to 4, there are some false-positive and false-negative results.
        1.1.1. We might improve the performance of our model with (1) more advanced data augmentation, (2) enlarged data set, (3) hyperparameter tuning, and (4) modifying architecture.
2. Meta-data generation
    2.1. Approach 1: GCP Vision API
        2.1.1. One thing in my mind is that the resultant metadata would be not that useful or end up with just entertaining result as my previous sentence generation idea. Since most of figures in newspapers are people, so most of the time, the GCP Vision API will label images to "person" or "people" as shown in Figure 5.
        2.1.2. For a large-scale data, there is a monthly usage.

Discussion of work that lies ahead

1. Segmentation
    ~~1.1. Training model with Beyond Words dataset to address data imbalance problem~~
    ~~1.2. Training model with enlarged ENP dataset~~
    1.3. Data augmentation
    1.4. Hyperparameter tuning
2. Meta-data generation
    2.1. Explore techniques to generate meta-data relevant to image quality
    ~~2.2. Explore techniques to generate meta-data relevant to image context~~
    2.3. Explore state-of-the-art methods in AIA field

Figures

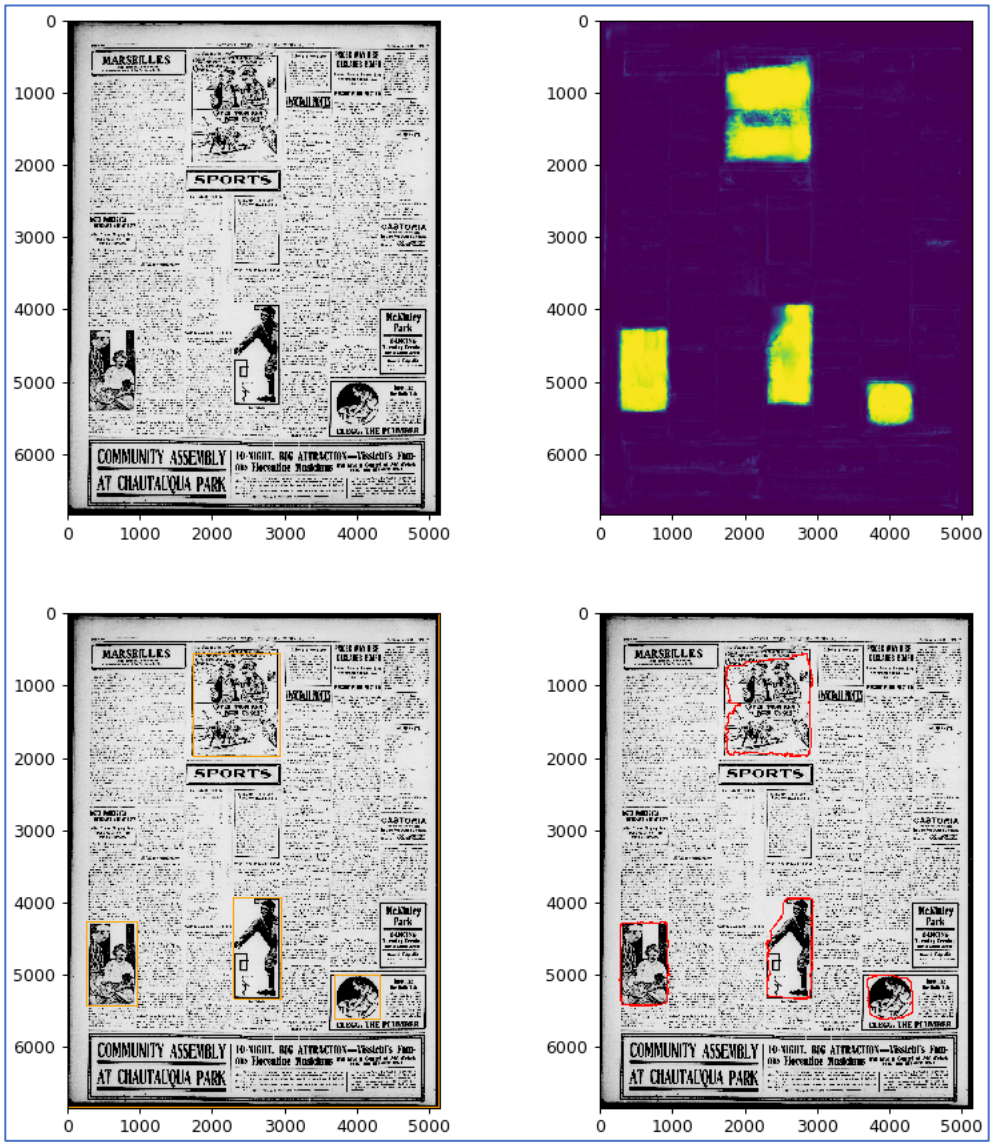


Figure 1. Segmentation result of ENP_500_v4 on Chronicling America image (sn92053240-19190805.jpg). Clockwise from top-left: (1) Input, (2) probability map for figure class, (3) detected figures in polygon, and (4) detected figures in bounding-box. In the probability map, pixels with higher probability to belong to figure class are shown with brighter color.

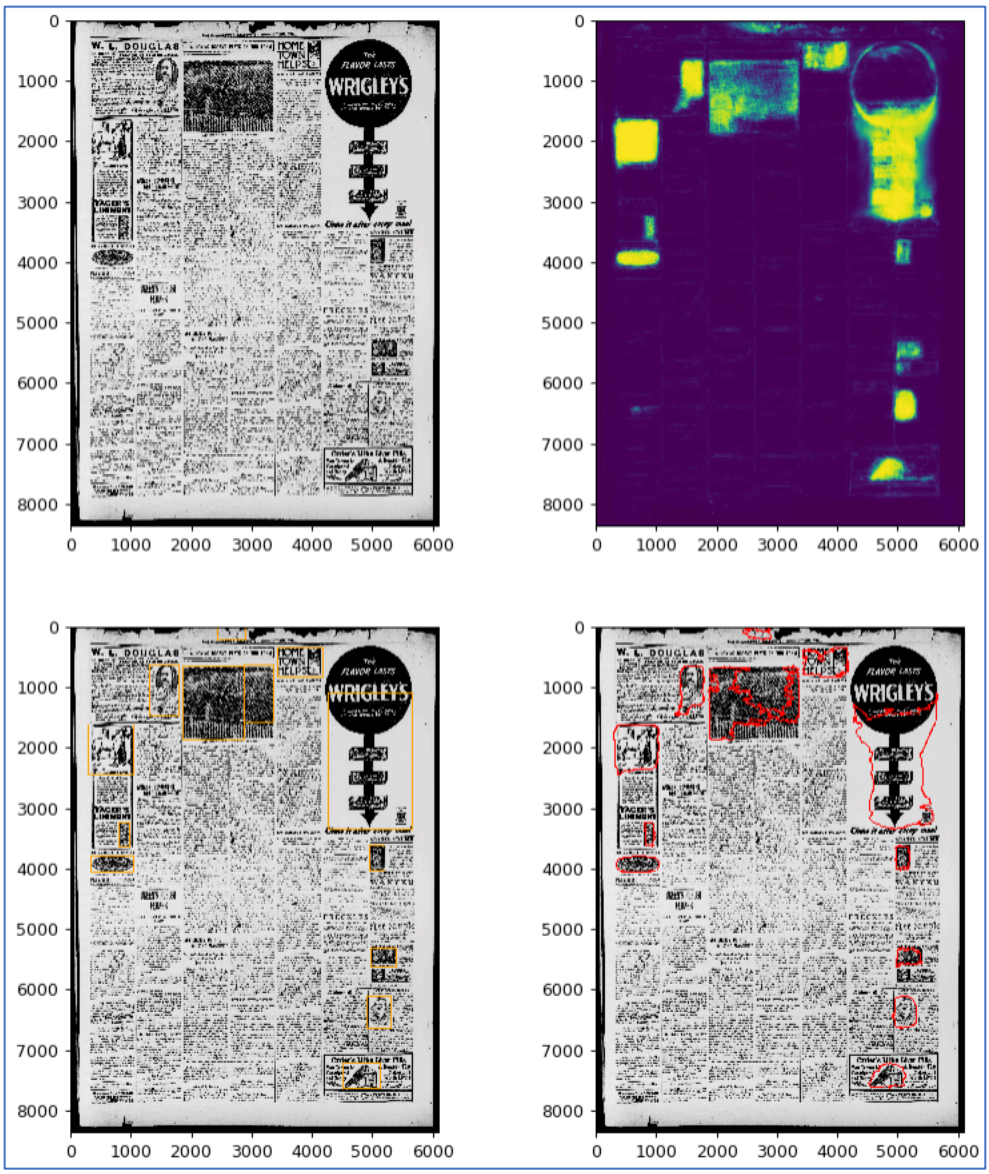


Figure 2. Segmentation result of ENP_500_v4 on Chronicling America image (ndnp-jpeg-surrogates_deu_descendo_ver01_data_sn84026820_00271765095_1917050501_0153.jpg). Clockwise from top-left: (1) Input, (2) probability map for figure class, (3) detected figures in polygon, and (4) detected figures in bounding-box. In the probability map, pixels with higher probability to belong to figure class are shown with brighter color.

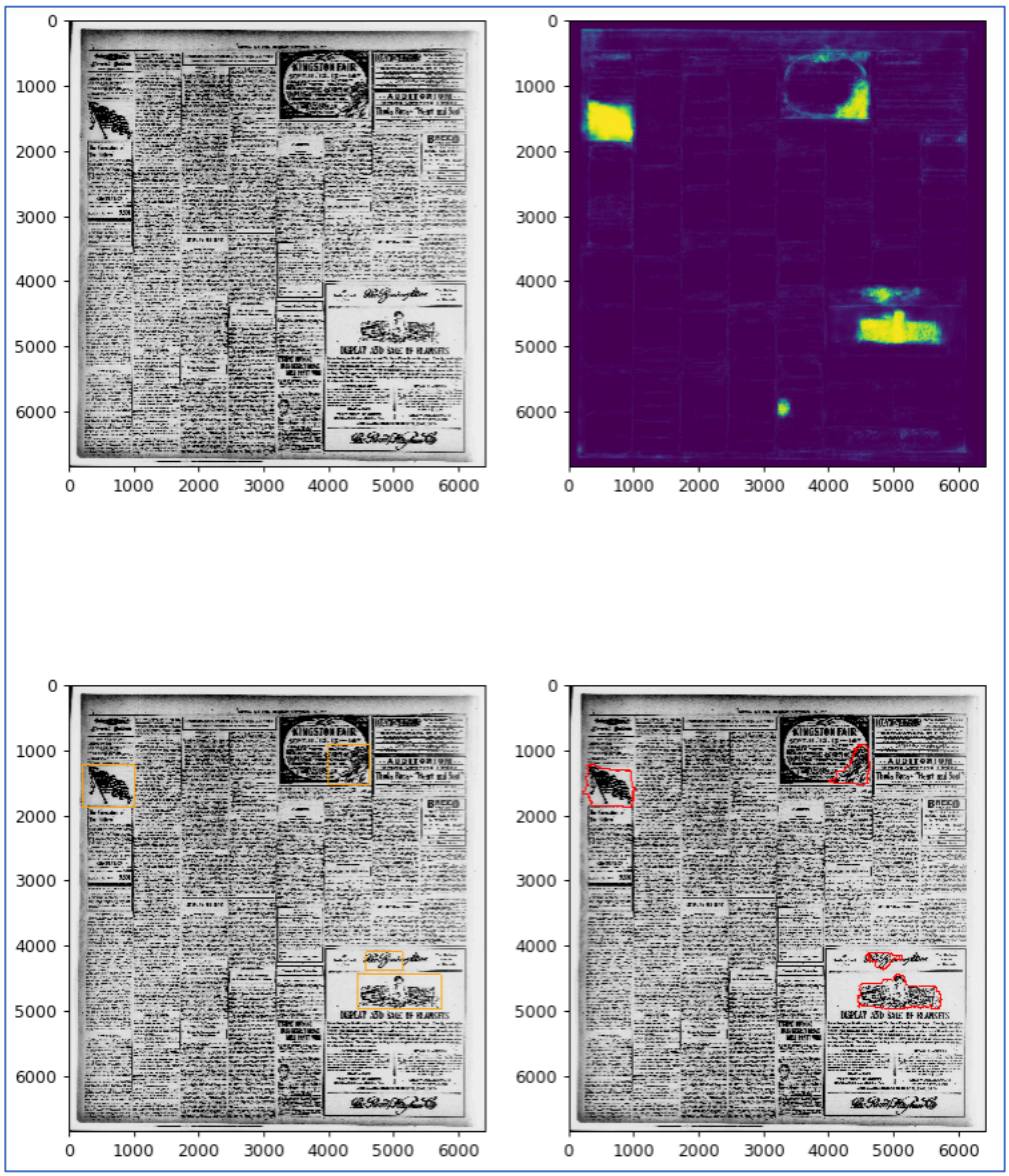


Figure 3. Segmentation result of ENP_500_v4 on Chronicling America image (ndnp-jpeg-surrogates_ct_berlin_ver01_data_sn82014086_00295866135_1917091301_0116.jpg). Clockwise from top-left: (1) Input, (2) probability map for figure class, (3) detected figures in polygon, and (4) detected figures in bounding-box. In the probability map, pixels with higher probability to belong to figure class are shown with brighter color.

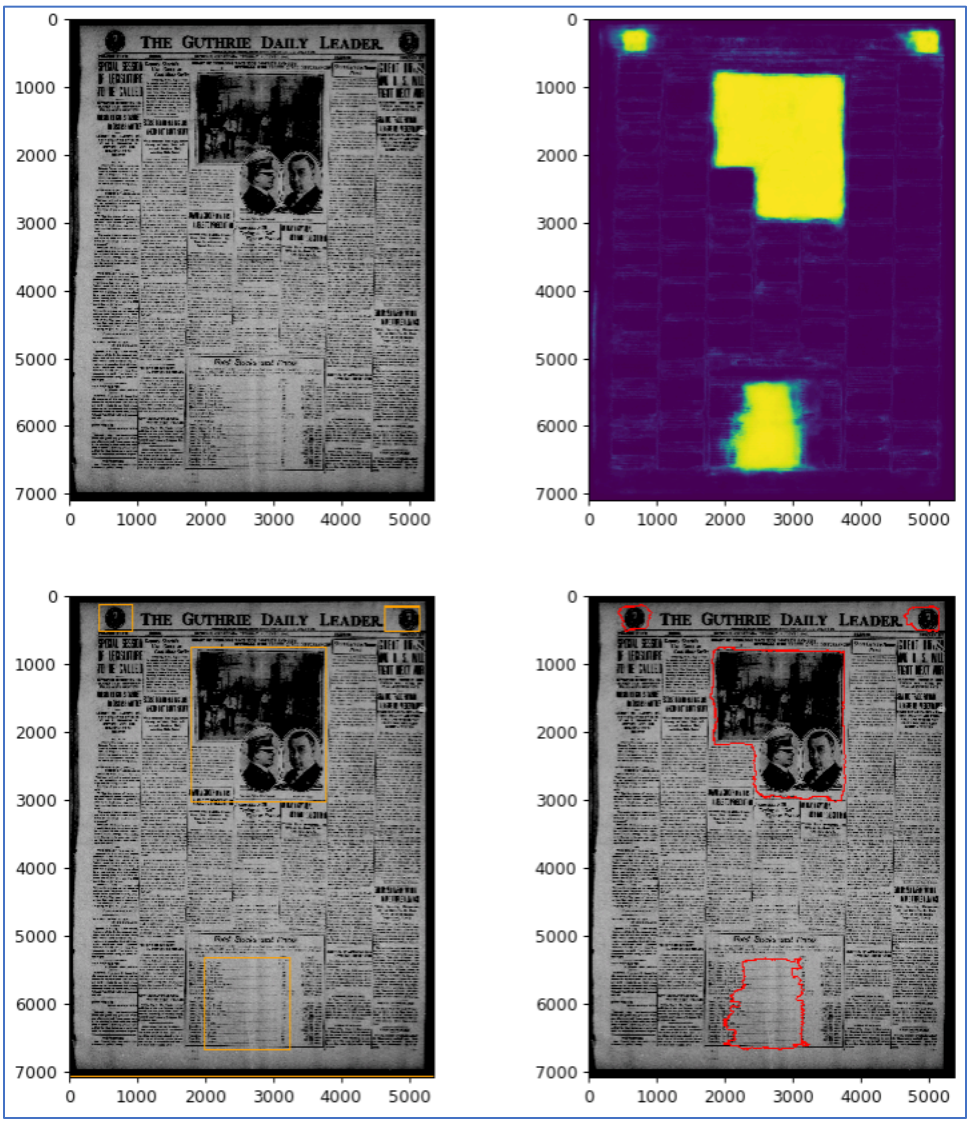


Figure 4. Segmentation result of ENP_500_v4 on Chronicling America image (sn86063952-19190805.jpg). Clockwise from top-left: (1) Input, (2) probability map for figure class, (3) detected figures in polygon, and (4) detected figures in bounding-box. In the probability map, pixels with higher probability to belong to figure class are shown with brighter color.

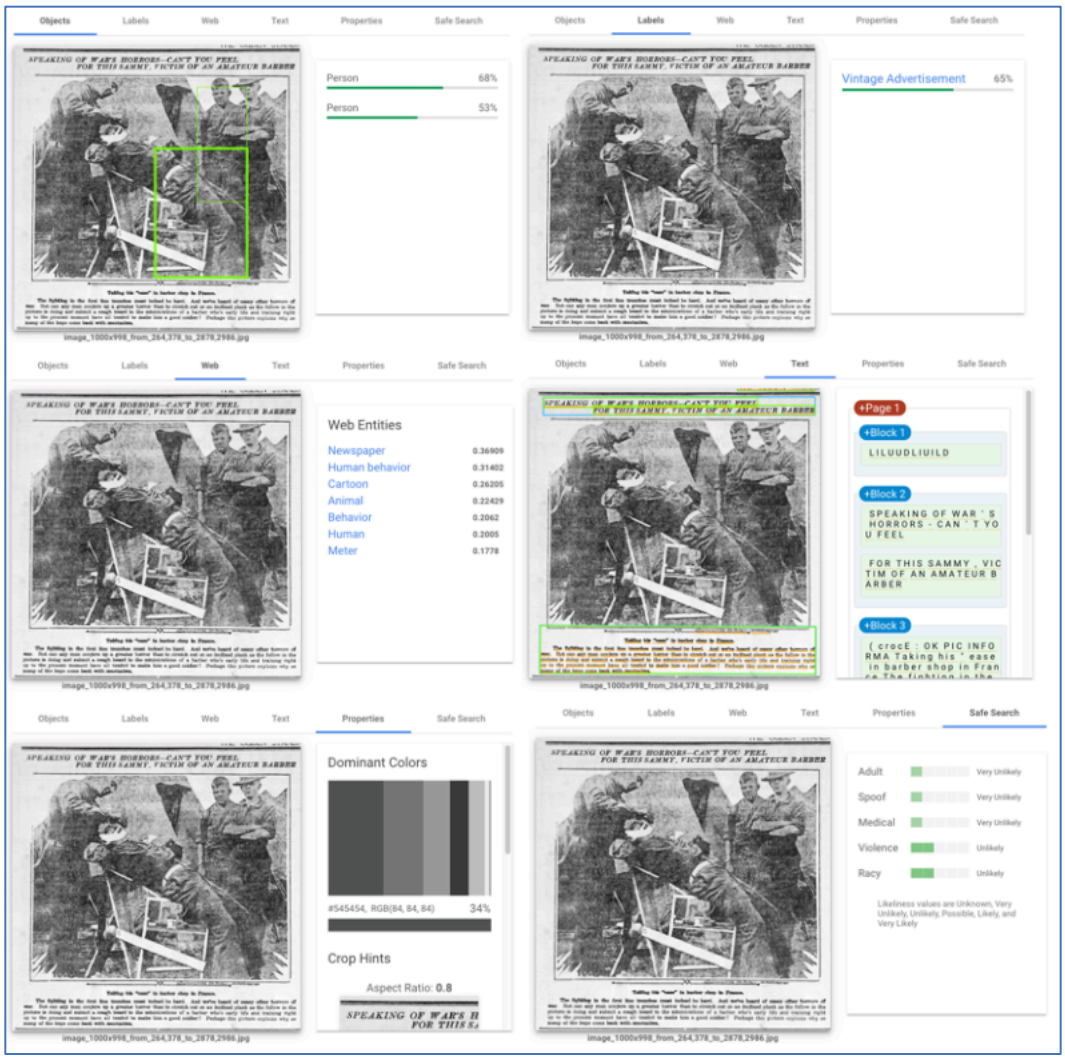


Figure 5. GCP Vision API demonstration on test image from Beyond Words (http://beyondwords.labs.loc.gov/#/view/5aa47eff639da00001002159). Top-left: Result of object detection. Note that two objects (i.e., two people) are detected and marked with green bounding-boxes. Top-right: Result of label assignment. Middle-left: A list of relevant keywords. Middle-right: Result of text recognition. Note that text blocks are marked with green bonding-boxes and words are marked with orange line. Bottom-left: A list of basic property of image (e.g., color and size). Bottom-right: Estimation of the likelihood that given image includes adult content, violence, and etc.

Progress report – Document image quality assessment for digital library collections

Yi Liu

Background

To better access digital library collections in terms of improving searchability, often document images—e.g., digitized or scanned from their original paper versions or microfilms—are tagged with metadata. Typically, metadata comes in two forms: (1) metadata about the document such as the original date of the document, publication venue, and so forth—also known as ancillary data in the realm of image processing; and (2) metadata about the texts directly discerned from the document, either through manual processing or natural language processing, such as keywords found in the texts, or titles extracted from the texts. However, as researchers begin to process large quantities of document images to develop robust classifiers or to develop generalizable automated systems, there is an increasing need for a third form of metadata: (3) metadata about the image quality of the document images such as average intensity of an image, contrast, range effects, layout structure, etc., such that researchers could query and retrieve specific subsets of document images based on these qualities for testing. It is this third form of metadata that motivates this report on document image quality assessment for digital library collections.

In general, image quality assessment includes both machine and human perceptions of quality of an image. For machine perception, quality assessment evaluates difficulties to predict or categorize an image for a machine. And for human perception, quality assessment evaluates difficulties to understand and interpret an image based on the visual appearance for human. In this report, our focus is document images.

In document image quality assessment (DIQA), there are two types of quality metrics [Ye and Doermann 2013]. One is called the *objective* quality metrics that is based on the ability to accurately predict the quality of a document image. For example, an optical character recognition (OCR) accuracy prediction model based on a convolutional neural network (CNN) [Kang et al. 2014] predicts the accuracy of OCR outcome for the document image. The other one is called the *subjective* quality metrics that define document image quality with respect to human perception. For example, a rating-based method assigns a categorical label to each image, such as the mean opinion score (MOS).

Problem Definition with respect to the Chronicling America's Repository

The Chronicling America repository (including Beyond Words and By the People) has little information on the document image quality. Note that there is an OCR accuracy quality score provided in the corresponding "ocr.xml" file. However, the score is not provided for all pages. Only some of the document pages have the corresponding score within its XML file. In addition, the score shows an objective score for OCR accuracy. It cannot intuitively indicate the quality of the image for human perception. Hence, a subjective score system is required to provide more quality information on human perception for further usage. For example, a school teacher might want to find some documents for his or her classroom activities. Intuitively, s/he might want document images with a clean background, good contrast, and less

content density. With a MOS system, the search query could be as easy as searching for document images with a good background, high contrast, and low density metadata.

State-of-the-Art

[Kang et al. 2014] proposed a shallow CNN model to predict the OCR accuracy for document images. The proposed model is shown in Figure 1. Further, they proposed to use a parallel min-max pooling before the dense layer. Such min-max pooling was able to maintain filter responses characterized by max and min values to capture statistical quality information for final score prediction.

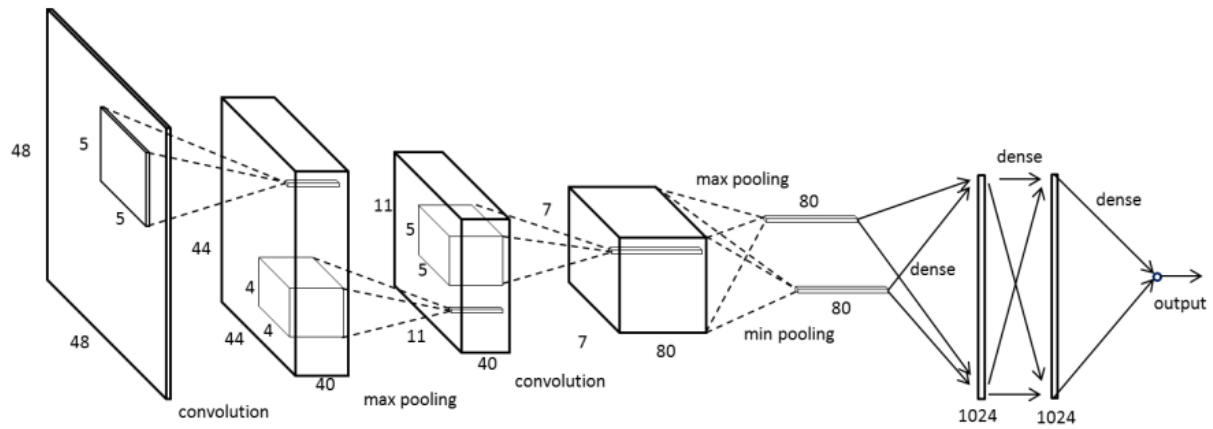


Figure 1 CNN model in [Kang et al. 2014] for OCR accuracy prediction.

MOS is widely used for subjective quality assessment [Ye and Doermann 2013]. However, there are challenges for subjective quality assessment. First, there is no existing human perception-based DIQA database to perform related experiments. Second, degradations could be present at different document levels, such as the character-level, the article-level, or the page-level. The appearance of multiple degradations increases the level of difficulty to design a global measurement. Third, a subjective quality assessment could be task-specific and might not be generalizable, as different tasks could command different values or emphases on how the quality of an image is judged or assessed.

Proposed Approach

Dataset Construction

Machine Learning, especially for deep learning, could require large amounts of labeled data to perform training. However, the lack of human perception-based DIQA database presents a challenge to investigations. We suggest **adding an interface to allow a user to describe the quality of the document images using five-level rating score, such as MOS (i.e., 5-Excellent, 4-Good, 3-Fair, 2-Poor, and 1-Bad), on aspects such as contrast, range-effect, background-cleanness, and content density.** Over time, a human perception based DIQA database could be established to support studies and experiments, and could even be made publicly available for research competition for academia.

Integrating Existing Work

In the work of Image Analysis for Archival Discovery (Aida), an objective DIQA experiment was carried out to evaluate historical newspapers pages from 1834 to 1922 in the Chronicling America repository. The

objective DIQA aimed to evaluate four metrics for the newspaper page across different languages in different eras. These four metrics that could be automatically computed included (1) the skewness, (2) the contrast, (3) the range-effect, and (4) the bleed-though (Examples are shown in Appendix). The results of the experiment were numeric scores ranging based on algorithmically understanding. As these metrics were numeric, it would require human expertise to better interpret the results such as the range of values for an image to be considered of high contrast or low contrast. Furthermore, it would require application-specific needs to leverage these metrics; for example, how high the range-effect would have to be for an image to be rendered not usable or interpretable for a particular application such as natural language processing?

However, this is not necessarily suggesting the existing work is useless for subjective DIQA. With additional works, the existing objective DIQA results of Aida could be helpful. These works include: (1) pre-defining the range of the score that makes sense to human users; e.g., numerical scores on range-effect within 0 to 1 may be considered excellent, within 1 to 2 good, within 3 to 5 fair, within 5 to 6 poor, and finally, larger than 6 bad; and (2) normalizing numeric scores based on the pre-defined range for each metric for subjective DIQA experiments.

Deep Learning-Based Experiment

We propose an inference multi-output U-NeXt to perform a subjective DIQA using MOS. The main architecture of the model is a combination of ResNeXt [Xie et al. 2017] and U-Net [Ronneberger et al. 2015] that is attached by a min-max pooling and two dense layers, shown in Figure 2. Each output corresponds to one aspect of a five-level MOS. Note that the U-NeXt model will not be trained from scratch. A pre-trained model using ImageNet [Russakovsky et al. 2015] and ENP [Clausner et al. 2015] database will be adopted. By using transfer learning, a pre-trained model can help us to reduce numbers of training parameters and to make the training process faster.

In the current stage, we could perform experiments based on the normalized objective DIQA scores from the project Aida. Hence, for each newspaper page from 1834 to 1922 in the Chronicling America's repository, four quality metrics are included in the ground-truth on the skewness, contrast, range-effect, and bleed-through using MOS. Then, based on the ground-truth and data, we can train the U-NeXt model to rate the newspaper page subjectively using MOS. Such configuration would be able to show the strength of the model on subjective DIQA tasks. However, because the subjective score is a pseudo-score based on algorithmic score, they are not necessarily able to accurately represent the actual human perception. Hence, further experiments to evaluate the effectiveness of the subjective DIQA using the U-NeXt requires an actual human perception-based DIQA database would be helpful.

Reference

[1] Clausner, C., Papadopoulos, C., Pletschacher, S., & Antonacopoulos, A. (2015, August). The ENP image and ground truth dataset of historical newspapers. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 931-935). IEEE.

[2] Kang, L., Ye, P., Li, Y., & Doermann, D. (2014, October). A deep learning approach to document image quality assessment. In *2014 IEEE International Conference on Image Processing (ICIP)* (pp. 2570-2574). IEEE.

[3] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham.

[4] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, A. C. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3), 211-252.

[5] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1492-1500).

[6] Ye, P., & Doermann, D. (2013, August). Document image quality assessment: A brief survey. In *2013 12th International Conference on Document Analysis and Recognition* (pp. 723-727). IEEE.

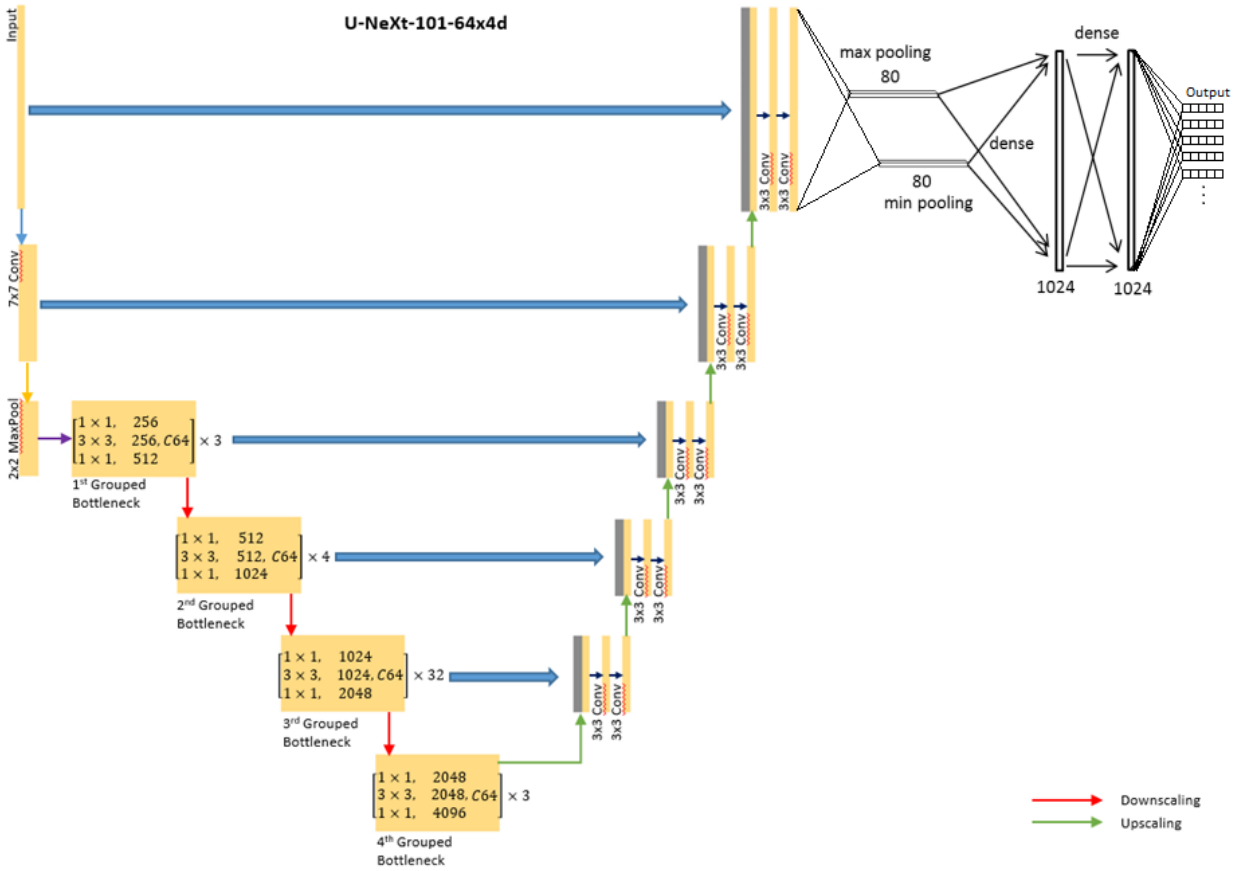


Figure 2 Inference multi-output U-NeXt model for subject DIQA.

Appendix

Table 1 Examples of newspaper pages having different levels of contrast, range-effect, and bleed-through.

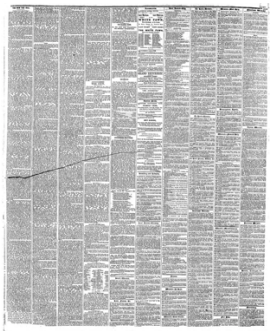
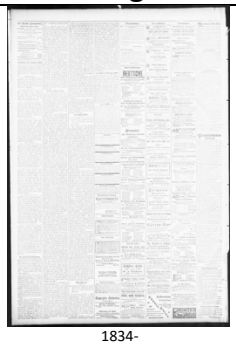
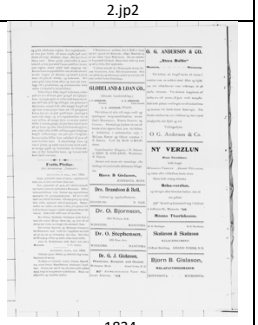
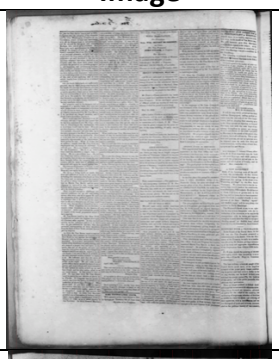
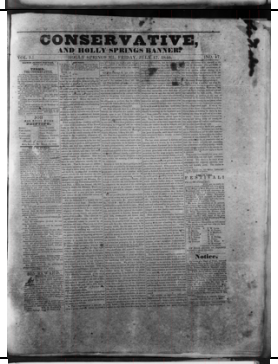
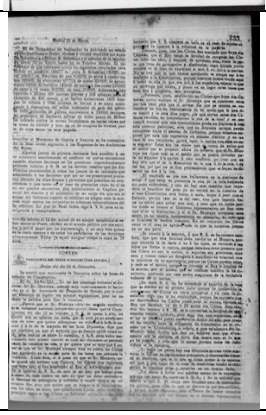
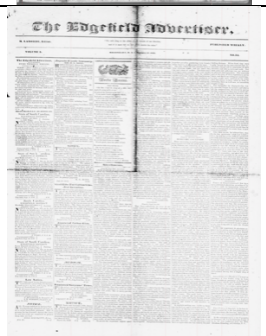
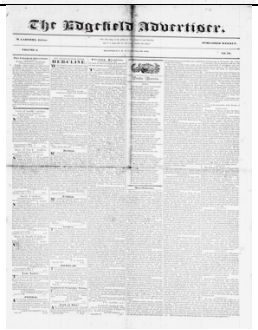
	High/Severe		Some		Low/None	
	Value	**Image**	**Value**	**Image**	**Value**	**Image**
Contrast	146.08	 1834-1922FullPages_20PerYr/1868_English/sn82014064/1868-07-18/ed-1/seq-4.jp2	55.2	1834-1922FullPages_20PerYr/1848_English/sn83035366/1848-03-16/ed-1/seq-1.jp2	3.11	 1834-1922FullPages_20PerYr/1898_German/sn83045081/1898-12-29/ed-1/seq-2.jp2
Range-effect	14.11	1834-1922FullPages_20PerYr/1896_English/sn88083938/1896-04-18/ed-1/seq-1.jp2	4.0	1834-1922FullPages_20PerYr/1867_Spanish/2013201074/1867-02-09/ed-1/seq-4.jp2	0.0	 1834-1922FullPages_20PerYr/1904_Icelandic/sn90060662/1904-12-01/ed-1/seq-12.jp2
Bleed-through	0.129	1834-1922FullPages_20PerYr/1861_Spanish/2013201074/1861-03-16/ed-1/seq-4.jp2	0.033	1834-1922FullPages_20PerYr/1856_English/sn85026050/1856-08-15/ed-1/seq-4.jp2	0.001	1834-1922FullPages_20PerYr/1907_Icelandic/sn90060662/1907-09-01/ed-1/seq-6.jp2

Table 2 Examples of newspaper pages having different levels of skewness.

Skewness			
Title	**Value**	**Image**	**Note**
/Archive/sn83016788_1840-05-26_ed-1_seq-2.jpg	0.0		No skewness
/Archive/sn83016788_1840-07-17_ed-1_seq-1.jpg	-0.5		Slightly tilting to left
/Archive/sn85025180_1837-10-14_ed-1_seq-1.jpg	0.5		Slightly tilting to right
/Archive/2013201074_1837-05-16_ed-1_seq-3.jpg	0.75		Slightly tilting to right

/Archive/2013201074_1837-01-24_ed-1_seq-3.jpg	-1.0		More tilting to left
/Archive/sn84026897_1838-09-27_ed-1_seq-1.jpg	1.0		More tilting to right
/Archive/sn84026897_1838-09-20_ed-1_seq-1.jpg	1.0		More tilting to right
/Archive/sn84026897_1840-05-28_ed-1_seq-3.jpg	2.0		More tilting to right

Progress Report – Document image classification for digital library collections

08/13/2019
Mike Pack

Background

Document image classification aims to classify a type of given document image into a certain category—email, letter, handwritten, etc.—based on its layout and visual structure. A successful document image classification can breakdown and categorize a large-scale digital document repository into a smaller subset, which is beneficial for maintenance, discoverability, etc.

The main challenge of document image classification arises from the fact that within each document type, there exists a wide range of visual variability, as shown in Figure 1. Another issue is that documents of different categories often have substantial visual similarities, as shown in Figure 2.

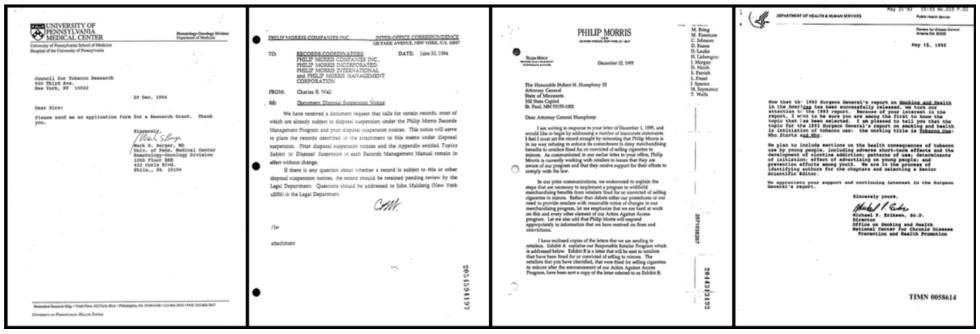


Figure 1. Examples of document images that show a wide range of visual variability within the same type (i.e., a *letter* type in this particular example). Note that no two documents show the exact same spatial arrangement of header, date, address, body, and signature; some of the documents even omit these components entirely.

Figure 2. Examples of document images that show visual similarities across different types (Left: *form*, right: *scientific publication*) Note that two different types of documents share similar spatial arrangement of title and body. Even the amount of contents, so-called density, is also similar.

In the past few years, using a deep convolutional neural network (CNN) to classify images has shown to be able to achieves substantially successful classification performances in various domain, such as natural image classification, natural image segmentation, etc. Inspired by the success of CNNs in other domains,

we would like to propose using the current state-of-the-art CNN model for document image classification problem.

State-of-the-Art

In this section, two papers, which used CNN for document image classification, are briefly reviewed. It is worth noting that all three papers used the same dataset, Ryerson Vision Lab Complex Document Information Processing (RVL_CDIP)[1] [1], and they achieved similar performance, around 90%.

Harley et al. (2015) investigated whether the features extracted from natural images (i.e., ImageNet) are general enough to be applied to document images [1]. The author also proposed a region-based CNN model, which consists of 5 different CNNs where each CNN is designed to be trained on particular regions: (1) holistic, (2) header, (3) footer, (4) left-body, and (4) right-body. Each of those CNN is VGG-16 [2] pre-trained on ImageNet. The dimension of each feature vector extracted from the corresponding CNN is reduced using principal component analysis, and they are concatenated into a single vector for the classification.

There are two interesting findings from their experimental result. First, *the features extracted from a CNN trained on ImageNet are powerful enough to be used for document image classification task that achieves approximately as well as a model fine-tuned on a subset of RVL_CDIP*, so-called *SmallTobacco*, 87.8% and 89.8%, respectively. There are two key implications from this finding. First, what the machine considers as distinctive features in natural images are also distinctive features in document images. Second, since we can easily transfer the knowledge (i.e., a set of filters capable of extracting distinctive features from an image) from one model to the other, we do not need to train our own model from scratch, which would allow us to reduce a significant amount of training time. Second, *given sufficient training data, enforcing region-specific feature-learning is unnecessary; a single CNN trained on entire images performed approximately as well as an ensemble of CNNs trained on specific subregions of document images*, 89.8% and 89.3%, respectively. This finding indicates that in the task of document classification, feeding large amount of data is more important than feeding fine-grained region-dependent representations. This result suggests that putting more efforts on collecting a larger amount of training dataset is *more* important than redesign a model's architecture for capturing a region-specific representation in the document classification.

Muhammad et al. (2017) investigated recent deep CNN architectures (i.e., AlexNet, VGG, GoogLeNet, and ResNet) and strategies for the task of document image classification [3]. Also, the author investigated the impact of transfer learning from a huge set of document images (i.e., RVL_CDIP). The outcome of this study can be summarized in two points as following: (1) VGG-16 performs slightly better than other networks by a small margin of 1-2%, and (2) with regards to the impact of transfer learning, all CNNs pre-trained on RVL_CDIP achieve higher accuracy than both ImageNet and random initialization (i.e., no transfer learning). The first outcome implies that there are no significant performance differences between recent CNN models, which allows one to use a computationally cost-effective model for practical deployment—if that is a concern. The second outcome is not a surprising result, which aligns with [1] in

[1] This dataset consists of 400,000 labeled document grayscale images from 16 classes. The images are sized, so their largest dimension does not exceed 1000 pixels.

that a model pre-trained on ImageNet outperforms a model trained from scratch. Overall, the key implication from this research is that using one of recent CNN model pre-trained on RVL_CDIP is a suitable preset for building our own document classification model.

Proposed Approach

As a first experiment for the task of document image classification, our goal is to build a model capable of distinguishing three different types of documents: (1) handwritten, (2) typed/machine-printed, and (3) mixed (both handwritten and typed). To this end, we propose to use a VGG-16 pre-trained on RVL_CDIP for the task of document image classification based on the two following findings: (1) a simple deep CNN architecture, especially VGG-16, showed better performance in the task of document classification than an ensemble model [1][3], and (2) a model pre-trained on RVL_CDIP outperformed both a model pre-trained on ImageNet and a model with random initialization.

The overall task can be detailed and broken down into two sub-tasks as below:

(1) Data acquisition: We first need to import datasets (i.e., campaigns) from By the People collection and manually label each image to construct a ground-truth. The number of data points in the smallest dataset in literature is 3,483 labeled images. So, hitting that number would be the best-case scenario. If this is not achievable, we can lower the bar to 1,000.
    a. **Subtask 1.** Write a script to download a bulk of images from LoC website using *loc.gov JSON API* to our *cdrhdev2* server.
    b. **Subtask 2.** Annotate each image with one of the following labels (integer format): (1) 0; handwritten, (2) 1; typed, and (3) 2; mixed.

(2) Training model: While we are doing the data acquisition, at the same time, we can setup and start Experiments 1 and 2. Once we have a dataset from the By the People collection, we can conduct Experiment 3.
    a. **Experiment 1.** In order to reproduce the results of aforementioned papers, we start training VGG-16 pre-trained on ImageNet with a *subset* of RVL_CDIP.
    b. **Experiment 2.** In order to generate a VGG-16 pre-trained on RVL_CDIP, we start training VGG-16 pre-trained on ImageNet with *full* RVL_CDIP.
    c. **Experiment 3.** We start training VGG-16 pre-trained on *full* RVL_CDIP with a dataset from By the People collection.

Reference

[1] Harley, A.W., Ufkes, A. and Derpanis, K.G., 2015, August. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 991-995). IEEE.
[2] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
[3] Afzal, M.Z., Kölsch, A., Ahmed, S. and Liwicki, M., 2017, November. Cutting the error by half: Investigation of very deep cnn and advanced training strategies for document image classification. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Vol. 1, pp. 883-888). IEEE.

Progress Report – Document image classification for digital library collections

08/20/2019
Mike Pack

Objective

In this report, we aim to report on the following two experimental results: (1) classification performances of *VGG-16*, pre-trained on ImageNet, trained and tested with *RVL_CDIP* dataset and (2) classification performances of *VGG-16*, pre-trained on *RVL_CDIP* in (1), trained and tested with *suffrage_1002* dataset collected from the By the People corpus. The remainder of this report is organized as follows: in Experiment 1, a configuration of a dataset (i.e., *RVL_CDIP*) and training process is described, followed by training and testing results. In Experiment 2, similar to Experiment 1, a configuration of a dataset (i.e., *suffrage_1002*) and training process is described, followed by training and testing results.

Experiment 1:
Training and Testing *VGG-16* pre-trained on *ImageNet* with *RVL_CDIP*

The objective of this experiment is to reproduce the result reported in the work of Harley et al. (2015) which is training a model, *VGG-16*, with a large-scale document image dataset (i.e., *RVL_CDIP*) using transfer learning [1]. The advantage of this experiment is that once we have a model trained on this large-scale document image dataset, we can reuse the rich features that this model has learned for many document analysis tasks, say, one of our main tasks, a document type classification.

Dataset: *RVL_CDIP*

The *RVL_CDIP* dataset, which is publicly available, consists of 400,000 document images that are divided into 16 evenly distributed classes. The dataset is provided in three different sets: training, validation, and test set. The training set contains 320,000 images of 16 different evenly distributed classes (i.e., about 20,000 images per class). Both validation and test sets together contain 40,000 images of 16 different evenly distributed classes (i.e., 2,500 images per class).

Network Architecture: *VGG-16*

We use the original *VGG-16* architecture, but the output tensor is adjusted to have a shape of 16, which is the number of classes found in the *RVL_CDIP* dataset.

Training

As a preprocessing step, in order to make the shape of our input to match with that of *VGG-16*, we convert grayscale images to three-channel images by simply copying the pixel values of the single-channel to three channels. Also, each image is resized to 224 by 224, and normalized to range from 0 to 1 by dividing each pixel's intensity value by 255. In accordance with the size of the training set and under a limited memory constraint, we use a batch size of 126. As an optimizer, we use adaptive momentum estimation, or so-called Adam, which is the state-of-the-art optimizer and also known as the rule-of-thumb [2]. The initial learning rate is set to a small value, 10^{-5}. This is because the model has been already pre-trained on *ImageNet*, so we want to prevent overshooting local minima of the loss function. The training is scheduled to run 80 epochs total, but we use early-stopping to terminate the training process if the validation loss is not improved than that of the previous iteration.

Results

Interestingly, the entire training process took only three epochs to converge with promising classification results. This indicates that features obtained from natural scene images (i.e., ImageNet) are general enough to be applied to documents. The resultant classification performance metrics—precision, recall, and f1-score—are shown in Table 1. On average, each metric shows around 87%, which aligns well with the result reported by Harley et al. (2015). In Figure 1, more detailed classification performance on the test set is visualized as a heatmap. *A series of high support values in diagonal elements indicates that the trained model is capable of producing many correct predictions.*

Table 1. Precision, recall, and f1-score of *VGG-16* trained on *RVL_CDIP* dataset. The alphabetic labels are corresponding to the following labels: *letter*, *form*, *email*, ***handwritten***, *advertisement*, *scientific report*, *scientific publication*, *specification*, *file folder*, *news article*, *budget*, *invoice*, *presentation*, *questionnaire*, *resume*, and *memo*. Our class of interest, ***handwritten***, is bolded.

(unit: %)	A	B	C	**D**	E	F	G	H	I	J	K	L	M	N	O	P	Avg
Precision	86	74	98	**89**	89	73	90	88	89	92	87	91	78	91	92	88	87
Recall	94	79	97	**96**	91	73	93	91	97	86	83	86	79	73	94	91	87
F1	86	77	97	**92**	90	73	91	90	93	89	85	88	79	81	93	90	87

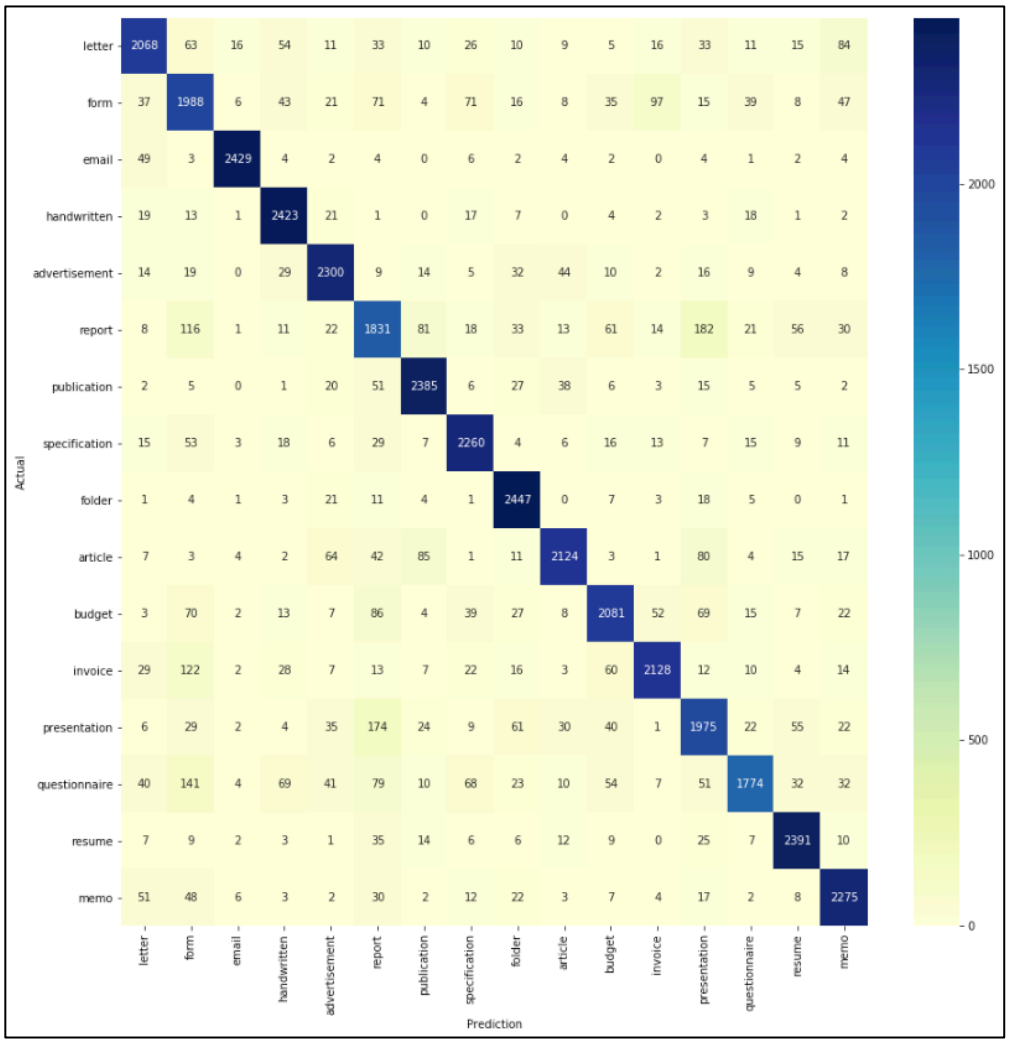


Figure 1. Heatmap of confusion matrices for classification performance of *VGG-16* trained on *RVL_CDIP*. Note that the diagonal elements represent the numbers of occurrences for which the predicted label is equal to the true label, while off-diagonal elements are those that are misclassified by the classifier. The higher the diagonal values of the confusion matrix the better, indicating many correct predictions.

Experiment 2:
Training and Testing *VGG-16* pre-trained on *RVL_CDIP* with *suffrage_1002*

The objective of this experiment is to generate our own model for this specific task—three-class document type classification; *handwritten*, *typed*, *mixed*—by retraining the model obtained from the previous Experiment 1 with our own *suffrage_1002* dataset.

Dataset: *suffrage_1002*

Thanks to Dr. Lorang and Ashlyn Stewart, we have collected a total of 1,002 images from a suffrage collection in By the People corpus[1]. This dataset is a fully balanced set (334 *handwritten*; 334 *typed*; 334 *mixed*) that has been compiled manually. The entire dataset is split into three sets—training, validation, and test—with the ratio of 8:1:1. Note here that in order to keep the class balanced during this split, it is inevitable to drop some datapoints (i.e., three datapoints). Our final dataset configuration is elaborated in Table 2.

Table 2. Configuration of *suffrage_1002* dataset.

	handwritten	**typed**	**mixed**	Total
train	267	267	267	801
validation	33	33	33	99
test	33	33	33	99
Total	333	333	333	999

Network Architecture: *VGG-16*

We use the same *VGG-16* architecture as in Experiment 1, but the output tensor is adjusted to have a shape of 3, which is the number of classes specified in *suffrage_1002* dataset.

Training

All the training configuration is the same as the previous Experiment 1, except for an initial learning rate and batch size. We use an initial learning rate of 10^{-6}, which is smaller than the one used in Experiment 1, since the model is pre-trained on *RVL_CDIP* on top of *ImageNet*. We also use a smaller batch size of 32 in accordance with the size of *suffrage_1002* dataset.

Results

Generally, one can diagnose whether a model is overfitted or underfitted to its training dataset based on a model's training and validation loss. For example, if a validation loss increases while training loss decreases, the learned model is speculated to have overfitted. Taking this into account, as shown in Figure 2, based on the overall decreasing trends of both training and validation loss, during the training, there is no symptom of overfitting or underfitting.
Overall, our model's classification performance on the testing set shows about 90% of precision, recall, and f1-score, as shown in Table 3. Compared to the other two classes, a *mixed* type shows relatively poor recall performance (i.e., 79%). We believe that this is due to challenging characteristics of *mixed* type document images; for example, too small amounts of handwriting

[1] https://crowd.loc.gov/topics/suffrage-women-fight-for-the-vote/

in a *typed* document, or vice versa, as shown in Figure 3. In Figure 4, more detailed classification performance on *suffrage_1002* test set is visualized as a heatmap.

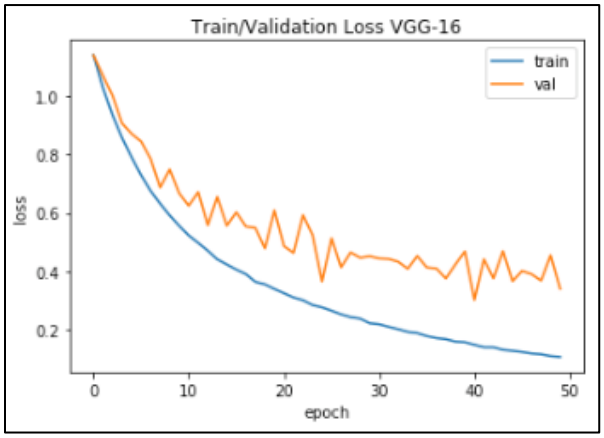


Figure 2. Training and validation loss of *VGG-16* with *suffrage_1002* training and validation set. In spite of some fluctuations, the overall trend of validation loss goes down.

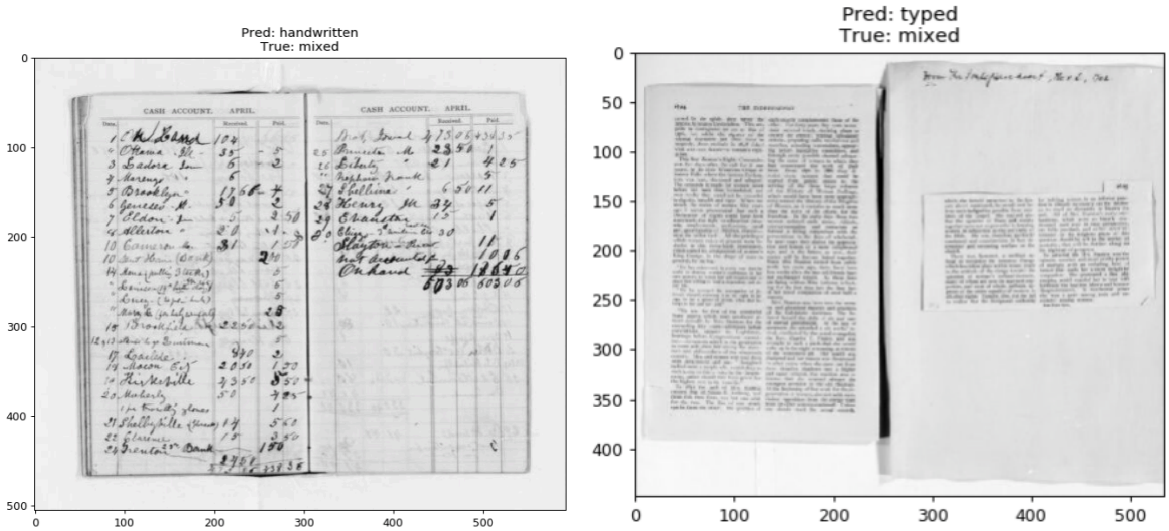


Figure 3. Failure prediction cases. On the left example, a typed region is relatively smaller than that of handwriting. On the right example, a handwriting region is relatively smaller than that of typing.

Table 3. Precision, recall, and f1-score of *VGG-16* on *suffrage_1002* testing set.

(unit: %)	**handwritten**	**typed**	**mixed**	**Avg**
Precision	89	91	90	90
Recall	97	94	79	90
F1	93	93	84	90

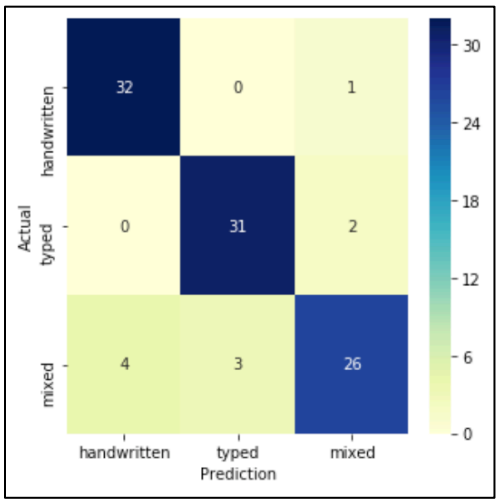


Figure 4. Heatmap of confusion matrices for classification performance of *VGG-16* trained on *suffrage_1002*. Note that diagonal elements contain most of datapoints, which indicates that most of our model's predictions are correct over all three classes.

Reference

[1] Harley, A.W., Ufkes, A. and Derpanis, K.G., 2015, August. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 991-995). IEEE.
[2] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Progress report

Yi Liu

Current Progress

Document Image Quality Assessment

In the last progress report (Document image quality assessment for digital library collections), we proposed to perform document image quality assessment (DIQA) to By the People dataset. Hence, in this report, we downloaded 36,003 images from the civil war collection (the dataset) of By the People. And we analyzed the outcome of the assessment results.

The DIQA algorithms used in this experiment were developed as part of the project Aida to assess qualities of newspaper page images from 1834 to 1922, which included four criteria: (1) skewness, (2) contrast, (3) range-effect, and, (4) bleed-through (background noise). We found that, for newspaper page images, a contrast score higher than 40 could be considered as having good contrast quality. And a range-effect score lower than three could be considered as having no or fewer range-effect issues. However, there was no clear indicator for skewness and bleed-through assessment. All we could say was that the lower the score on skewness or bleed-through, the better the quality.

In this statistical analysis, there were 35,990 out of 36003 images that successfully passed the quality assessment program. 13 images failed due to exceptions of the program caused by incorrect assumptions. We will later dig into the program to find the detailed reasons causing these exceptions.

Skewness. For skewness evaluation shown in Figure 1, there are 43.63% (15,703 out of 35990) images in the dataset with the maximum skewness score (i.e., score of 2). Hence, there are 43.63% images that are significantly skewed. There are 7.25% images that are lightly skewed (i.e., skewness ~1-2) in the dataset. Further, 2.48% of the images are trivially skewed (i.e., skewness < 1) in the dataset. And there are 43.63% images that are not skewed at all. Note that the larger the absolute value of the score, the more skewed the document image. And a positive or negative score indicated the skewness orientation. In Figure 1, "|score|" means the absolute value of the skewness score.

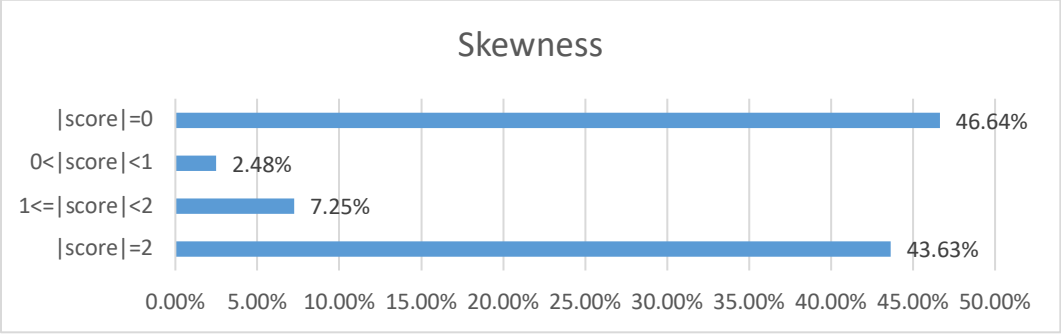


Figure 1 Skewness analysis

Contrast. For contrast evaluation, shown in Figure 2, images from 1930 to 1939 result in lowest contrast score (i.e., score of 23.87). And images from 1910 to 1919 result in highest score (i.e., score of 70.88). Note that, in this analysis, the higher the contrast the better the visual quality. Hence, based on the study

of Aida (i.e., score above 40 indicating good quality in contrast evaluation), except for images from 1860 to 1869 and from 1930 to 1939, the collection has a good contrast quality. However, there are 90% images from 1860 to 1869 in the collection. Hence, the 10-year chart (Figure 2) is not a good representation of the overall collection. As a result, we break the 10-year period from 1860 to 1869 into a year-by-year chart, shown in Figure 3.

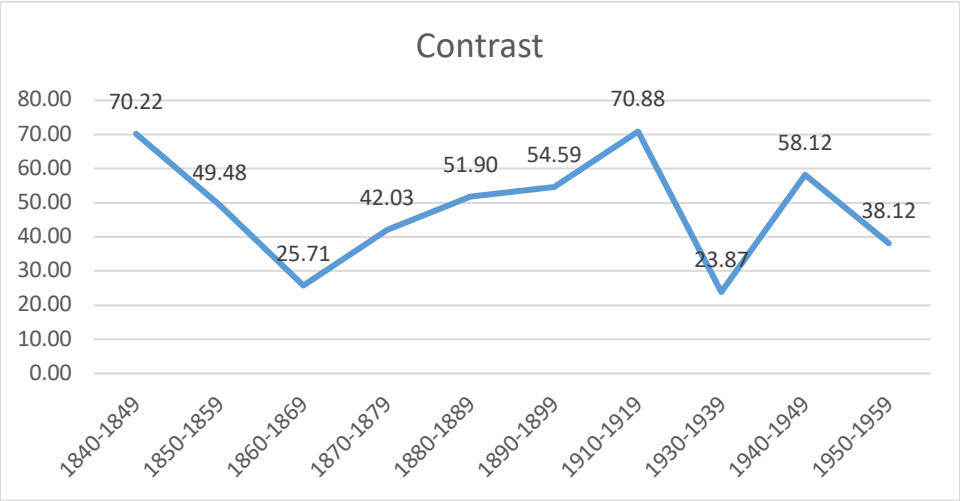


Figure 2 Contrast Score Analysis

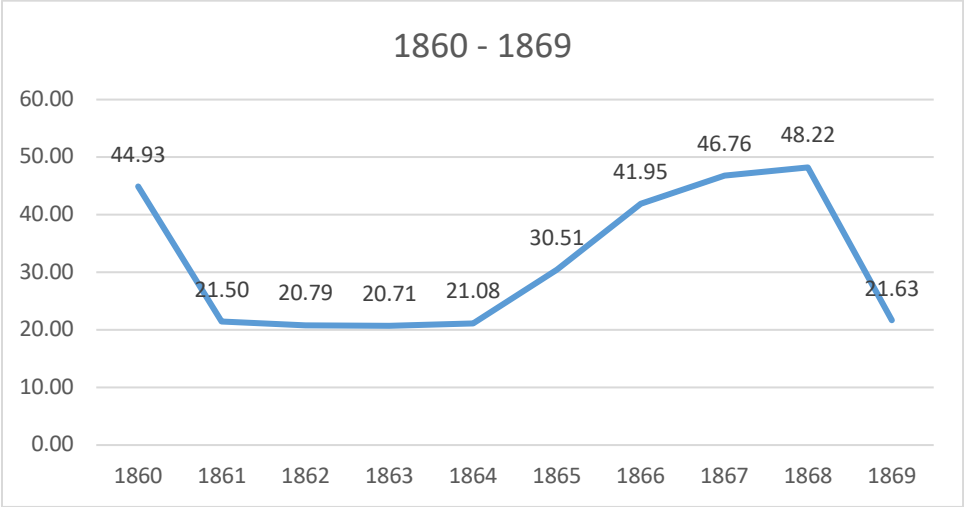


Figure 3 Contrast Score Analysis from 1860 to 1869

The breakdown chart shows that images with low score are from years 1861 to 1865. We suspect that the low score could be document images that are digitized from handwritten letters, shown in Figure 4. There are two problems among these letters that could lower the contrast score. First, the background largely suffers from yellowing. And, second, the ink is significantly faded. Further, we see that the appearance of low scores overlaps with the civil war years. Hence, the low score may also due to the degradation of the document considering the plausible challenges in newspaper preservation during the war.

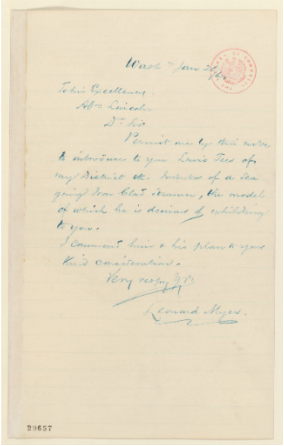


Figure 4 Handwritten letter with fading out ink

Range-effect. For range-effect evaluation, shown in Figure 5, images from all (but one) year ranges have relatively low scores. For the year range of 1930-1939, there are two with relatively high scores and hence the score of 27.33. Note that, for range-effect evaluation, the lower the score the better the quality. However, compared to our baseline study done in the Aida project that found any score below three implied good quality in range-effect evaluation, the civil war collection suffers from relatively more range-effect problems than the newspaper collection previously evaluated by Aida. This does not mean that the visual quality is necessarily visually for human perception. But it indicates that the collection could need substantial preprocessing to reduce range-effect before in-depth analysis.

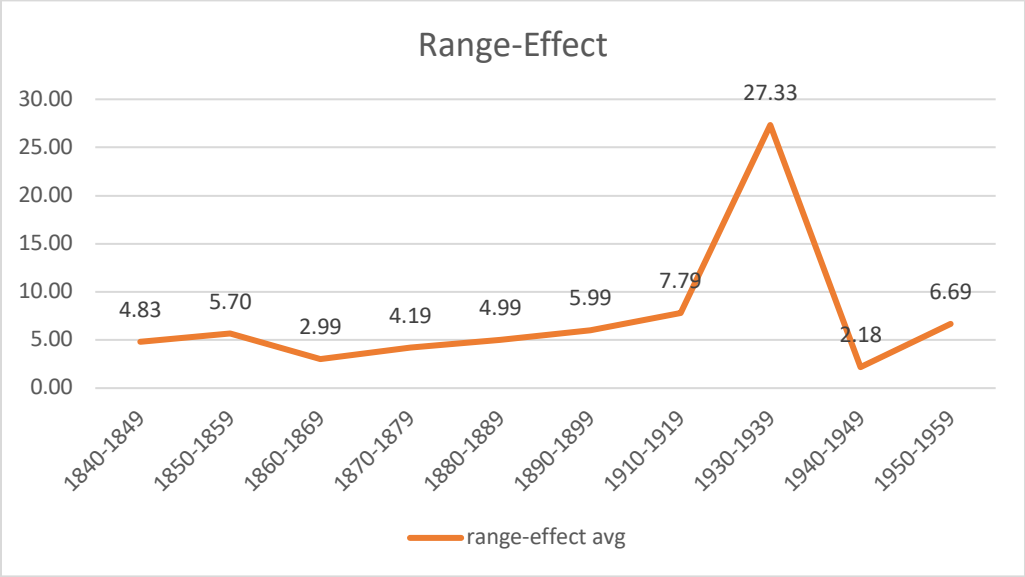


Figure 5 Range-effect Score Analysis

Bleed-through. For bleed-through (background noise) evaluation, shown in Figure 6, again, images from all year ranges (except one) have relatively low scores on bleed-through evaluation. For the year range 1940-1949, there are 76 images with high scores and hence the score of 12.10). Note that, the lower the bleed-through score the better the quality. However, a score identifying generally good quality does not exist for bleed-through evaluation. We can only confidently say that the score of zero is ideal.

Further, based on observation, the "paper yellowing" issue is a major problem in the collection. In our processing, a document image is first converted into a grayscale image by the evaluation algorithm. Hence, the yellowing paper results in a dark background after the conversion. A dark background would affect bleed-through evaluation, even, might result in a faulty evaluation. However, this does not mean that the bleed-through evaluation is not useful. Considering, in a way, the bleed-through evaluation represents the quality of background cleanliness, and thus, a high score can suggest that the background may need a noise removal process.

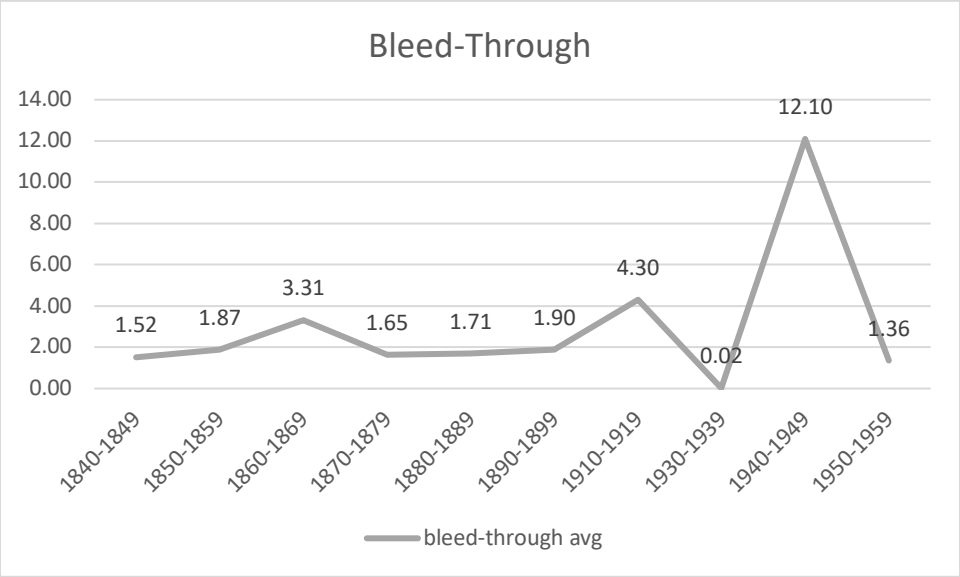


Figure 6 Bleed-through Score Analysis

Differentiation between Microfilm and Scanned images

The types of digitization that generated the document images are mixed. There are both microfilms and scanned images in the collection. As a result, techniques developed for one type might not work for the other. In our DIQA suite of image processing tools, for example, we assume that the document images were scanned images, with white or brighter pixels as background, and darker pixels as texts. However, documents from microfilm sometimes have inverted range of pixel intensities, rendering our image processing tools not effective. Hence, we propose a way to differentiate the digitization type of document to metatag them for further processing.

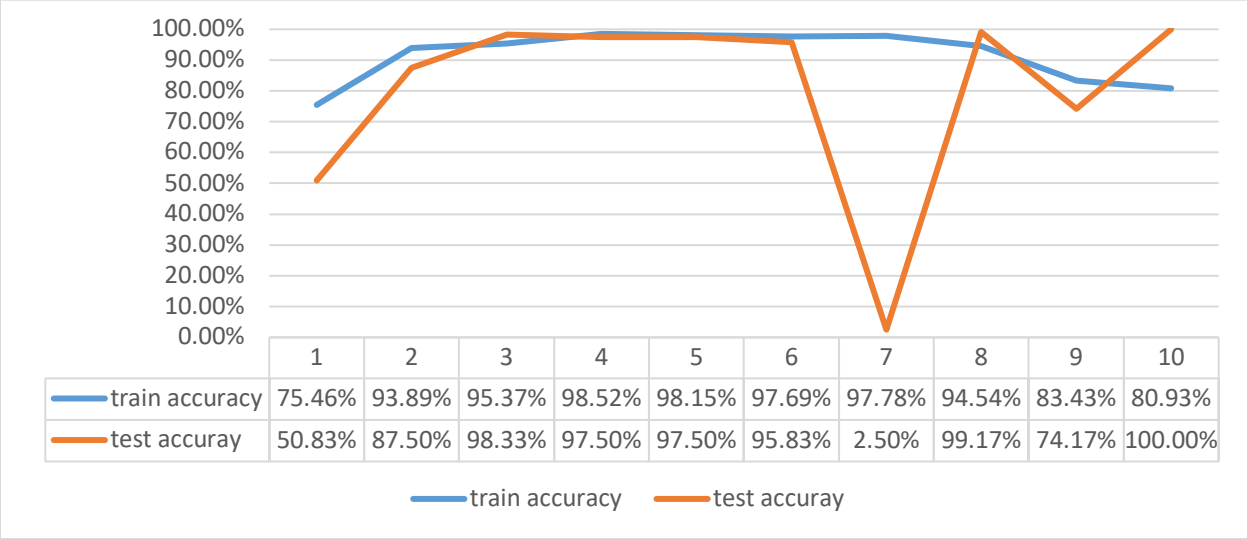
We propose to adopt the current state-of-art image classification model, called ResNeXt, to classify the digitization type of documents. In addition, to train the model, we need a set of labeled images. Hence, we manually build a database containing 1200 images from the civil war dataset. In this database, there are 600 scanned document and 600 documents from microfilm. A balanced database is built so that the training will not be biased.

Further, in a general idea of a machine learning training process, we want to keep the database as balanced as possible to prevent bias problem. This applies to not only numbers of instance for each label, but also other aspects such as skewness, contrast, range-effect, and bleed-through. In other words, we want our model to "see" as many conditions as possible during the training. Hence, during the creation of

the database, we randomize the file list to make each image in the collection has a fair chance to be included by the database.

Moreover, we also want to maintain replicability for future studies. So, the randomization was performed with a fixed random seed using a pseudo-randomization algorithm. By taking advantage of the randomization algorithm, we can reproduce the result as needed.

Shown in Figure 7, the ResNeXt model works very well on differentiating the two digitization types. The training process took only two iterations to reach over 90% accuracy. And test accuracy reached 100% correct at the 8th iteration. We see that the test accuracy at 7th iteration drops to 2.5%. This may be caused by the optimizer of the training process. The optimizer keeps a momentum to make the training process to be able to jump out of a local minimum. Hence, it may result in abnormal test accuracy. However, the test accuracy in these iterations does not necessarily affect the final performance of the classification as long as the training does not stop on these iterations.



	1	2	3	4	5	6	7	8	9	10
train accuracy	75.46%	93.89%	95.37%	98.52%	98.15%	97.69%	97.78%	94.54%	83.43%	80.93%
test accuray	50.83%	87.50%	98.33%	97.50%	97.50%	95.83%	2.50%	99.17%	74.17%	100.00%

Figure 7 Digitization Type Differentiation using ResNeXt-100 64x4d

Work That Has Been Done

Task 1

36,003 images from the civil war collection were downloaded through the website of By the People. And the downloaded image was backed up and stored in the CDRH server of the Aida team.

@cdrhdev2.unl.edu/var/local/aida/by-the-people_civil-war

Task 2

Collect information of creation/publication years of the corresponding item of the civil war collection for DIQA analysis.

@cdrhdev2.unl.edu/var/local/aida/by-the-people_civil-war/civil-war-images-info.csv

Task 3

Manually create a database containing 1200 images to perform training and evaluation for the digitization type classification.

@cdrhdev2.unl.edu/var/local/aida/by-the-people_civil-war/microfilms.txt

@cdrhdev2.unl.edu /var/local/aida/by-the-people_civil-war/scans.txt

Task 4

Adopt ResNeXt model from ImageNet-1000

Task 5

Create corresponding code to fine-tune and classify the digitization type.

Progress Report on Second Iteration

Yi Liu

1. Differentiation between Microfilm and Scanned images

In the first iteration of this project, we proposed to adopt the state-of-art image classification model, ResNeXt, using transfer learning, to classify the digitization type of documents. In addition, to train the model, we created a labeled database containing 1,200 images from the Civil War dataset. In the database, there are 600 scanned document and 600 scanned documents from microfilm.

In the second iteration, we continue to fine-tune the model for a better classifier. Based on the observation during the labeled database construction, the ratio of the number of digitized materials from microfilm to those from the scanning process is about 1 to 16. Hence, there are three metrics to evaluate the fine-tuned classifier.

Further, three evaluation metrics are (1) training performance, (2) validation performance, and (3) prediction performance. First, the training performance is the classification performance on the training set, which is 90% of the 1200-database. This metric represents the ability of the classifier on classifying data has been seen by the model. Second, the validation performance is the classification performance on the validation set, which is the rest 10% of the 1200-database. And this metric validates the training process to compute an expected prediction performance using a small set of labeled unseen-data. And third, the prediction perfrmance is an evaluation of the entire Civil War collection. Based on the previous observation, the entire Civil War is expected to have about 2,256 document images digitized from microfilm. Hence, by comparing the predicted ratio of microfilmed and scanned document images, the strength of the classifier can be observed.

In the experiment, at which time to stop the training process and save the trained weights of the classifier is based on the training performance and validation performance. The general idea is to stop the training when both training and validation performances are good (i.e., the harmonic mean of training and validation F1 scores is greater than 99%). At the same time, we want to avoid overfitting and underfitting. Overfitting means the training performance is higher than the validation performance. Hence, the classifier could suffer from picking up noise when overfitting occurs. Underfitting means the validation performance is higher than the training performance. This is where the prediction could be biased. Considering the harmonic mean of two metrics has high response if two metrics have high values, and, at the same time, they are close to each other. Therefore, we compute the harmonic mean of the training and validation performance to decide the stopping point of the training.

Shown in Figure 1-4, the model started to converge usually after around 30 epochs. And both training and testing performances on the accuracy, precision, recall and F1 score are very promising. After convergence, the best epoch is the 44th training epoch, where the training accuracy is 98.52%, and the validation accuracy is 100%. Hence, the 44th epoch is stored for analyzing prediction performance.

In the prediction performance analysis, the stored classifier made predictions on the entire Civil War collection. Table 1 shows the prediction results. The prediction ratio of microfilmed document images to scanned document images is roughly 12:1. Hence, the classifier is more generous in classifying a document

image as digitized from microfilms than the expectation (i.e., 16:1). Figure 5-8 shows 4 types of typical mis-classifications.

The four types of "problematic" document images, are: (1) one that is largely "blank" (e.g., Figure 5); (2) one that has poor contrast quality (e.g., Figure 6); (3) one that is a picture of a physical item (e.g., a coin in Figure 7); and (4) one that is a graphical photo (e.g., a portrait photo in Figure 8). We suspect that there are two possible reasons. First, for type (1), there is little information for the classifier to make prediction since the document image contains largely background pixels. Second, for type (2), the poor quality could weaken the visual features that are required for the classifier to make the prediction. Third, these four types are rare or missing from the training database. Hence, the classifier was not trained sufficiently to make predictions.

Therefore, for future iterations of this project, two options could effectively improve the performance further. First, we can expand the training database to include more examples of the four-type document image to increase the variety. Second, we can apply a pre-processing step to normalize the document image quality for the collection before the prediction stage.

2. Conclusion

We found that classification performance for the digitization type differentiation to be promising. There are some mis-classified cases. However, the problem could be fixed by increasing the variety of the training database and applying pre-processing techniques. Further, although the microfilmed photo was not included in the training database, the classifier was able to correctly predict such photo as microfilmed material, shown in Figure 9. This suggests that the model has the generality to apply on a large collection for digitization type prediction.Type equation here.

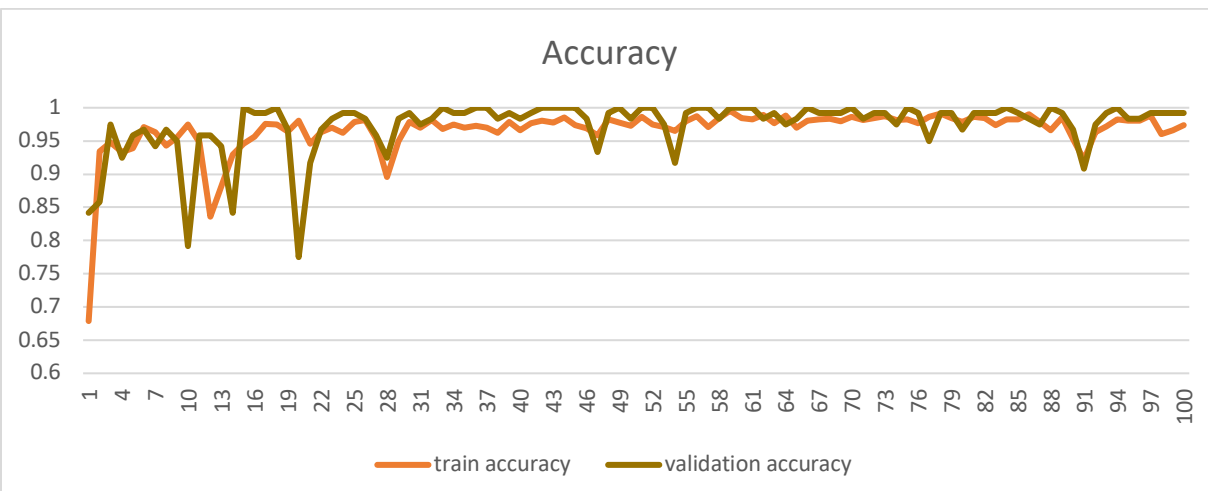


Figure 1 Training and validation accuracies

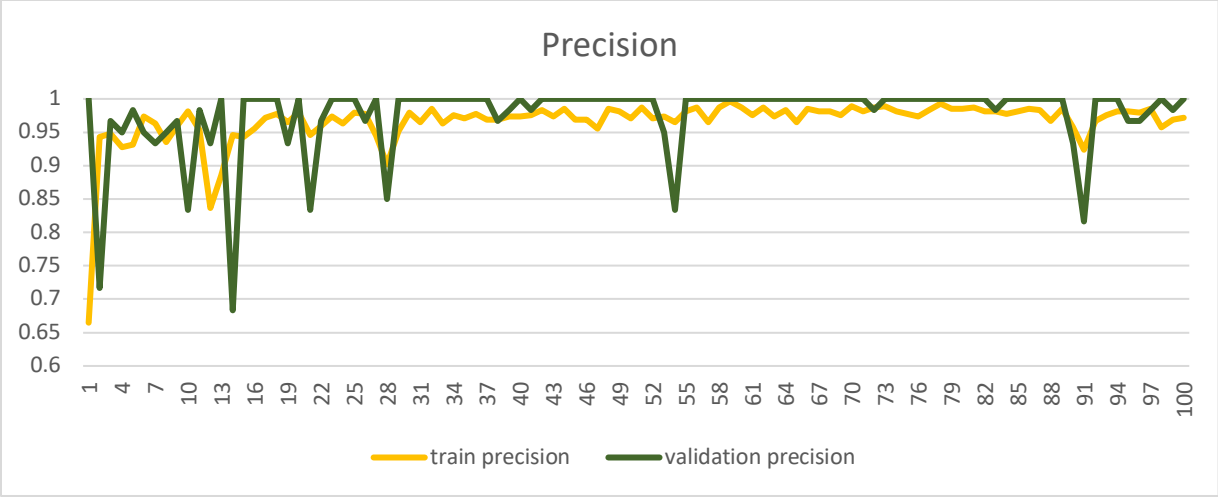


Figure 2 Traing and validation precisions

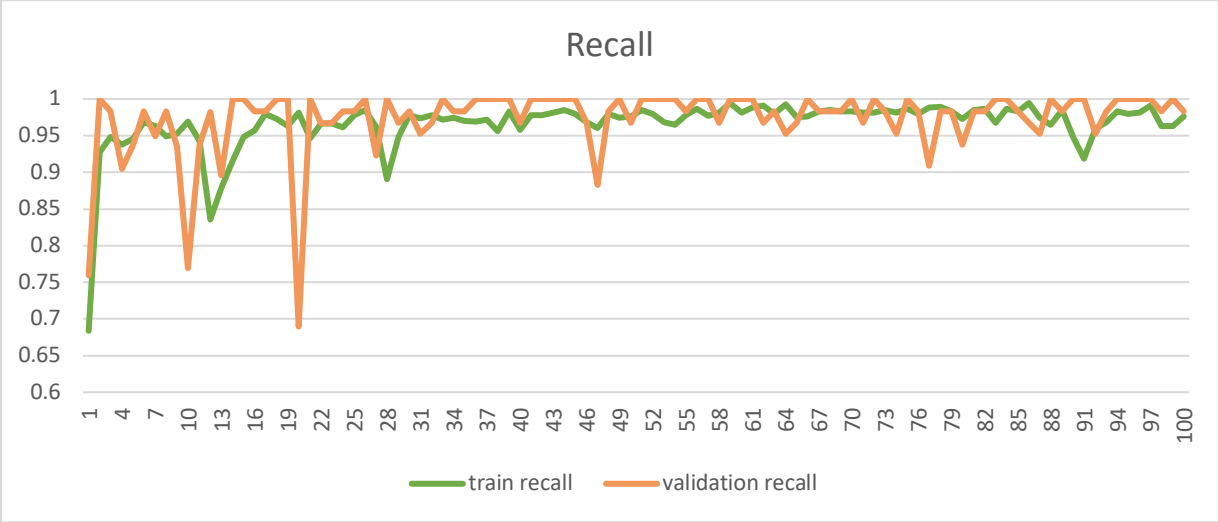


Figure 3 Training and validation recall

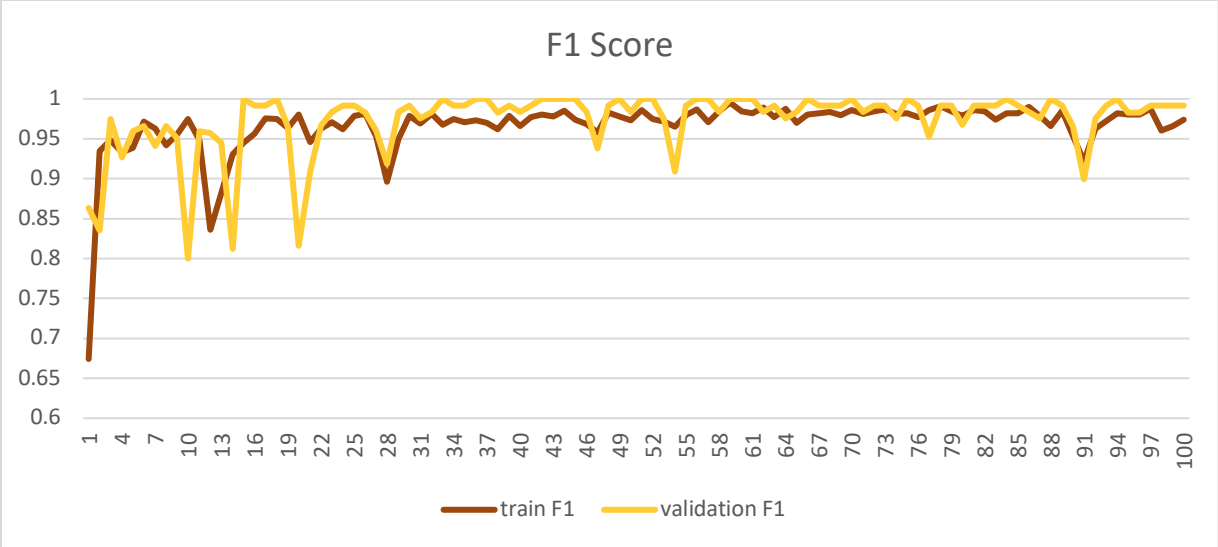


Figure 4 Training and validation F1 score

Table 1 Prediction Results		
Total	Predicted Microfilmed Documentation	Predicted Scanned Documentation
36103	2834	33269

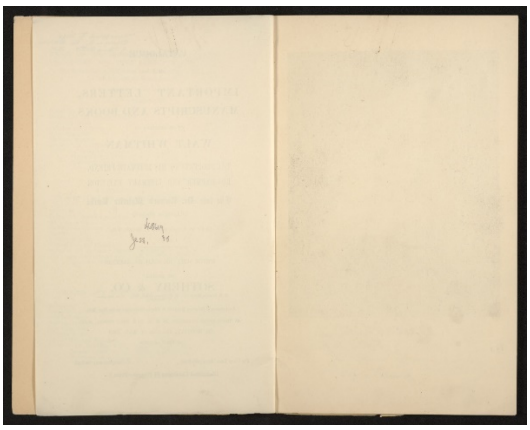


Figure 5 Type (1) mis-classification: "blank" document image

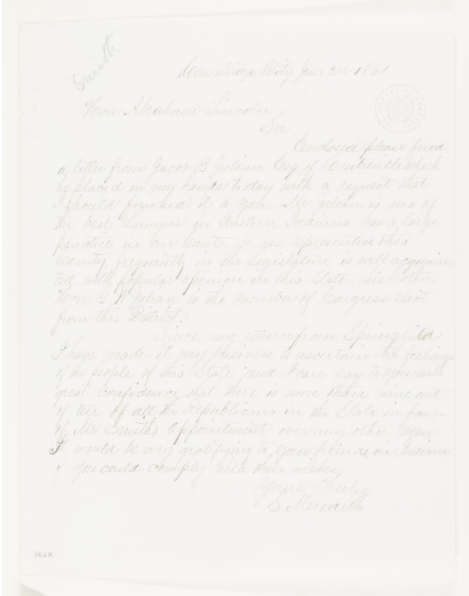


Figure 6 Type (2) mis-classification: poor contrast quality

Figure 7 Type (3) mis-classification: item images

Figure 8 Type (4) mis-classification: graphical images (photo)

Figure 9 Microfilmed frame-photo being correctly classified

Progress Report on Second Iteration

Yi Liu

1. Introduction to Weak Supervision

Machine learning models, especially for deep learning models, needs large amounts of labeled data for the training purpose. However, these data usually involves the high-level supervision, such as manually visual supervision. The traditional approach to get such supervision involves subject matter experts whom is familiar to the database and the task. Considering the amount of the labeled data required, it could cost months or even years to build the groundtruth database for the experts.

In fact, one of the most important advantages that the deep learning model is that, such model allow developers to get the state-of-art performance without the hand-engineered features of the data. Such features involves the high-level supervision and abstraction of the subject matter experts. Hence, the deep learning models free the labor of the expert during the model design. However, they are not compeletely freed from the loop of the machine learning, as we still need high-level supervised dataset for training.

Therefore, a tool is needed to get the large training dataset for machine learning techniques. The waek supervision is a new programming paradigm designed form machine learning. It provides a bridge between high-level supervision and dataset required by machine learning models. The weak supervision introduces programming tools such as heuristic rules, constraints, and invariances to get labeled data based on subject matter experts' knowledge. Hence, the basic procedure is, first, get detailed description on how to label data from the subject matter expert. Second, developers convert the description to computer language using the weak supervision. Third, the computer runs to build the labeled dataset.

2. Motivation

2.1 A large training dataset is commonly required

A general idea to train a better machine learning model, especially a deep learning model, is the more data the better.

In the first iteration of this project, we proposed a two-step approach to extract figure/graph and generate metadata for beyond words collection. The first step, an FCN (U-NeXt) combining ResNeXt and U-Net was built and trained to segment and classify graphic snippets on newspaper pages based on ground truth extracted from Beyond Words. Besides, the ResNeXt part of the model was transferred from pre-trained ImageNet ResNeXt-101 to reduce training parameters. Based on dhSegment, using transfer learning is able to boost training effectiveness, and preserve a good performance. The second step, a text segmentation, and recognition model retrieved textual content in the graphic snippets (i.e. extracted graphic snippets from the first step). Specifically, EAST text detection was applied to find text regions for an OCR process to retrieve words within graphic snippets. And the retrieved word was encoded into metadata for further usages, such as search queries.

In the second iteration, we focus on evaluating and improving segmentation step using U-NeXt model. The U-NeXt model is an extension on dhSegment model. The dhSegment used pre-trained ResNet while our U-NeXt used pre-trained ResNeXt model. Noet that, ResNeXt is an improved version of ResNet. In the study of dhSegment on Beyond Words collection, the classification accuracy was 88% and the mean

intersection over union (mIoU) was 26%. The U-NeXt is expected to have a better performance than dhSegment. Further, the EAST text dection largely depends on the performance of the segmentation step. Hence, improving segmentation step is a key opponent in this project.

Note, two metrics are used to evaluate the performance in this project. First, the classification accuracy is a pixel-wise accuracy. It computes the percentage of correctly labeled pixels to total numbers of pixels for each class. Second, the mIoU evaluates if the predicated region accurately covers the true region in the groud-truth overall classes. And for both metrics, the higher the better.

2. Dataset

Two datasets were used to train and evaluate the segmentation step using U-NeXt model. In Beyond Words collection (BW), some graphic region appeared on the page are missing in the ground-truth. And the marked region in the ground-truth does not tightly bond to the actual shape of the graphic region. Hence, we pre-train our model on a more comprehensively labeled dataset called Eurapean Newspaper collection (ENP). By doing so, during training, some local minimum, which created by the issue in Beyond Words collection, could be avoided. Specifically, the ENP contains 480 images in total, in which, there are 384 images in training set and 96 images in validation set. And the BW contains 1,532 images in total, in which, there are 1,226 images in training set and 306 images in validation set.

Further, the similarity shared by ENP and BW collections is the crutial reason why the ENP can be used for pre-training. First, both ENP and BW collections are document images that are digitized from newspapers. Hence, they share similar content layout and density. Second, the ground-truth of the ENP marked five classes: (1) background, (2) text, (3) figure, (4) layout separator, and (5) table; while the ground-truth of the BW marked background and five detailed type of figures. Hence, the learned knowledge from the pre-trained model on figures using the ENP prevides a good reference for U-NeXt to identify the figure region. Then, the fine-tuning using the BW could focus on detailed figure type differentiation than identify the figure region.

3. Experimental Results

In this experiment, early stopping is not applied since the expectation on the performance is unknown. Hence, we set the pre-training process on the ENP up to last 700 iterations, and the fine-tuning process on BW up to last 80 iterations to oberserve the performance as the preliminary result for observation.

Shown in Figure 1, the training performance of pre-training stage on the ENP reached 72% on training accuracy and 63% on training mIoU. And the testing performance of the pre-training stage reached 68% on testing accuracy and 53% on testing mIoU.

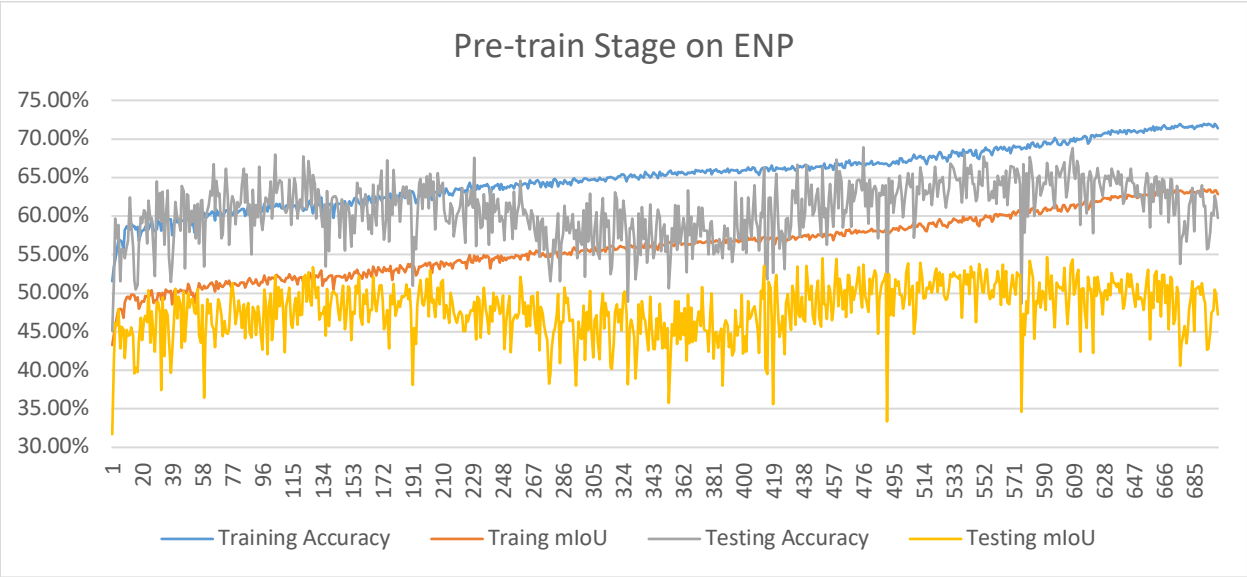
Shown in Figure 2, the fine-tuning stage on the BW was able to reach 64% on training accuracy and 47% on training mIoU, and reach 59% on testing accuracy and 52% on testing mIoU at an early iteration (i.e., at third iteration). However, the fine-tuning stage tried to classify all pixels as background pixels after convergence. This suggests that the Beyond Words ground-truth is severely biased. Amount of pixels from non-background classes is too small. During fine-tuning stage, a small weight was applied to background class (i.e., less than 0.1), the fine-tune still wants to classify all pixels as background.

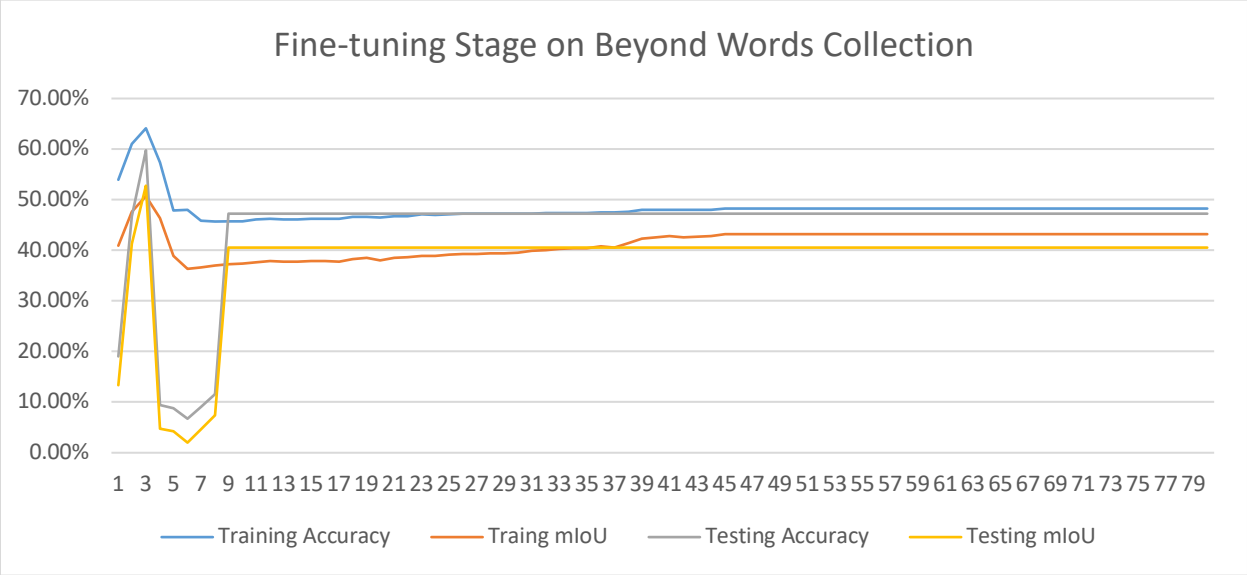
Therefore, based on the observation, by combining all non-background class as one class (i.e., graphics), splitting six-class segmentation task to a pipline of two tasks: (1) extraction of graphics and (2)

classification of graphic types could be a better approach. In preliminary experiment, the extraction of graphics had >60% on both training and testing mIoU.

2. Conclusion

We found that classification performance for the digitization type differentiation to be promising. There are some mis-classified cases. However, the problem could be fixed by increasing the variety of the training database and applying pre-processing techniques. Further, although the microfilmed photo was not included in the training database, the classifier was able to correctly predict such photo as microfilmed material, shown in Figure 9. This suggests that the model has the generality to apply on a large collection for digitization type prediction.





Progress Report – Deep Clustering for Metadata Generation

09/23/2019
Mike Pack

1. Introduction and Objectives

In this study, we explore and evaluate the potential for deep-visual-representation-based clustering approach to generate and analyze metadata from a large-scale digitized document image collection (in this case, European historical newspapers.) In particular, we utilize deep visual representation extracted from image classification/segmentation model and from which datapoints are clustered using t-SNE, which is one of the state-of-the-art clustering methods. The idea behind this approach is fairly simple. If we can obtain a manifold of a large-scale dataset, then many of meta related tasks (e.g., metadata generation, suggestion, and refinement) can become relatively trivial tasks based on the following two assumptions:

Assumption 1. The deep *visual* representation of each datapoint contains enough feature information to be clustered.
Assumption 2. In the clustered manifold, datapoints resides in the same *neighborhood* will share similar *visual* metadata to each other.

2. Method

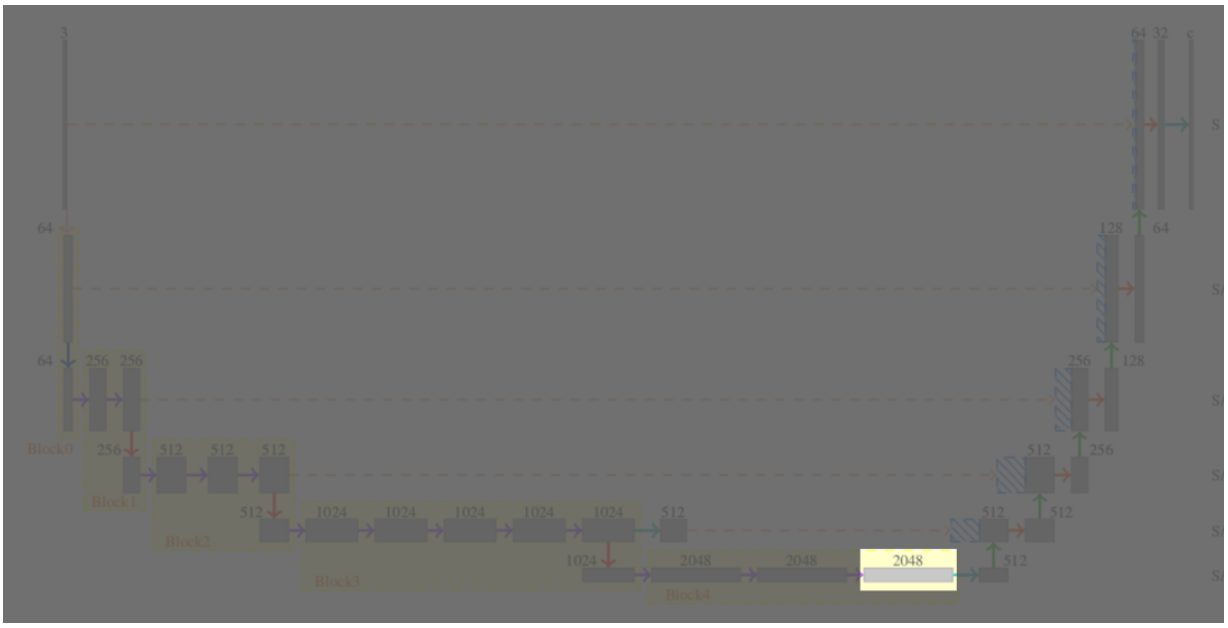


Figure 1. Latent space of *dhSegment*. This model is a combination of ResNet-50 and U-net. From this model, we extract only the highlighted block, which is known as a latent space.

First, unlike traditional clustering approaches using a set of hand-crafted features extracted from images, we extract a set of feature maps—so-called, latent space—learned by a deep model, in particular, a ResNet-50 + U-net that we trained for our first page segmentation project, as shown in Figure 1. Here, the size of latent space is $W/32$ x $H/32$ x 2048, visualized in Figure 2, here W and H are the width and height of input image, respectively. It is unsuitable to use this latent space directly for clustering task due to the following two aspects.

- First, the dimension of the latent space is *inconsistent* since the width and height of the input images can vary, whereas clustering method usually requires a fixed size of the dimension of features.
- Second, the dimension of the latent space is too *large*. For example, the resolution of our input images is usually about 1800 x 2400, and the corresponding latent space becomes about 1800 x 2400 x 2048, which would contain redundant information that degrades clustering performance in both quality and computation time.

Thus, for dimensionality reduction purpose, we perform an average pooling for each of 2048 feature maps to keep only the intensity of activation but ignore *where* that activation has occurred. By doing this, the dimension of our latent space is reduced down to 1 x 1 x 2048.

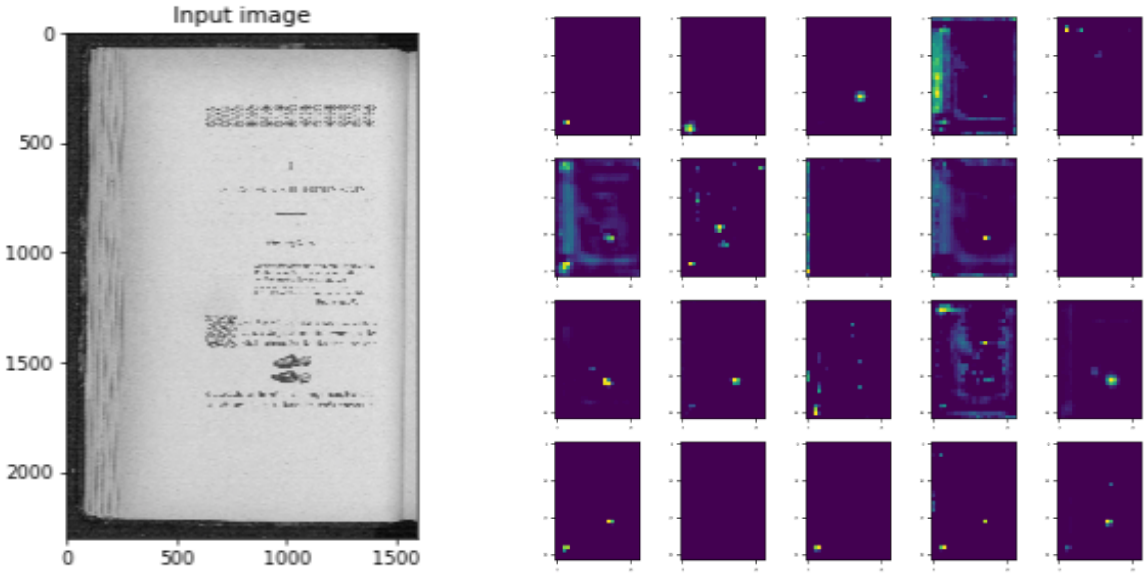


Figure 2. **Left**: Input image; **Right**: Visualized latent space down-sampled to *W*/32 x *H*/32 x 25

Our reduced latent space is then clustered using t-SNE, which is known to be an efficient and effective clustering algorithm that can map a high-dimensional data into a low-dimensional space, 2-d in our case. The way t-SNE works is by (1) converting the pairwise distances in both latent space and the low-dimensional space into probabilities that measures the similarity of two series of data points and (2) minimizing the Kullback-Leibler divergence between those two probabilities.

Once we obtain the manifold of our dataset in a low-dimensional space, we can then visually inspect and analyze whether datapoints residing in the same neighborhood (i.e., cluster) shares similar visual metadata, such as density, the existence of figure, layout, visual quality, etc.

3. Experimental Results

3.1. Initial Study

A total of 96 document images from European historical newspaper dataset are randomly collected and tested. The clustering result is shown in Figure 3. From the result, we can observe that 96 datapoints in 2048-dimensional space are grouped into roughly 3 clusters in 2-dimensional space. For each of 3 clusters, 4 datapoints are picked and colored in red, yellow, and green, respectively, to visually inspect the following two points: (1) *intraclass correlation*; whether datapoints in the same cluster share the similar visual features and (2) *interclass correlation*; whether different clusters show dissimilarity to each other.

First, as shown at the bottom row in Figure 3, the four sampled images in the same cluster does share similar visual features; for example, all four images in each color box (i.e., red, yellow, and green) show similar degree of brightness (i.e., white, gray, and dark) and contrast (i.e., high, medium, and low). This result implies that there is a certain amount of intraclass correlation; images in the same cluster somewhat resemble to each other.

Second, as shown at the bottom row in Figure 3, images in different clusters does show distinctive visual features; for example, images in the red box show a sparse layout (i.e., 2-column) whereas images in the yellow box show denser layout (i.e., more than 2 columns with various figures). Note here that image 8 shows a rather sparse layout and this is captured in the visualized cluster: its location is somewhat far apart from the rest of three datapoints in the yellow cluster (i.e., images 11, 66, and 5). Similarly, images in the green box show relatively dense layouts compared to the images in the red cluster and they also contain numerous figures.

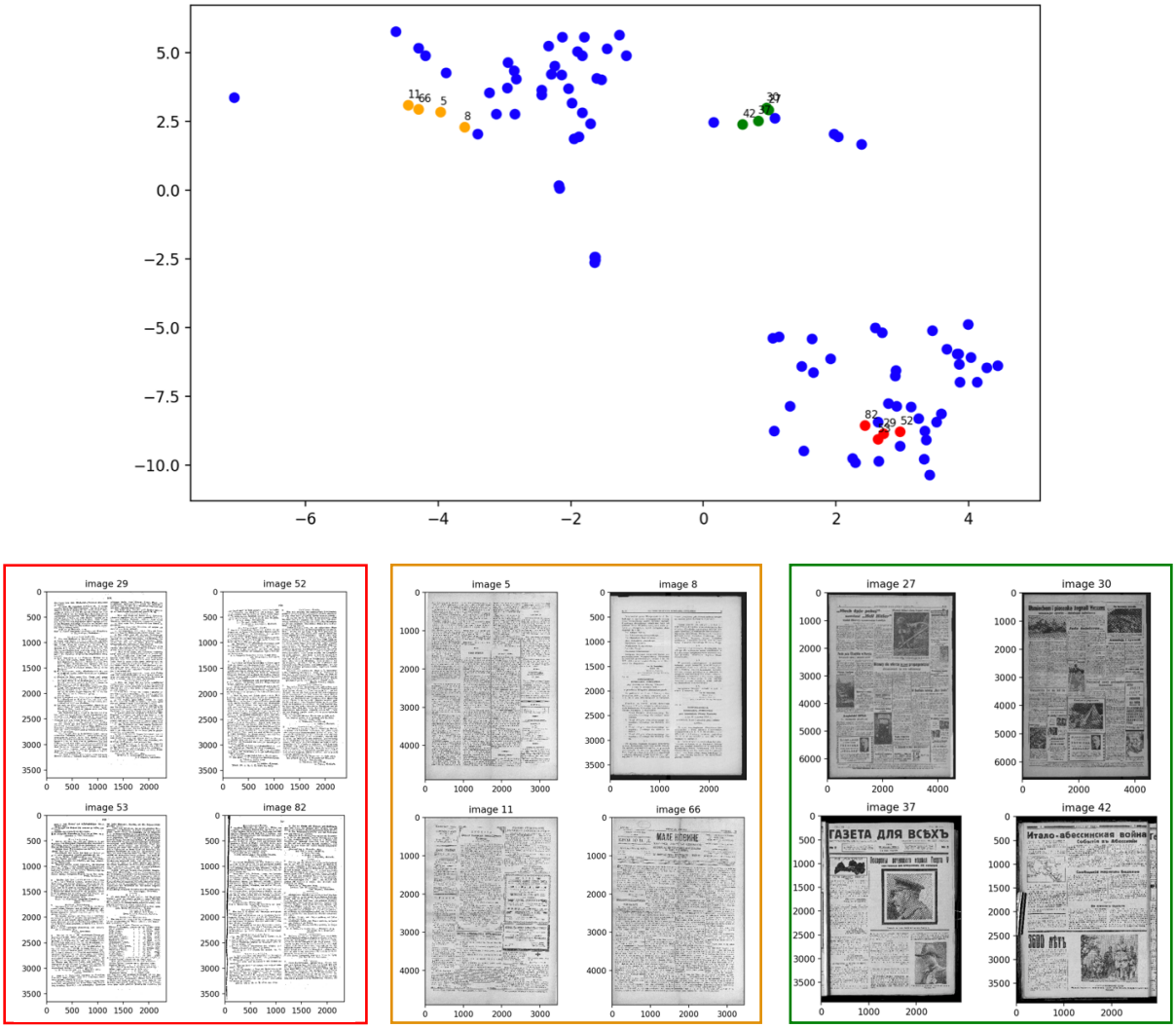


Figure 3. **Top row:** Visualization of latent space of ResNet-50 mapped into a low-dimensional space using t-SNE. For visual similarity inspection purpose, we selected four datapoints from different clusters labeled them in different colors; orange, green, and red. On the top-right of each exemplary datapoint, its image identification number is displayed. **Bottom row:** Actual image of each exemplary datapoint. Images are grouped in the bounding box in a color corresponding to that of datapoint. Note that images in the same cluster share similar characteristics, whereas different cluster shows different characteristics. For example, images in the red group show high contrast and simple layout structure. The images in the orange group show relatively grayish appearance without figure components. The images in the green group show relatively darker appearance with figure components.

3.2. Normalized Study

From the above first experimental results, it is reasonable to question whether this clustering result is simply based on the intensity value of image, and thus we performed the second experiment to cluster deep visual representations extracted from images that are normalized to have zero mean and a unit standard deviation of intensity value, as shown at the bottom row in Figure 4.

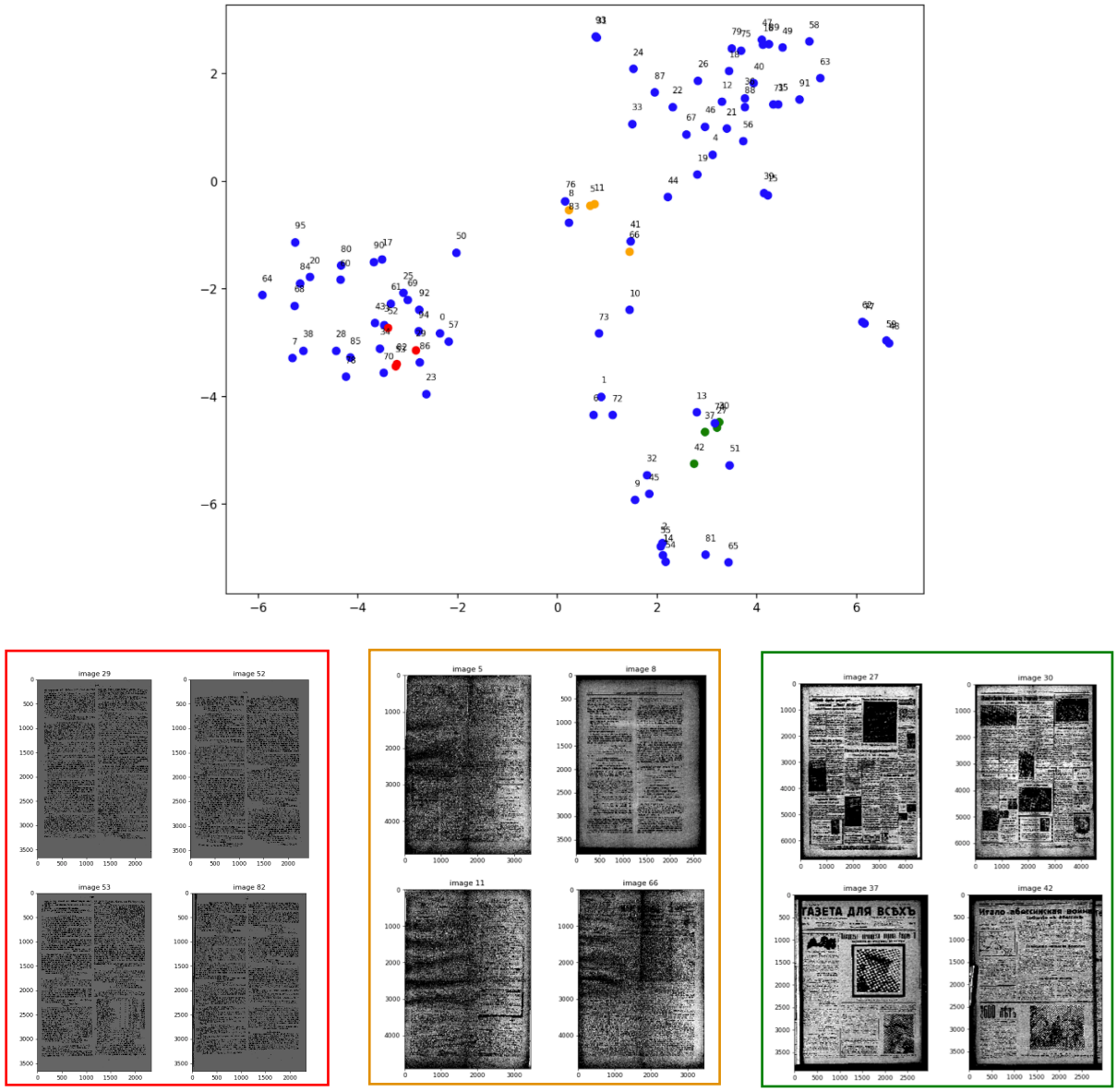


Figure 4. **Top row:** Visualization of latent space of ResNet-50 with normalization mapped into a low-dimensional space using t-SNE. For visual similarity inspection purpose, we selected four datapoints from different clusters labeled them in different colors; orange, green, and red. On the top-right of each exemplary datapoint, its image identification number is displayed. **Bottom row:** Actual image of each exemplary datapoint. Images are grouped in the bounding box in a color corresponding to that of datapoint. Note that at this time, images are normalized first, and then the deep visual representations are extracted and clustered. The clustering result shows similar clustering pattern as the previous clustering result; however, some datapoints sharing similar layout structure are slightly separated from each other, for example, image 66 and image 11 in the yellow cluster.

As shown at the top row in Figure 4, from the second experiment, we make two observations. First, the clustering result using the deep visual representation excluding intensity features shows a similar pattern to that of using the deep visual representation including intensity features. This outcome implicates that the performance of our clustering approach is not based primarily on intensity features. Second, based on the observation that some datapoints sharing similar layout structures are slightly separated from each other compared to the first experiment clustering result; for example, image 66 and image 11 in the yellow cluster, intensity feature does have an effect on the clustering process.

As another step to investigate and analyze the relationship between clusters and visual metadata, we can generate DIQA and complexity scores—that measure how dense and busy a document image is—for each datapoint and explore whether they are statistically significant or not for a more objective evaluation on the clustering result, which will be our 3rd iteration on DIQA and Segmentation/Classification.

4. Conclusions

In this study, we have presented a viable solution for visual metadata generation using a deep-visual-representation-based clustering approach. As shown in our first experiment, a set of deep visual representations of document images can be mapped into a low-dimensional space efficiently and effectively in which neighboring datapoints show considerable visual similarity. Also, as shown in our second experiment, this visual similarity is not based primarily on simple intensity features; rather on high-level visual features, such as layout density.

For better comprehensive understanding of deep-visual-representation-based clustering as a solution for visual meta generation, three additional experiments are needed: (1) investigate the use of another set of deep visual representations extracted by the *unsupervised* deep model (e.g., VAE) to build a more generic or universal deep-visual-representation-based clustering solution that is not limited to a specific document domain (European historical newspapers, in this case), (2) explore a more sophisticated way of *dimensionality reduction techniques* rather than a simple average pooling so as to retain spatial information for more accurate metadata generation, and (3) generate and analyze *DIQA* and *document complexity score* to examine whether clustering result is statistically significant or not, so that we can evaluate our solution in a more objective manner.

Progress Report on Second Iteration

Yi Liu

1. Figure/Graph Extraction for Beyond Words Collection

In the first iteration of this project, we proposed a two-step approach to extract figure/graph and generate metadata for the Beyond Words collection. The first step, an FCN (U-NeXt) combining ResNeXt and U-Net was built and trained to segment and classify graphic snippets on newspaper pages based on ground truth extracted from Beyond Words. Besides, the ResNeXt part of the model was transferred from pre-trained ImageNet ResNeXt-101 to reduce training parameters. Based on dhSegment, using transfer learning is able to boost training effectiveness, and preserve a good performance. The second step, a text segmentation, and recognition model retrieved textual content in the graphic snippets (i.e. extracted graphic snippets from the first step). Specifically, EAST text detection was applied to find text regions for an OCR process to retrieve words within graphic snippets. And the retrieved word was encoded into metadata for further usages, such as search queries.

In the second iteration, we focus on *evaluating and improving the segmentation step using the U-NeXt model*. The U-NeXt model is an extension of dhSegment model. The dhSegment model used pre-trained ResNet while our U-NeXt used pre-trained ResNeXt model. Note that ResNeXt is an improved version of ResNet. In the study of dhSegment on the Beyond Words collection, the classification accuracy was 88% and the mean intersection over union (mIoU) was 26%. The U-NeXt is expected to have a better performance than dhSegment. Further, the EAST text detection largely depends on the performance of the segmentation step. Hence, improving segmentation step is a key component of this project.

Note that two metrics are used to evaluate the performance of this project. First, the classification accuracy is a pixel-wise accuracy. It computes the percentage of correctly labeled pixels to total numbers of pixels for each class. Second, the mIoU evaluates whether the predicated region accurately covers the true region in the ground-truth overall classes.

2. Datasets

Two datasets were used to train and evaluate the segmentation step using the U-NeXt model, the Beyond Words collection and the European Newspapers collection. In the Beyond Words collection (BW), some graphic regions appeared on a page are missing in the ground-truth. And the marked region in the ground-truth does not tightly map to the actual shape of the graphic region. This lack of reliable ground-truth in the BW collection led us to pre-training our model on a more comprehensively labeled dataset called the European Newspapers collection (ENP). By doing so, during training, some local minimum, created by the aforementioned issues in the Beyond Words collection, could be avoided. Specifically, the ENP contains 480 images in total, in which, there are 384 images in training set and 96 images in validation set. And the BW contains 1,532 images in total, in which, there are 1,226 images in training set and 306 images in the validation set.

Further, the similarity shared by ENP and BW collections is the crucial reason why the ENP can be used for pre-training. First, both ENP and BW collections are document images that are digitized from newspapers. Hence, they share a similar content layout and density. Second, the ground-truth of the ENP marked five classes: (1) background, (2) text, (3) figure, (4) layout separator, and (5) table; while the ground-truth of the BW marked background and five detailed types of figures. Hence, the learned

knowledge from the pre-trained model on figures using the ENP provides a good reference for U-NeXt to identify the figure region. Then, the fine-tuning using the BW could focus on detailed figure type differentiation than identify the figure region.

3. Experimental Results

In this experiment, early stopping is not applied since the expectation of the performance is unknown. Here we report on two sets of results: from the pre-training experiment and from the fine-tuning experiment:

- The pre-training experiment involved training and testing on the ENP dataset (up to 700 iterations).
- The fine-tuning experiment involved four different approaches.
  - The first approach trained and tested on the BW dataset without using the ENP-trained classifier. This is meant to serve as a baseline design.
  - The second approach used the above ENP-trained classifier as the beginning classifier, and training and testing it on the BW dataset (up to 80 iterations). We added this design because using a pre-trained classifier for a similar task could help the fine-tuning experiment address the issue of lack of ground truth data mentioned in the previous section. This second approach is a variant of the first approach.
  - The third approach replaced the deconvolutional layer with a resizing layer in the deep learning model, and training and testing on the BW dataset. Since the deconvolutional layer is known to suffer from the "checkerboard" issue [Distill 2016], the resizing layer is seen as a potential improvement technique. This third approach is thus a variant of the first approach.
  - The fourth approach performed a two-class segmentation, instead of six classes on the BW dataset for both training and testing. This is because the training dataset is biased where there is a predominantly large number of background pixels compared to other classes of pixels[1]. By collapsing all the object pixels into one class, we hope to reduce the imbalance in the number of pixels in each class during training. This fourth approach is thus also a variant of the first approach.

3.1 Pre-training Experiment

Figure 1 shows the training performance of the pre-training experiment reaches 91.30% on pixel-wise accuracy and 57.19% on mIoU. And the testing performance is 81.90% on pixel-wise accuracy and 48.18% on mIoU. From the result, the convergence is observed (i.e., the tendency of accuracy gets close to 100% percent). The observed convergence indicates the parameters are getting trained to fit the task; hence, the model is ready for fine-tuning.

3.2 Fine-tuning Experiment 1: without pre-trained ENP classifier

Figure 2 shows the performance of the experiment without using the pre-trained ENP classifier reaching 89.08% on training pixel-wise accuracy, 50.43% on training mIoU, 80.11% on testing pixel-wise accuracy and 38.00% on testing mIoU. The experiment lasts 80 epochs, and the convergence is observed on both training and testing curves. However, the testing curve shows instability, that, although the tendency towards higher testing accuracy, the testing accuracy varies high and low rapidly during the experiment. And Table 2 (row 1 - 4) shows the class-wise testing performance on accuracies and mIoUs. The class-wise stats show that the classifier failed to recognize classes of editorial cartoons, illustrations, and maps. These

[1] There are 88.21% pixels in background class, but for the rest of classes, only 0.71% in editorial cartoon class, 2.89% in comics/cartoon class, 1.38% in illustration class, 6.64% in photograph class, and 0.18% in map class.

three classes happen to be the top three rarest classes. Hence, the misrecognition issue is likely caused by the rareness of corresponding classes. However, overall, the performance of the classifier is promising since both training and testing accuracies reached 80% within only 80 training epochs.

3.3 Fine-tuning Experiment 2: using pre-trained ENP classifier

Figure 3 shows the performance of the experiment using the pre-trained ENP classifier reaching 89.41% on pixel-wise training accuracy, 41.21% on training mIoU, 85.53% on testing accuracy and 38.57% on testing mIoU. Though the performance indicators above might look promising, upon further investigation, the classifier trained during the fine-tuning experiment attempted to classify as many pixels as background pixels after training convergence. Table 2 (row 5 - 8) shows the class-wise stats. We see that, after the convergence, all training and testing stats for non-background classes are zero. Hence, the performance stats are better than the first fine-tuning experiment numerically, but the actual performance is worse since none of the objective class was recognized. As previously mentioned, the background pixel is the majority over all pixels of the BW dataset. Such imbalance could create a "deep" local minimum. We suspect that the classifier fell into the "deep" local minimum. And it could not "jump" out from the minimum. In fact, the large fluctuations at the beginning epochs are indirect evidence. It shows that the classifier tried but failed to "jump" out from the minimum. However, the advantage of using pre-trained ENP classifier is the faster converging speed. Therefore, by taking such advantage, the computational resources could be saved comparing to others.

3.4 Fine-tuning Experiment 3: using resizing layer

Figure 4 shows that, for testing performance, the pixel-wise accuracy reached 86.69% and the mIoU reached 37.84%. The performance did not show clear improvement, because the pixel-wise testing accuracy is higher while the testing mIoU lower than the experiment 3.2. More, similarly, in the class-wise performance, shown in Table 2 (row 9 - 12), we also found that pixel-wise accuracy and mIoU of the editorial cartoon, illustration, and map classes are zeros. However, the curve in Figure 4 did not show the instability like experiment 3.2. Hence, the instability likely came from the "checkerboard" issue since the resizing layer was introduced to solve the issue. Therefore, from the perspective of stability, using the resizing layer has better performance than experiment 3.2.

3.5 Fine-tuning Experiment 4: combined two-class segmentation

Training a classifier to learn information from rare classes is very hard. Hence, combining five non-background classes into one class could decrease the complexity of the task, which could lead to improvements. In fact, pixels in non-background classes only 11.79% of the entire training dataset in total. And in this experiment, Figure 5 shows the combined class segmentation outperformed all other fine-tuning experiments. That is, for training performance, the pixel-wise accuracy was 91.76% and the mIoU was 71.44%; and, for testing performance, the pixel-wise accuracy was 88.89% and the mIoU was 64.97%.

Table 1 Average performance of fine-tuning experiments

	Without Pre-trained ENP Classifier		Using Pre-trained ENP Classifier		Using Resizing Layers		Combined Two-class Segmentation	
	Train	Test	Train	Test	Train	Test	Train	Test
Accuracy	89.08%	80.11%	89.42%	85.53%	88.90%	86.69%	91.76%	88.89%
mIoU	50.43%	38.00%	41.21%	38.57%	51.31%	37.84%	71.44%	64.97%

Table 2 Class-wise statistics of fine-tuning experiments

			Background	Editorial Cartoon	Comics/ Cartoon	Illustration	Photograph	Map
Without Pre-trained ENP Classifier	Train	Accuracy	92.70%	0.00%	10.66%	0.00%	92.11%	0.00%
		mIoU	90.81%	0.00%	7.00%	0.00%	54.46%	0.00%
	Test	Accuracy	84.43%	0.00%	44.82%	0.00%	72.38%	0.00%
		mIoU	79.99%	0.00%	24.97%	0.00%	52.09%	0.00%
Using Pre-trained ENP Classifier	Train	Accuracy	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		mIoU	89.42%	0.00%	0.00%	0.00%	0.00%	0.00%
	Test	Accuracy	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		mIoU	85.53%	0.00%	0.00%	0.00%	0.00%	0.00%
Using Resizing Layers	Train	Accuracy	90.87%	6.32%	41.85%	3.26%	88.22%	0.00%
		mIoU	89.46%	5.16%	29.07%	2.61%	47.90%	0.00%
	Test	Accuracy	97.83%	0.00%	4.19%	0.00%	40.60%	0.00%
		mIoU	87.38%	0.00%	0.20%	0.00%	34.24%	0.00%

			Background	Non-Background
Combined Two-class Segmentation	Train	Accuracy	91.02%	90.45%
		mIoU	90.22%	52.66%
	Test	Accuracy	92.82%	68.18%
		mIoU	86.64%	43.29%

4. Conclusion

In this second iteration of the figure/graph extraction task, we tested our proposed U-NeXt model during the first iteration of exploration. The pre-training stage used the ENP collection. Though the pre-training performance was promising, it was not very strong. In addition, the fine-tuning stage with several experiments used the BW collection as well as other improvement techniques reported in machine learning. The fine-tuning experiments showed evidence that the issue in BW collection affected the performance.

Further, according to the visualized extraction result, we found two widespread issues in the BW ground truth. First, the missing component issue appears to be quite widespread in the BW ground truth data. For example, shown in Figure 5, a large portion of a photograph in the document is missing from the ground truth, but is captured by our U-NeXt classifier. Second, there are inaccurate rectangular regions. For instance, shown in Figure 6, the ground truth region includes incorrectly a large portion of the text content. In future work, these issues are a good starting point for improving the BW ground truth.

However, we found a very interesting occurrence where the U-NeXt classifier tried to fit the exact shape of the figure/graph region. For example, shown in Figure 7, the classifier prediction tried to fit the exact shape of the eagle on the right-hand side of the newspaper page. We speculate that the light background of the figure/graph region might have confused the classifier. And this may suggest that the actual performance of the U-NeXt may be better than the statistical evaluation used in our experiments (i.e., pixel-wise accuracies and mIoU).

Hence, we propose two ways to continuously improve the performance of this figure/graph extraction task.

- First, splitting the figure/graph extraction task to a pipeline of two tasks: (1) extraction of graphics from the background and textual content and (2) classification of the extracted graphics to detailed graphic types. Such arrangement would reduce the complexity of the task.
- Second, there is still room to improve the U-NeXt model for the extraction task directly. For example, the resizing layer can improve the performance of our experiment.

At this stage, it is hard to say which of the above solutions would yield better results. They all have advantages. The model using a pre-trained ENP classifier converges faster; The resizing layer avoids the "checkerboard" issue and improves stability. And the combined class segmentation can decrease the task difficulty while the direct six-class segmentation can avoid introducing complexity from pipelining two tasks.

Furthermore, we found that, because the classifier tries to fit the exact shape of the graphical content, the actual classification performance may be higher than the statistical evaluation indicated. However, comparing to the issues from the U-NeXt model, the major problem is that the BW ground truth *has two widespread quality issues to be fixed*. We believe that performance improvement will be observed if the ground truth issues can be removed.

- Therefore, in the next iteration of this projection, work should also be done on the BW ground truth to fix the quality issues as well as the imbalance class issue. Specifically, increase the number of pixels for rare classes.

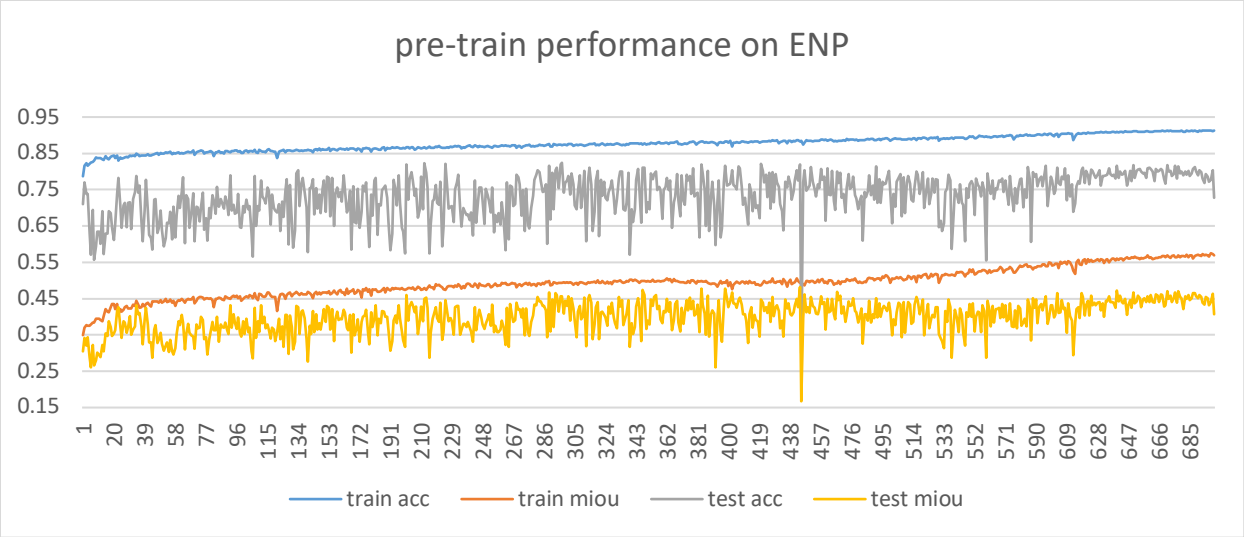


Figure 1 Pre-train ENP classifier performance.

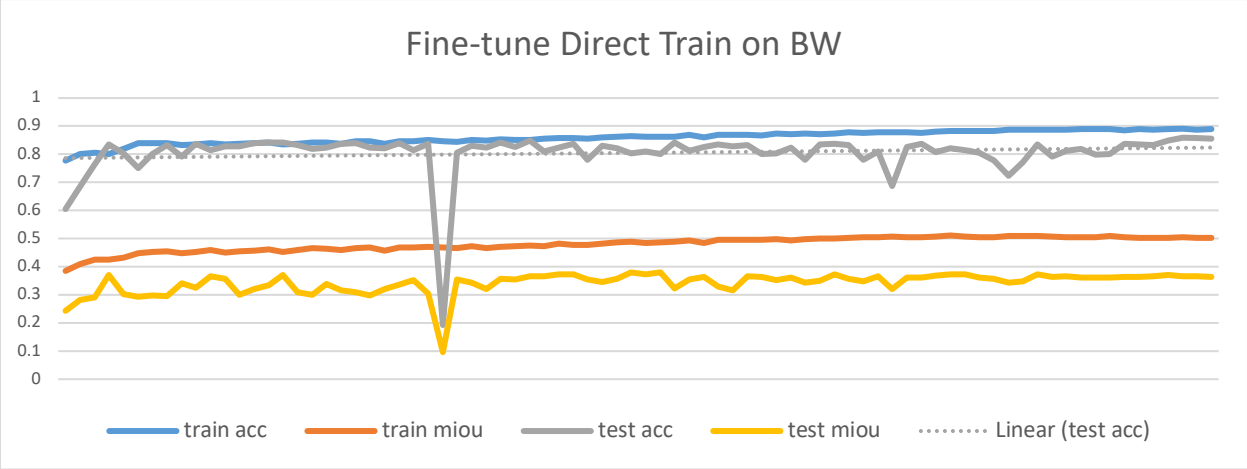


Figure 2 Fine-tuning experiment 1 - the baseline.

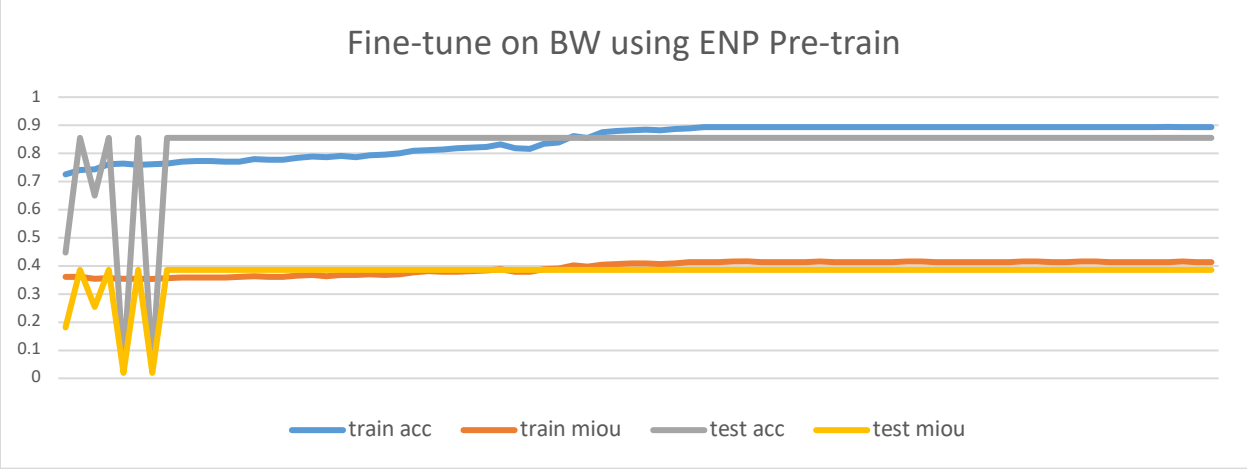


Figure 3 Fine-tuning experiment 2 - using pre-trained ENP classifier.

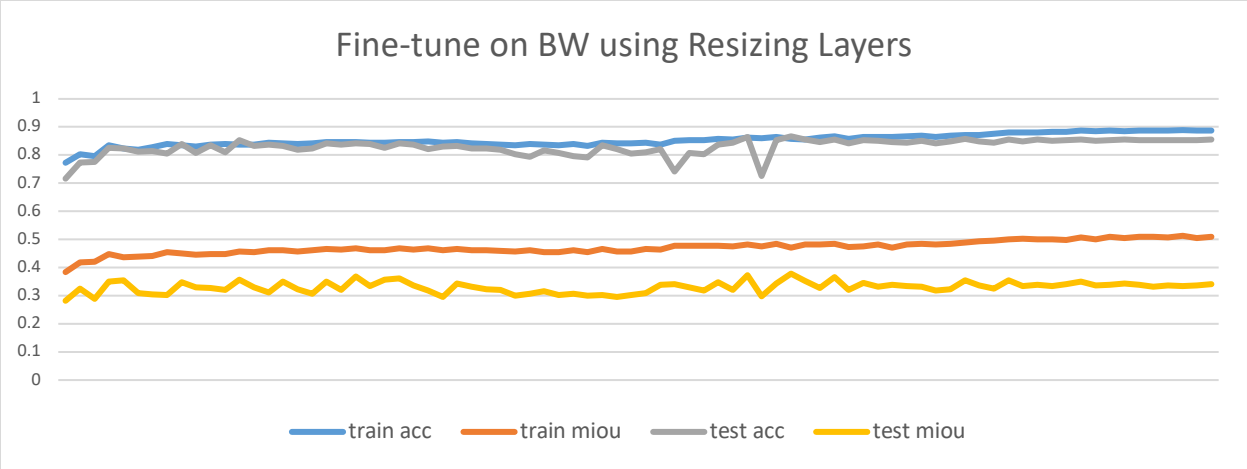


Figure 4 Fine-tuning experiment 3 - using resizing layers.

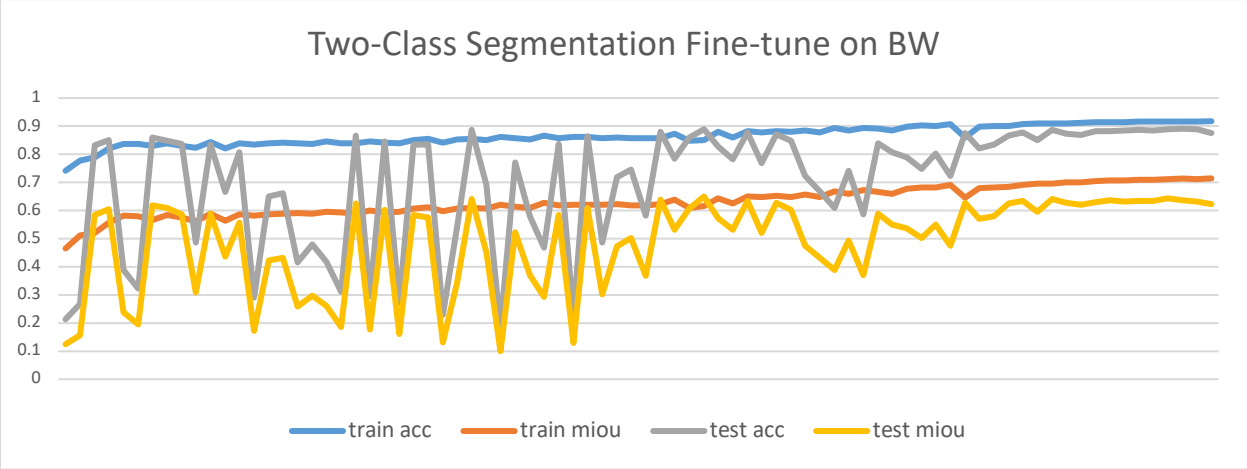
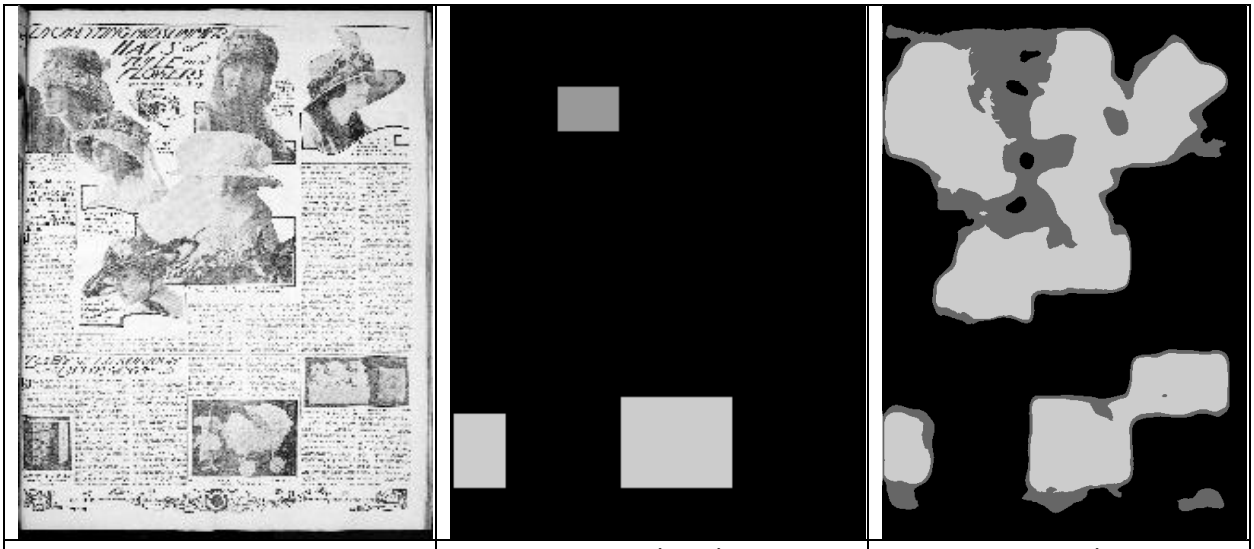
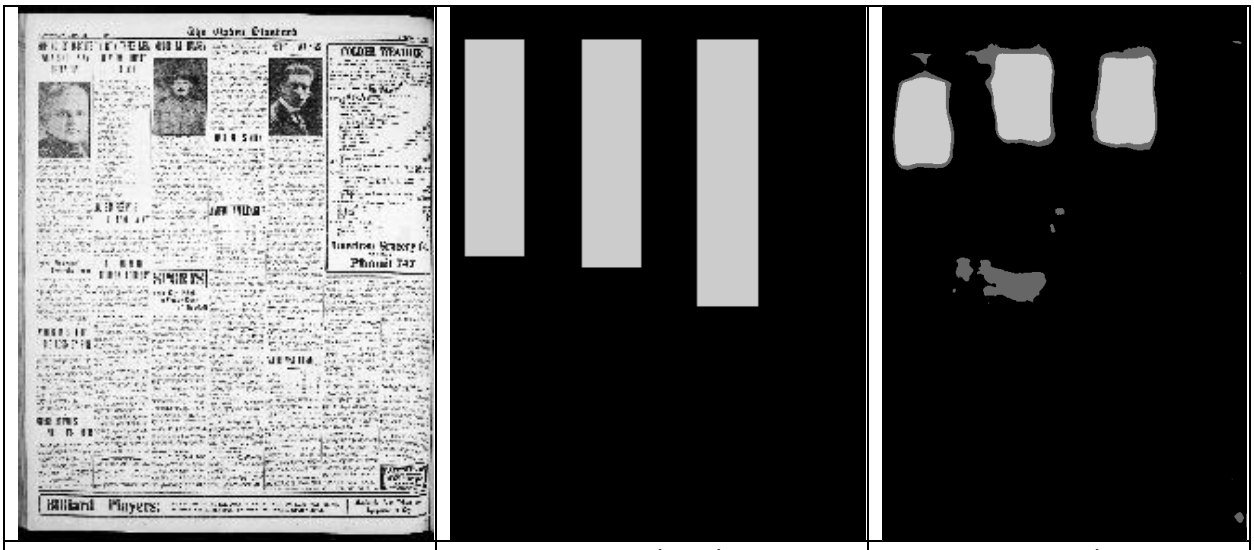


Figure 5 Fine-tuning experiment 4 - combined two-class segmentation.



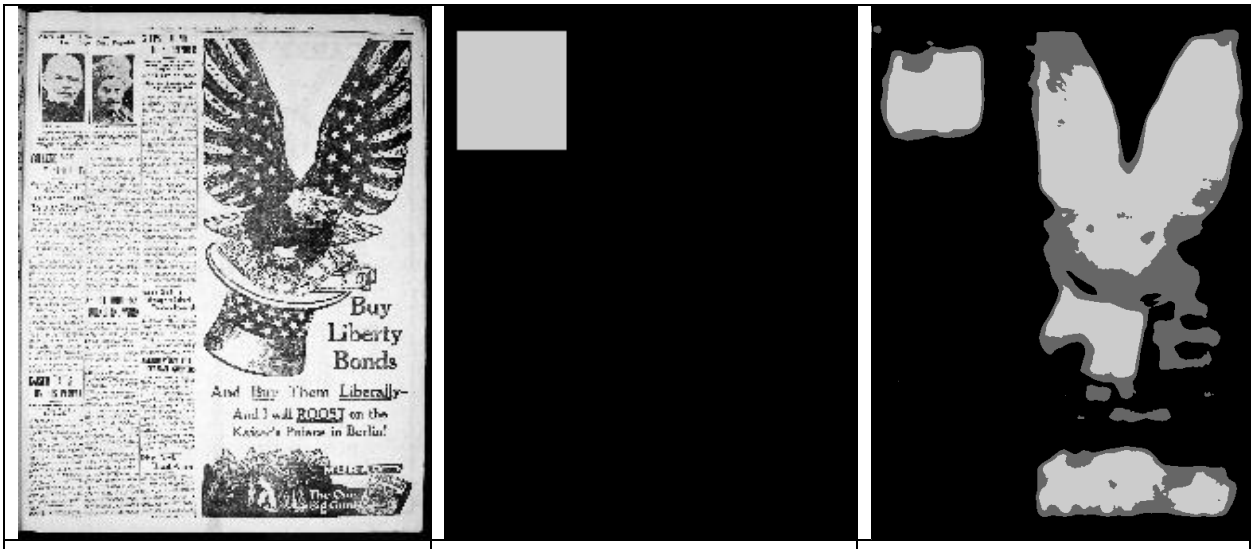
Newspaper page	Ground truth	U-NeXt Prediction

Figure 6 The missing component issue



Newspaper page	Ground truth	U-NeXt Prediction

Figure 7 The extra text content issue



Newspaper page	Ground truth	U-NeXt Prediction

Figure 8 Classifier tried to fit the exact shape of the graphic content

Progress Report - Analysis on Relationship between Document Difficulty Score and Visual Features

10/31/2019
Mike Pack

1. Introduction

As the 3rd iteration of Project 5 (Document Image Quality Assessment), in this experiment, we aim to reveal a relationship between a *difficulty score* and *visual features*. One of the expected beneficial outcomes from this experiment is to build a *difficulty score prediction model* based on the revealed relationship that would give the Library of Congress the capability of controlling and managing challenging document images, especially for human perception involved tasks such as transcription.

2. Dataset

The dataset used in this experiment is a subset of document images (15,592 images) collected from the Library of Congress archive along with corresponding difficulty score.
The difficulty scores here—collected by Library of Congress—is the number of trials on transcription by human volunteers based on the intuition behind that poorly readable document images due to various visual artifacts (e.g., noise or ugly handwriting) would have a higher number of resubmissions by multiple transcription volunteers. Note here that the scores are not verified by the human experts.

3. Experiment 1: Visual Inspection

Before directly diving into the numerical correlation analysis between visual features and difficulty scores, we first visually inspect a handful of images to investigate to what extent the difficulty scores reflect the human perception of difficulty, particularly for transcription-like tasks. Particularly, we focus on finding any notable visual cues that makes distinctive differences between different difficulty scores.
To this end, we sampled two images (i.e., acceptable and not acceptable for human perception of difficulty to the difficulty score) from two different types (i.e., *handwritten* and *typed* document images) for six different difficulty scores as shown in Table 1. From the inspection, we found the following two observations:

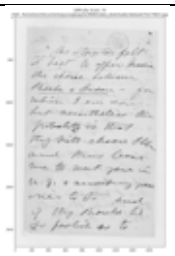
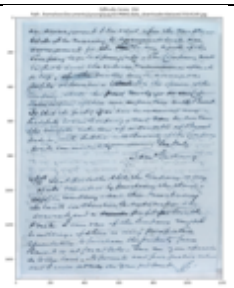
Observation 1. The same *visual feature* deemed to related to the *difficulty score* in *typed* documents is not deemed to related to the difficulty score in *handwritten* documents.

Observation 2. It is hard to expect to find correlation between a simple standalone *visual feature* (e.g., number of characters or low contrast) and a *difficulty score.*

For a better comprehensive understanding of the above observation, consider the following examples. About the first observation, for the *typed* documents, note that the amounts of contents/characters in an image (i.e., density) is subtly deemed to related with the difficulty score (see the rightmost column in Table 1), meanwhile it is not the case for the *handwritten* documents (see the third column in Table 1.)
About the second observation, note that it is hard to find notable visual similarity between *not acceptable* images and *acceptable* images within the same difficulty score. For example, document images with the difficulty score of 9, *not acceptable handwritten* image show poor image quality in terms of low contrast and higher density compared to *the acceptable handwritten* image, as shown in Table 1. This is also the case in the *typed* document images.

Table 1. Document samples for different difficulty scores. The empty cells meaning no more images exist for the corresponding difficulty score.

Difficulty Score	**Not Acceptable**		**Acceptable**	
Type	**Handwritten**	**Typed**	**Handwritten**	**Typed**
9	Low contrast + comparably large amounts of contents, but ONLY 9	Complicated layout + decent amounts of contents, but ONLY 9		
20	Range-effect, but comparably small amounts of contents, but 20	Amounts of characters looks similar to 9		

70	 Small amounts of contents, but ONLY 70	 Amounts of characters is way small		
135	 Low contrast, but relatively pretty writing, but 135	 Looks quite similar to the difficulty score of 9 or 20 images		
350	-	-	 Bleed-through + Ugly writing	
748 (handwritten) 3064 (typed)	-	-	 Ugly writing + Medium contents	

From the above observations, we can set the following two assumptions:

Assumption 1. A feature indicating whether an image is *handwritten* or *typed* seems promising to be somewhat related with the *difficulty score.*

Assumption 2. It is necessary to find more high-level visual features (e.g., expert knowledge-based engineered features or deep-features learned by a deep-learning model) hard to expect to find a correlation between a simple standalone *visual feature* (e.g., number of characters or low contrast) and the *difficulty score.*

4. Experiment 2: Pearson's Correlation

In this experiment, we perform numerical analysis to find a set of visual features showing a meaningful correlation to the difficulty score using the Pearson's Correlation.

A set of features here is low-level visual features obtained by relatively simple image processing techniques, such as contrast measure or counting the number of connected components (i.e., letters or characters.)

Along with these low-level visual features, based on the above two assumptions, we added four additional high-level visual features: (1) *prediction*, (2) *density*, (3) *number of zones*, and (4) *zone size abnormality*.

First, the *prediction* feature is a categorical value indicating whether the type of document image is *handwritten*, *typed*, or *mixed*. This feature is obtained by our deep-learning-based document type prediction model developed in our Project 2, which showed promising classification performance with 0.9 of f1-score (best value at 1 and worst value at 0.)

Second, the *density* feature measures how dense the document is by considering the area of non-background regions. This feature is obtained by dividing the area of non-background regions by the resolution of the image.

Third, the *number of zone* feature represents how many zones (i.e., visually homogeneous regions) are presented in the image. This feature is obtained by segmenting the image by our deep-learning-based document segmentation model developed in our Project 1, which showed promising segmentation performance with 0.7 of mIoU (best value at 1 and worst value at 0.)

Fourth, the *zone size abnormality* feature measures the size of zones and calculates the degree of outliers. This feature is obtained by counting the number of outliers in terms of zone size and divide it by the resolution of the image for the normalization purpose. The intuition behind this feature is that the output of our segmentation algorithm tends to generate the relatively regular and uniform size of zones for straight forward and clear document images whereas it tends to generate a number of abnormal size of zones (i.e., extremely small zones and big zones simultaneously) for noisy document images.

After obtaining the whole visual features, before conducting Pearson's correlation, we carry out histogram analysis to visually inspect how images are distributed for each visual feature, as shown in Figure 1. From this analysis, we can observe that some visual features that follow a normal distribution at a certain level, such as *density*, *contrast*, and the *number of letters*. Note that one assumption behind Pearson's correlation is that variables (i.e., visual features) should be normally distributed. In this regard, we can expect that those three features are likely to show relatively high correlation coefficient values. We can observe that this expectation does actually match with the result of Pearson's correlation, as shown in Table 2.

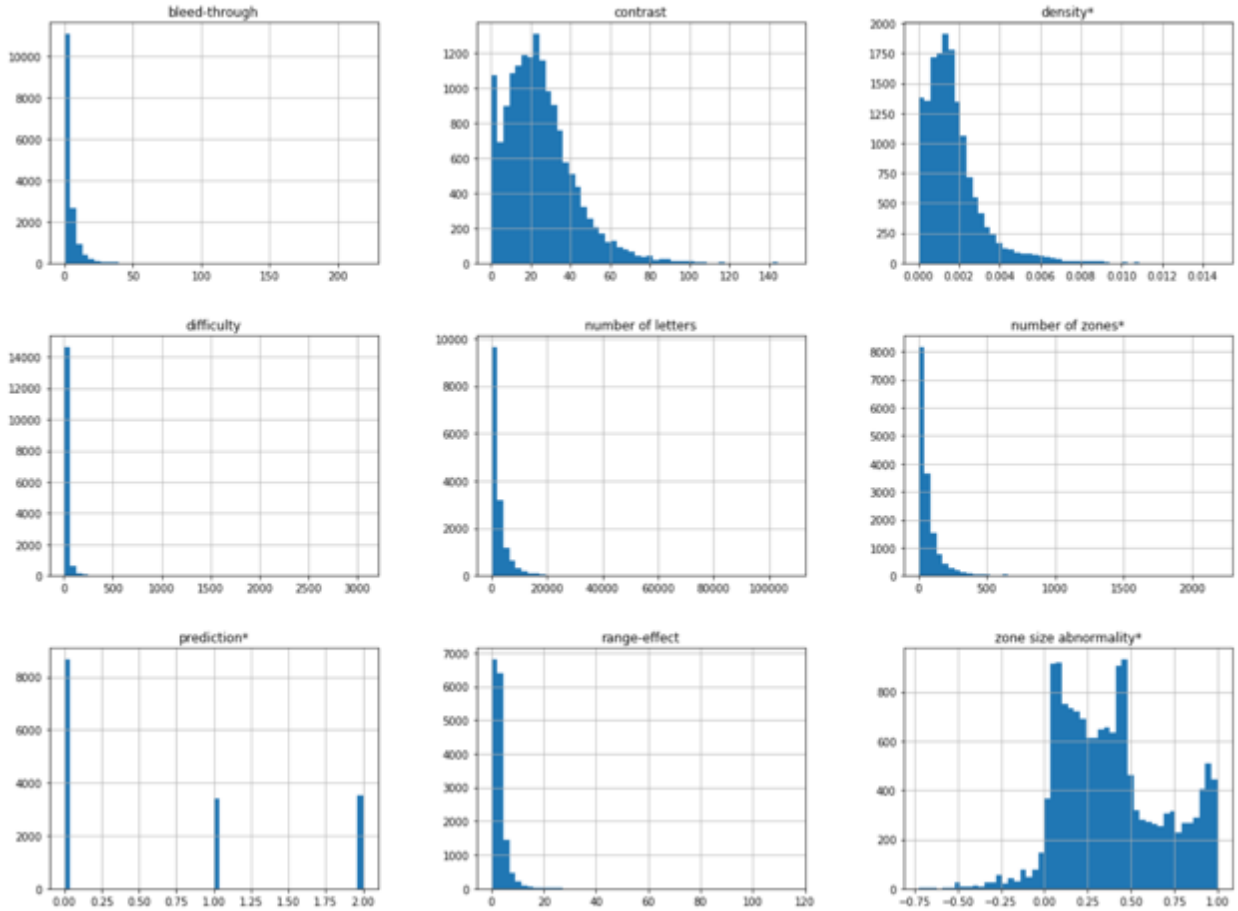


Figure 1. Distribution of visual features over 15,592 images. Note that one of the assumptions behind the Pearson's correlation coefficient is that variables (i.e., visual features) should be normally distributed.

Table 2. The size of correlation for various visual features. Note that a visual feature with asterisks (*) is high-level engineered features using low-level visual features.

Visual features	Size of correlation
Density*	0.16
Contrast	0.15
Number of Letters	0.15
Number of Zones*	0.10
Zone Size Abnormality*	0.07
Bleed-through	0.03
Range-effect	0.02
Prediction*	0.01

It is worth noting that there is no rule for determining what size of correlation is considered strong, moderate, or weak. The interpretation of the coefficient depends, in part, on the topic and context of the study. When we are conducting research that is difficult to measure, in our case, the difficulty of the document image in the context of human perception, we should expect the correlation coefficient to be lower. With this in mind, we can interpret this result as follows.

First, we can observe that one of our high-level engineered visual features, *density*, shows the highest size of correlation. This implicates that if we end up finding a more sophisticatedly engineered visual feature, its size of correlation will superior to the low-level visual features by a large margin.

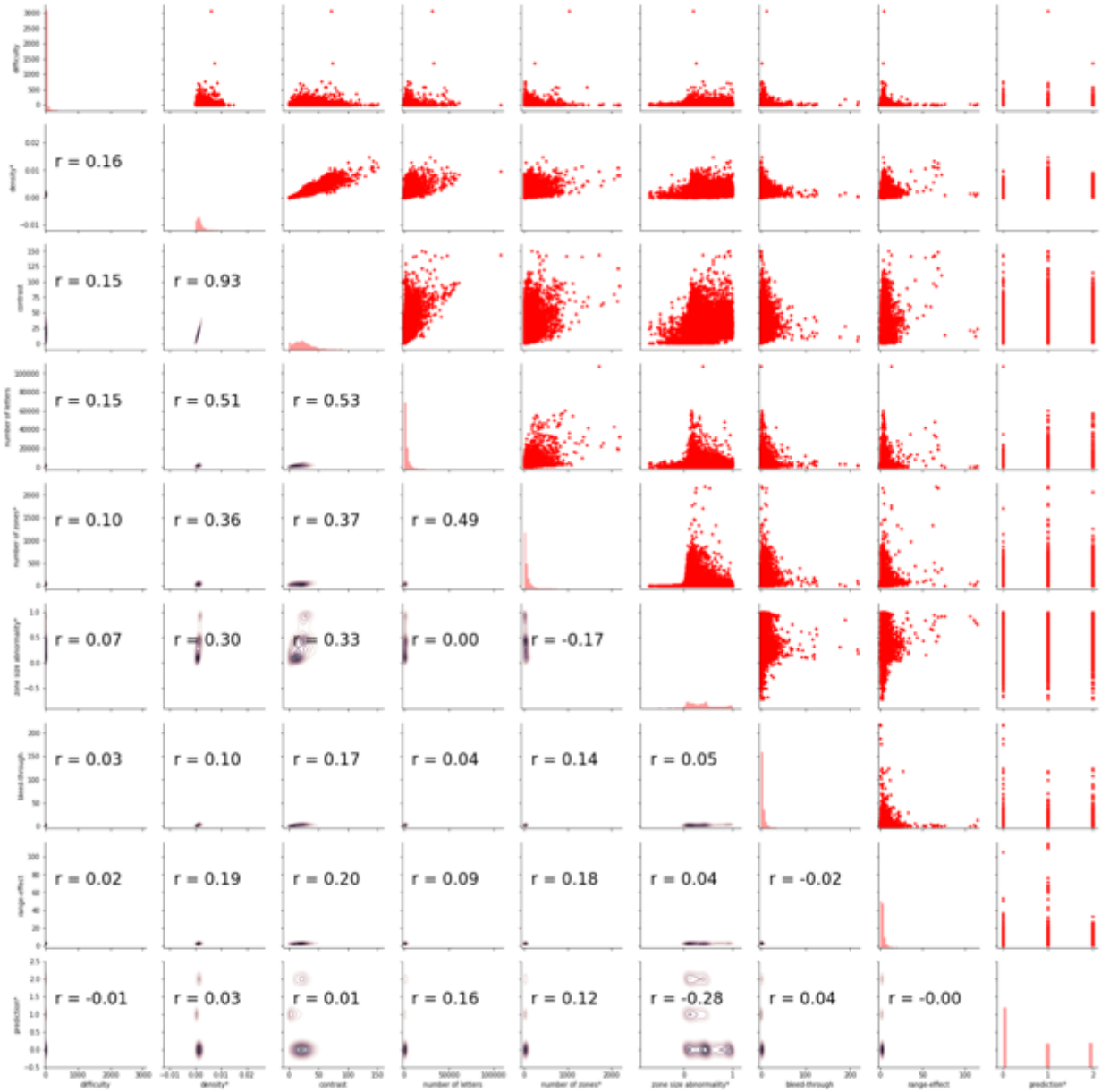


Figure 2. Pairs plot to visualize scattered distribution and relationship between two visual features. The *r* value represents the size of correlation (range from -1 to 1, best value at -1 or 1, worst value at 0.) It is worth noting that the relationship between most of the *visual features* with the *difficulty score* is not linear (first row). Especially, the *prediction* and the *difficulty score* does not show any linear relationship (top right cell.)

Second, a standalone *prediction* feature shows a very weak (or even neglectable) correlation. This result is expected since its distribution (see Figure 1) does not follow the normal

distribution. Also, since the Pearson's correlation is limited to reveal a "linear relationship" between variables, if there is a non-linear relationship between variables (see Figure 2), the size of correlation can be very low. In this regard, based on our Assumption 1 and 2, we expect that this *prediction* feature should be combined with other variables in a non-linear way, for example, by using the polynomial regression or support vector machine, to reveal the correlation to the *difficulty score*.

5. Conclusion

In this experiment, we show that both low-level and high-level engineered visual features are capable of capturing a certain level of correlation to the difficulty score. However, as shown in the pairs plot, most of the visual features rarely show any linear relationship with the *difficulty score* (see the first row in Figure 2). From this outcome, we can think of two future directions to reveal a more comprehensive understanding of the relationship between visual features and the difficulty score.

First, instead of low-level or engineered visual features, we can explore deep features, which is learned by a neural network model. Because of non-linearity property inherent in the neural network model, the features extracted by the model are known to be significantly high-level non-linear property.

Second, we can explore models that are capable of dealing with non-linear data, such as polynomial regression, support vector machine, or neural network. These models are mapping the low-level features into high dimensional space, which has an effect of embedding non-linearity property and the interaction between different low-level visual features, and we can expect a better understanding of the relationship between *visual features* and the *difficulty score*.

APPENDIX 2: WORK-IN-PROGRESS PRESENTATIONS

Document Images & ML

A COLLABORATORY BETWEEN THE LIBRARY OF CONGRESS AND THE IMAGE ANALYSIS FOR ARCHIVAL DISCOVERY (AIDA) LAB AT THE UNIVERSITY OF NEBRASKA, LINCOLN, NE

Overview of Projects

- Project 1: Document Segmentation (Mike & Yi)
- Project 2: Document Type Classification (Mike & Yi)
- Project 3: Quality Assessment (Yi)
- Project 3.1: Figure/Graph Extraction from Document (Yi)
- Project 3.2: Text Extraction from Figure/Graph (Yi)
- Project 4.1: Subjective Quality Assessment (Yi) (Work In Progress)
- Project 4.2: Objective Quality Assessment (Yi)
- Project 5: Digitization Type Differentiation: Microfilm or Scanned (Yi)

Background | State-of-the-Art CNN models

- **Convolutional Neural Network** (CNN) Models (deep learning)
  - Classification [Dataset; Top-1 / Top-5]
    - 2014, VGG-16 (Classification) [ImageNet; 74.4% / 91.9%]
    - 2015, ResNet-50 (Classification) [ImageNet; 77.2% / 93.3%]
    - 2018, ResNeXt-101 (Classification) [ImageNet; 85.1% / 97.5%]
  - Segmentation [Dataset; Intersection-over-Union (IoU)]
    - 2015, U-net (Segmentation/Pixel-wise classification) [ISBI; 92.0%]

- So, we now know that CNNs achieve *remarkable* performances in both classification and segmentation tasks.
- ***What about document images then?***

Project 1: Document Segmentation

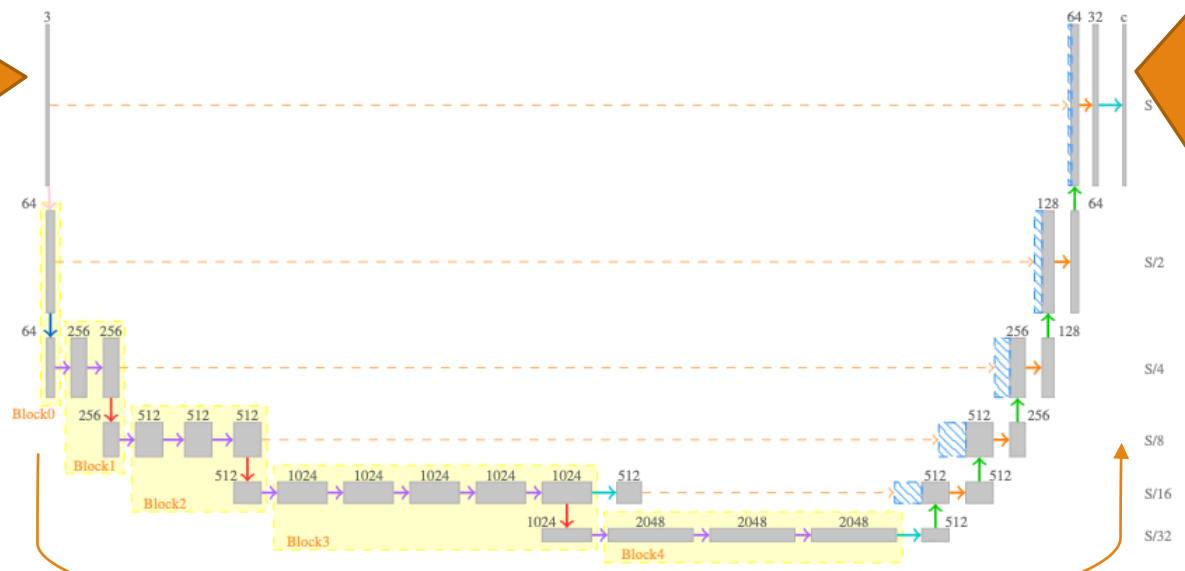
Objectives | Find and localize *Figure*/*Illustration*/*Cartoon* presented in an image

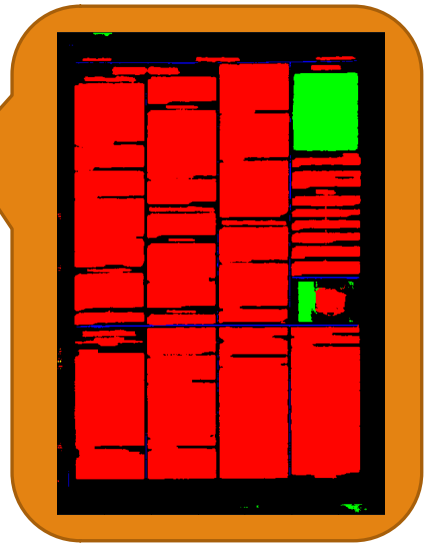
Applications | metadata generation, discover-/search-ability, visualization, etc.

Document Segmentation | Technical Details

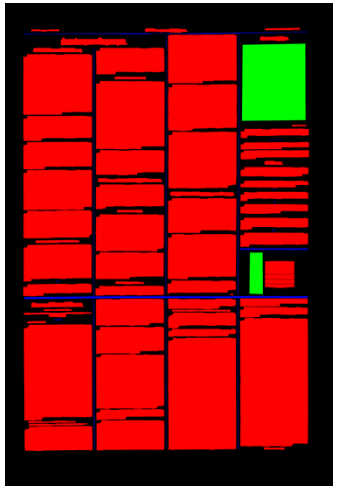
- ***Training*** is a process of finding the <u>optimal value weights between artificial neurons</u> that minimizes a pre-defined ***loss*** function

Input





Prediction



Ground-truth

1. Convolution & Down-sampling:
understand "***WHAT***" is present in the image
(i.e., feature extraction)

2. Up-sampling:
understand "***WHERE***" it is present in the image

3. Calculate per-pixel loss
4. Update weights between neurons
5. Repeat the process

Document Segmentation | Dataset

Beyond Words

- Total of 2,635 image snippets from 1,562 pages (as of 7/24/2019)
  - 1,027 pages with single snippet
  - 512 pages with multiple snippets
- Issues
  - Inconsistency (Figure 1)
  - Imprecision (Figure 2)
  - Data imbalance (Figure 3)

Figure 1. Example of inconsistency. Note that there are more than one image snippets in the left image (i.e. input) while there is only a single annotation in the right ground-truth.

Figure 2. Example of imprecision. From left to right: (1) ground-truth (yellow: Photograph and black: background) and (2) original image. Note here that in the ground-truth, non-photograph-like (e.g., texts) components are included within the yellow rectangle region.

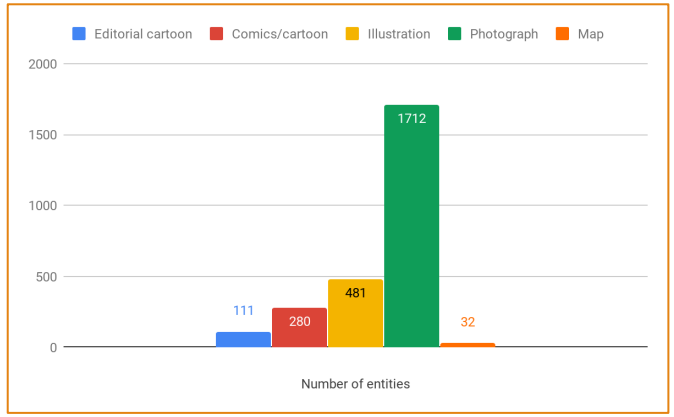


Figure 3. Number of snippets in Beyond Words. Note here the data imbalance

Document Segmentation | Dataset

European Historical Newspapers (ENP)

- Total of 57,339 image snippets in 500 pages
  - All pages have multiple snippets
- Issues
  - Data imbalance
    - Text: 43,780
    - Figure: 1,452
    - Line-separator: 11,896
    - Table: 221

Figure 4. Example of image (left) and ground-truth (right) from ENP dataset. In the ground-truth, each color represents the following components: (1) black: background, (2) red: text, (3) green: figure, (4) blue: line-separator, and (5) yellow: table.

Document Segmentation | Experimental Results

- A U-net model trained with ENP dataset shows better segmentation performance than that with Beyond Words in terms of pixelwise-accuracy and IoU score
  - IoU score is a commonly used metric to evaluate segmentation performance
  - The three issues—inconsistency, imprecision, and data imbalance—of Beyond Words dataset need to be improved for better use in training

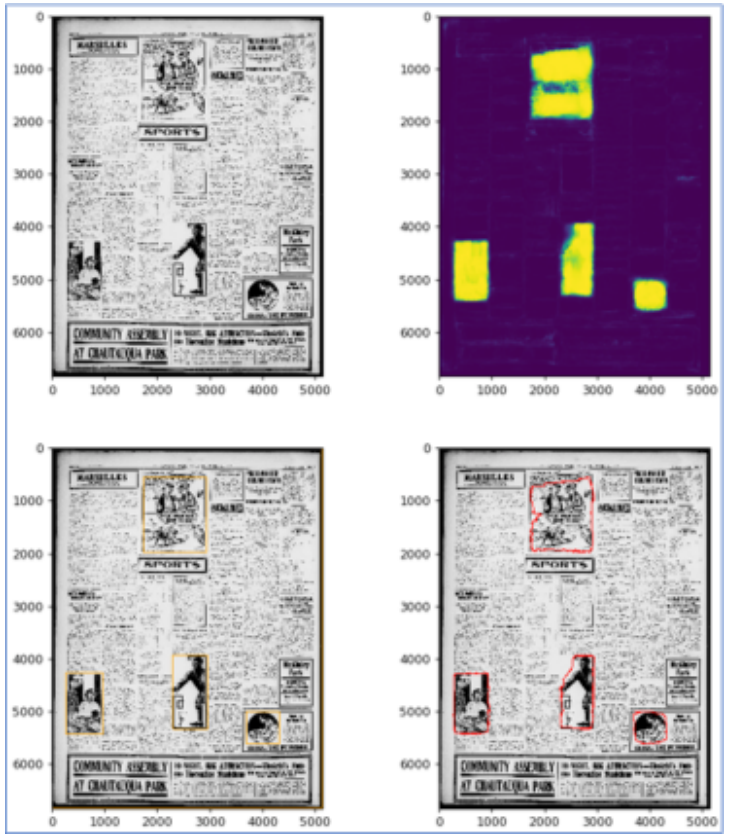
Model	train/eval size	Classes	Weighted training	Pre-processing (Normalization)	Best Score	
					Accuracy	mIoU
BW_1500_v1	1226/306	0: Background 1: Editorial cartoon 2: Comics/cartoon 3: Illustration 4: Photograph 5: Map	No	No	0.87	0.24
BW_1500_v2			Yes [10;22;20;18;8;22]		0.88	0.26
ENP_500_v1	385/96	0: Background 1: Text 2: Figure 3: Separator 4: Table	Yes [5;10;40;10;35]	No	0.88	0.64
ENP_500_v2				Yes	0.89	0.64
ENP_500_v3			**No**	**No**	**0.91**	**0.69**
ENP_500_v4				**Yes**	**0.91**	**0.69**

*Accuracy: Pixel-wise accuracy.
*mIoU: Average intersection over union.
*Normalization: Zero mean unit variance

- Assigning different weights per class to mitigate data imbalance did *not* show performance improvement
  - ***Future Work:*** Explore a different way of weighting strategy to mitigate a data imbalance problem

1

Document Segmentation | Potential Applications 1



- Enrich page-level metadata by cataloging the types of visual components presented on a page
- Enrich collection-level metadata as well
- Visualize figures' locations on a page

Figure 5. Segmentation result of ENP_500_v4 on Chronicling America image (sn92053240-19190805.jpg). Clockwise from top- left: (1) Input, (2) probability map for figure class, (3) detected figures in polygon, and (4) detected figures in bounding-box. In the probability map, pixels with higher probability to belong to figure class are shown with brighter color.

Document Segmentation | Potential Applications 2

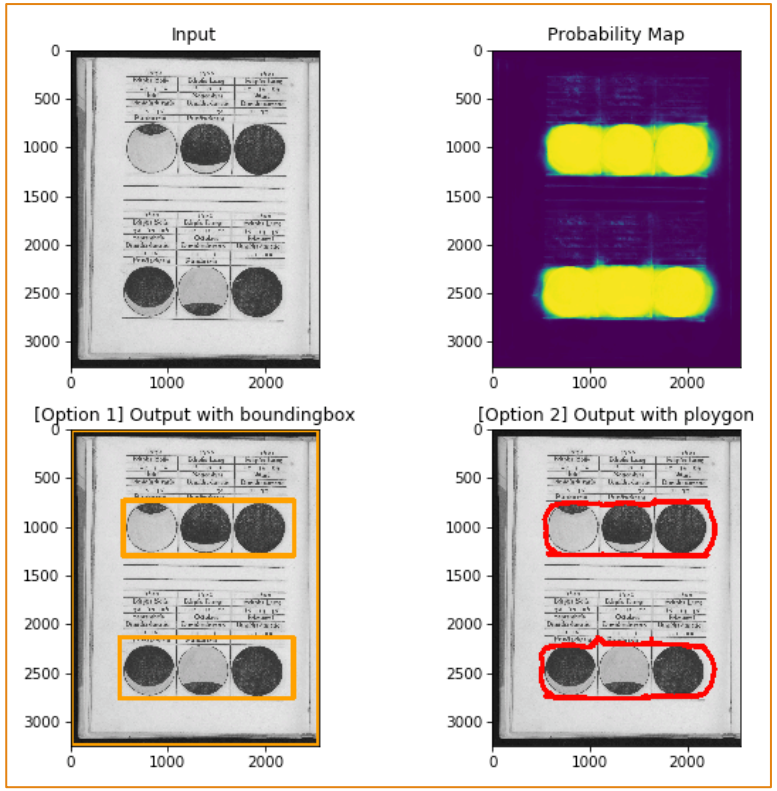


Figure 6. Successful segmentation result of ENP_500_v4 on book/printed material (https://www.loc.gov/resource/rbc0001.2013rosen0051/?sp=37).

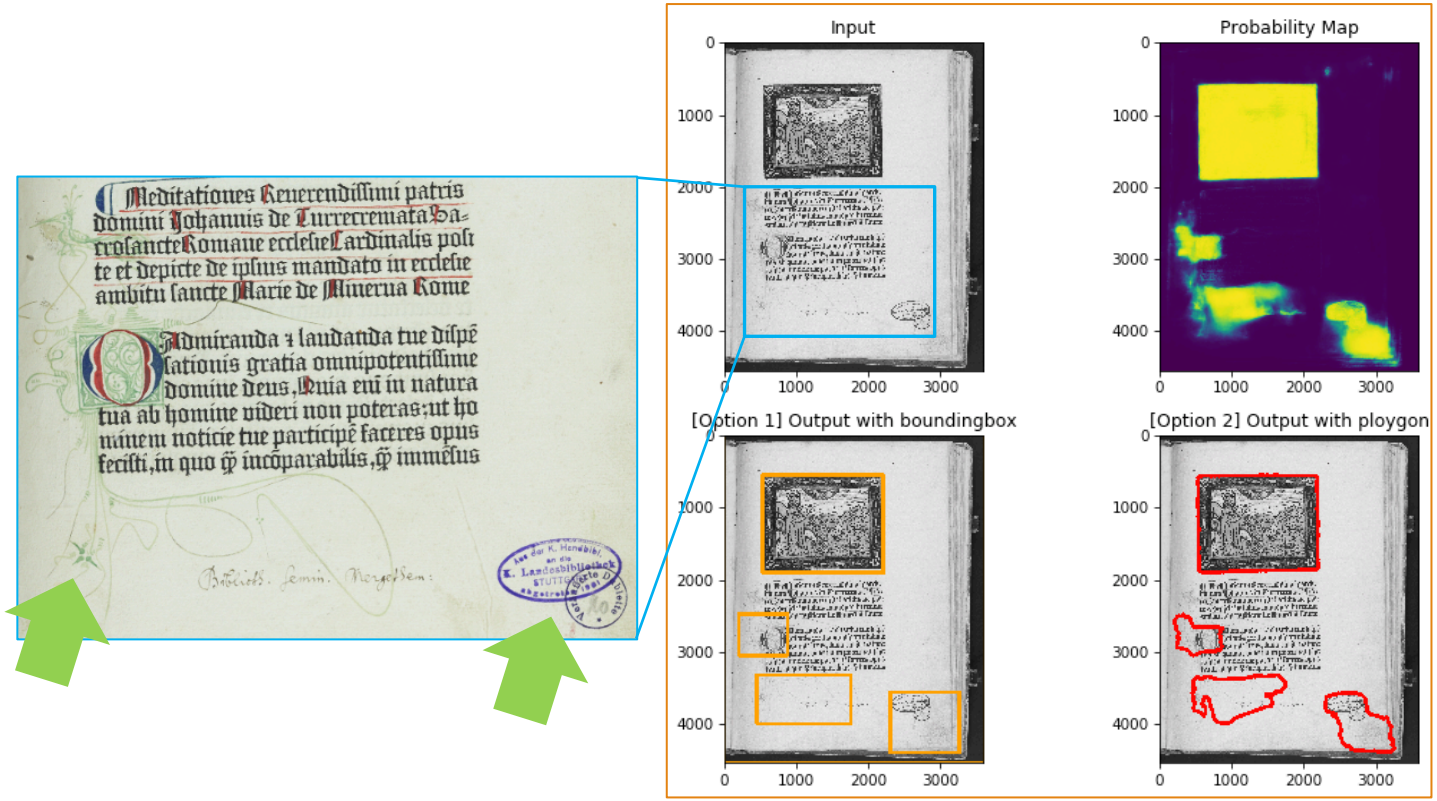


Figure 7. Failure segmentation result of ENP_500_v4 on book/printed material (https://cdn.loc.gov/service/rbc/rbc0001/2010/2010rosen0073/0005v.jpg). Note that there is light drawing or stamps (marked in green arrows) on the false positive regions.

Document Segmentation | Conclusions

- As a preliminary experiment, a state-of-the-art CNN model (i.e., U-net) shows promising segmentation performance on ENP document image dataset,
  - There is still room for improvement with more sophisticated training strategies (e.g., weighted training, augmentation, etc.)
- To make Beyond Words dataset more as a valuable training resource for machine learning researchers, we need to address the following issues:
  - Consistency
  - Precision of the coordinates of regions

Project 2: Document Type Classification

Objectives | (1) Classify a given image into one of *Handwritten*/*Typed*/*Mixed* type; (2) Classify a given image into one of *Scanned*/*Microfilmed*

Applications | metadata generation, discover-/search-ability, cataloging, etc.

Document Type Classification | Technical Details

*Note that we do not need up-sampling in this task, since **WHERE** is not our concern*

- A simple VGG-16 is used (Figure 8)
  - Afzal et al. reported that most of state-of-the-art CNN models yielded around 89% of accuracy on document image classification task
- **Transfer learning?**
  - Why don't we initialize our model's weights from a model that has been already trained on a large-scale data, such as *ImageNet* (about 14M images)?
  - ***Why?*** (1) training a model from the scratch (i.e., the value of weights between neurons are initialized to random number) takes too much time; (2) we have too small a dataset to train a model

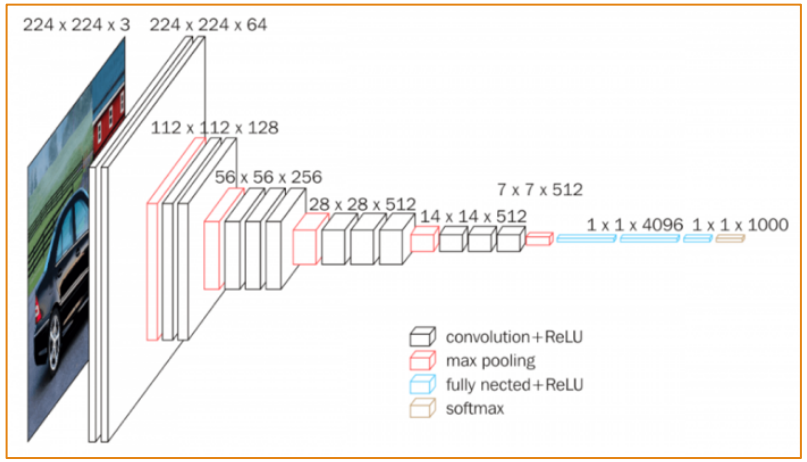


Figure 8. Architecture of original VGG-16. In our project, the last softmax layer is adjusted to have a shape of 3, which is the number of our target classes; handwritten, typed, and mixed

Afzal, M. Z., Kölsch, A., Ahmed, S., & Liwicki, M. (2017, November). Cutting the error by half: Investigation of very deep CNN and advanced training strategies for document image classification. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*(Vol. 1, pp. 883-888). IEEE.

Document Type Classification | Datasets

- We have two datasets:
  - Experiment 1: *RVL-CDIP* (400,000 document images with 16 different balanced classes); publicly available
  - Experiment 2: *suffrage_1002* (1,002 document images with 3 different balanced classes); manually compiled from ***By the People: Suffrage*** campaign (Table 1)

	handwritten	typed	mixed	Total
train	267	267	267	801
validation	33	33	33	99
test	33	33	33	99
Total	333	333	333	999

Table 1. Configuration of *suffrage_1002* dataset.

Figure 9. Example document images from each 16 different classes

Document Type Classification | Datasets

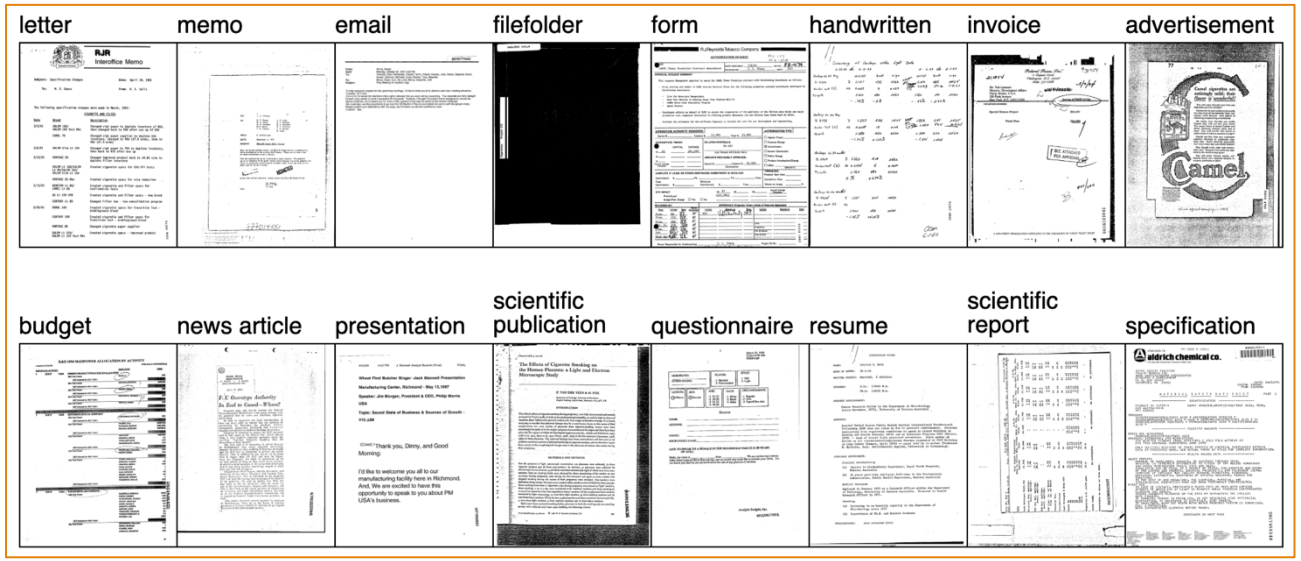


Figure 9. Example document images from each 16 different classes in *RVL_CDIP* dataset

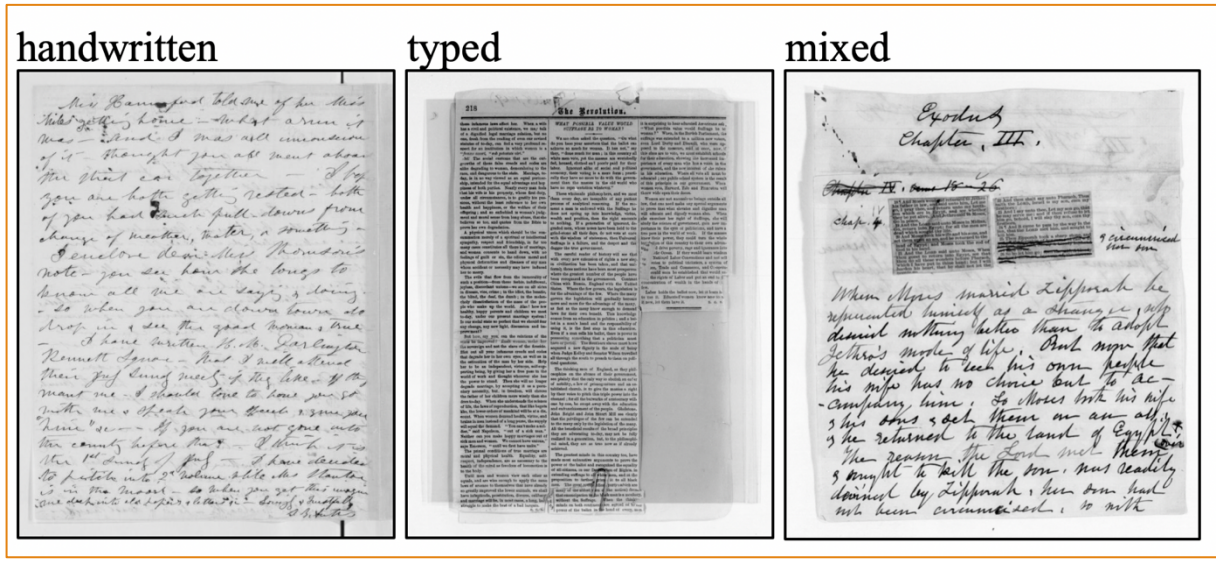


Figure 10. Example document images from each 3 different classes in *suffrage_1002* dataset

Document Type Classification | Experimental Results

Table 1. Precision, recall, and f1-score of *VGG-16* trained on *RVL_CDIP* dataset. The alphabetic labels are corresponding to the following labels: *letter*, *form*, *email*, ***handwritten***, *advertisement*, *scientific report*, *scientific publication*, *specification*, *file folder*, *news article*, *budget*, *invoice*, *presentation*, *questionnaire*, *resume*, and *memo*. Our class of interest, ***handwritten***, is bolded.

(unit: %)	A	B	C	**D**	E	F	G	H	I	J	K	L	M	N	O	P	Avg
Precision	86	74	98	**89**	89	73	90	88	89	92	87	91	78	91	92	88	87
Recall	94	79	97	**96**	91	73	93	91	97	86	83	86	79	73	94	91	87
F1	86	77	97	**92**	90	73	91	90	93	89	85	88	79	81	93	90	87

Table 2. Precision, recall, and f1-score of *VGG-16* on *suffrage_1002* testing set.

(unit: %)	**handwritten**	**typed**	**mixed**	**Avg**
Precision	89	91	90	90
Recall	97	94	79	90
F1	93	93	84	90

- Experiment 1: We obtained a model trained on a large-scale document image dataset, *RVL-CDIP* with promising classification performance, as shown in Table 1
  - ***Implication***: Features learned from natural images (ImageNet) are general enough to apply to document images
  - Now we can utilize this model by retraining it with our own *suffrage_1002* dataset in Experiment 2
- Experiment 2: The retrained model shows even better classification performance, as shown in Table 2

Document Type Classification | Conclusions

- In both experiments, the state-of-the-art CNN model is capable of classifying document images with promising performance
  - ***Potential Applications***: help tagging an image type
- A main *challenge*: classifying a mixed type document image, as shown in Figure 11
  - ***Future Work:*** Perform a confidence level analysis to mitigate this problem
- ***Future Work:*** We expect that the classification performance can be further improved with a larger large-scale dataset

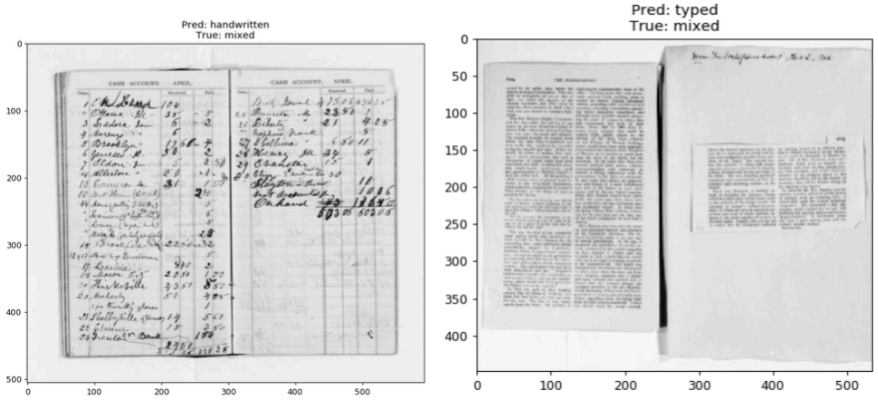


Figure 11. Failure prediction cases. On the left example, a typed region is relatively smaller than that of handwriting. On the right example, a handwriting region is relatively smaller than that of typing.

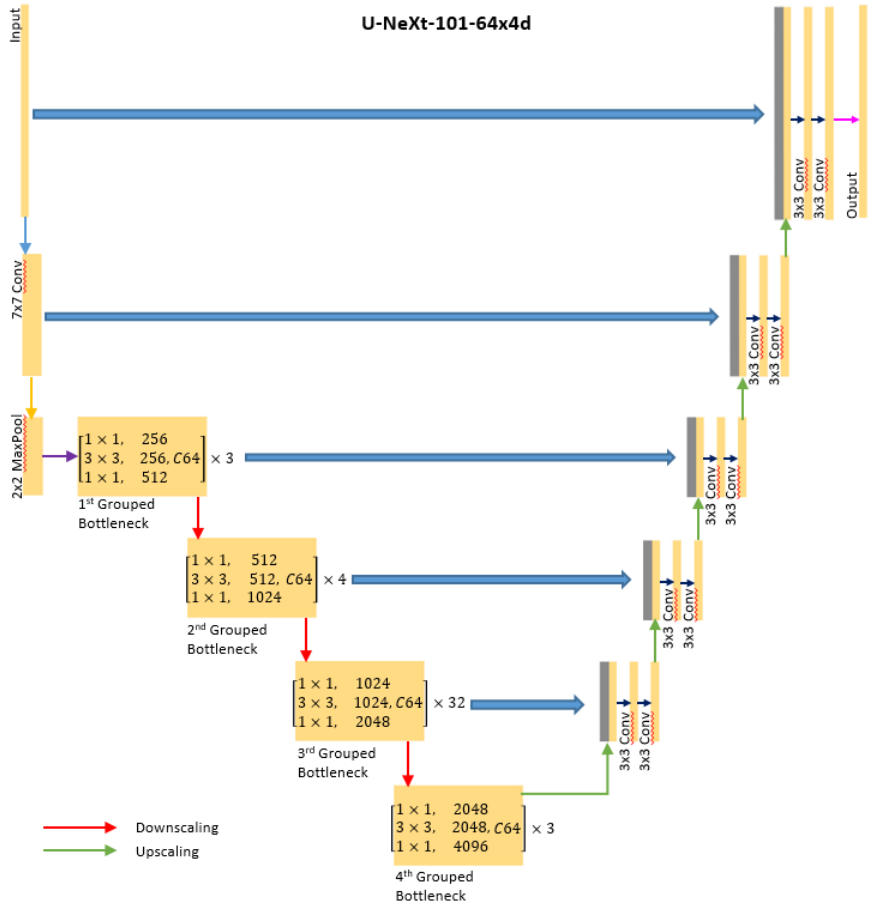
Afzal, M. Z., Kölsch, A., Ahmed, S., & Liwicki, M. (2017, November). Cutting the error by half: Investigation of very deep cnn and advanced training strategies for document image classification. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*(Vol. 1, pp. 883-888). IEEE.

Project 3.1: Figure/Graph Extraction from Document

Objectives | Find and localize *Figure*/Graph in a document image

Applications | Graph retrieval, document segmentation based on content type

Figure/Graph Extraction from Document | Technical Details



An FCN (U-NeXt) is used

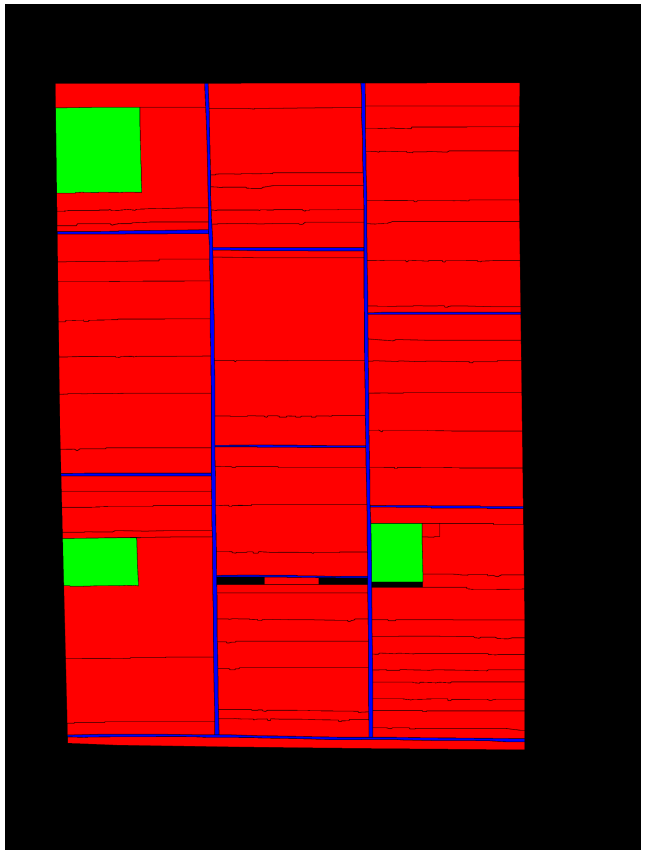
- U-NeXt combines ResNeXt and U-Net
  - ResNeXt101_64x4d
- Why ResNeXt101_64x4d?
  - Current state-of-art
  - Accessible pre-trained model
- **Transfer learning**
  - ResNeXt101_64x4d
  - Number of parameters:
    - 114.4 million → 32.8 million

Figure/Graph Extraction from Document | Datasets

- **ENP collection**: European newspaper collection
  - A subset used for the International Conference on Document Analysis and Recognition competition
- **Beyond Word collection**: Transcribed collection
  - But cannot be used for training directly …
    - Problem 1: missing figures in ground-truth
    - Problem 2: inaccurate ground-truth

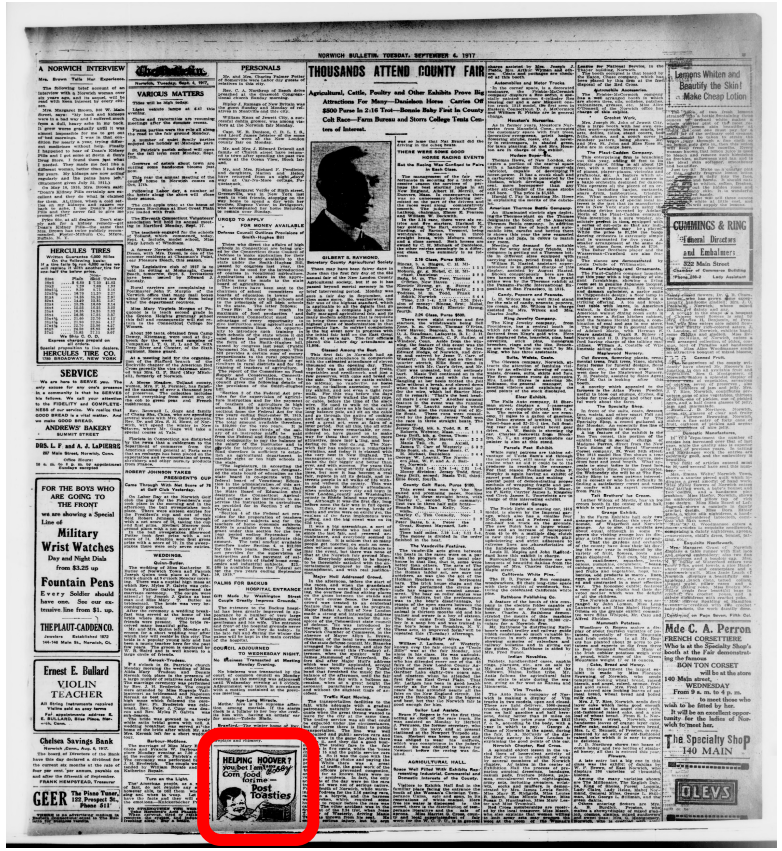
Figure/Graph Extraction from Document | Datasets: ENP

Document Image

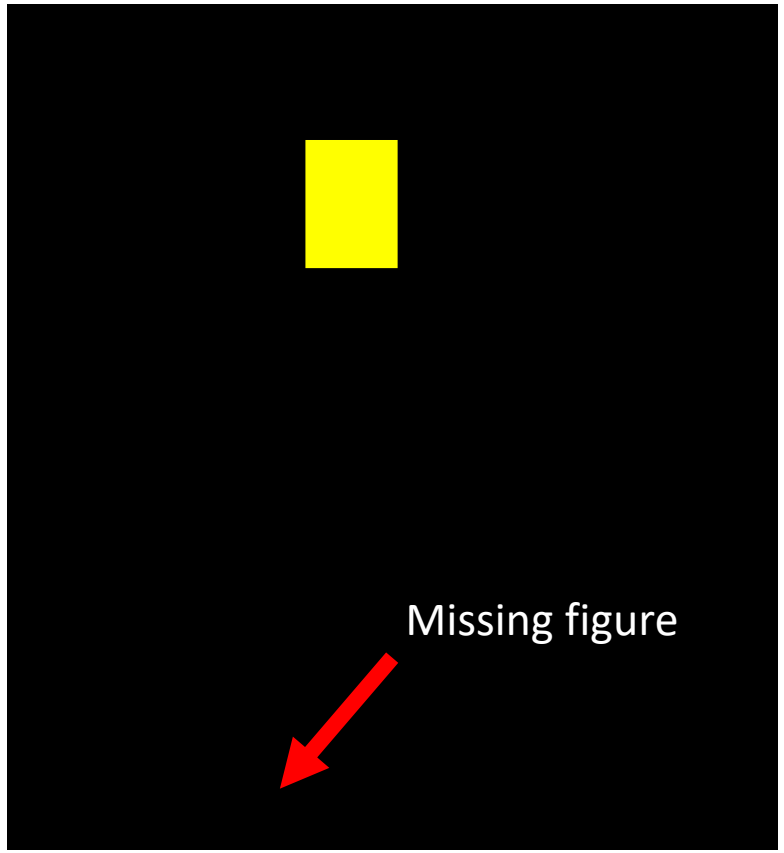


Ground-truth

Figure/Graph Extraction from Document | Datasets: Beyond Words



Document Image

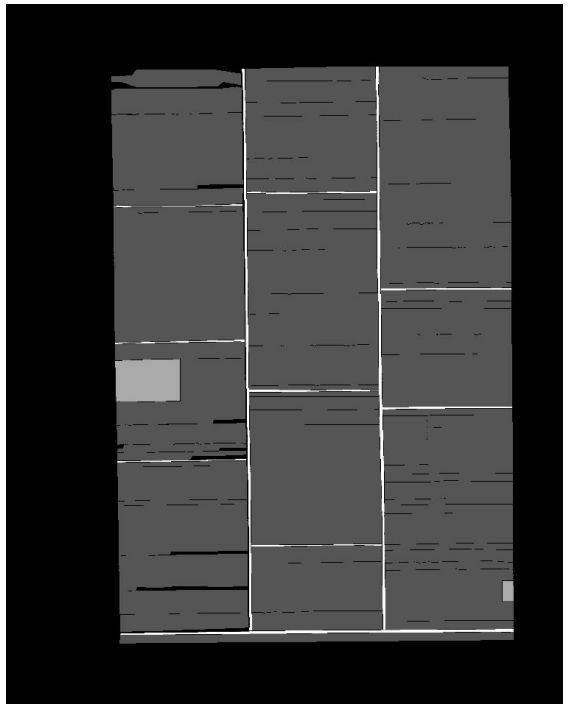


Ground-truth

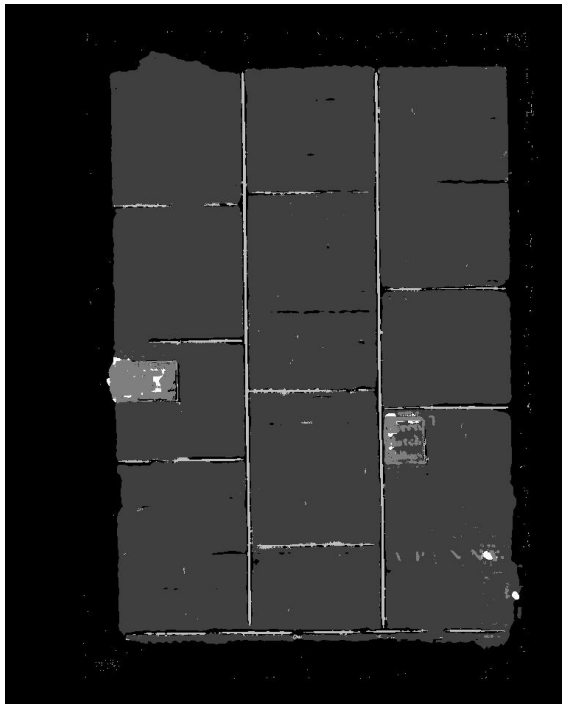
Figure/Graph Extraction from Document | Preliminary Results

- Transfer parameters from pre-trained ResNeXt101 64x4d
- Trained on ENP dataset

Document Image



Ground truth

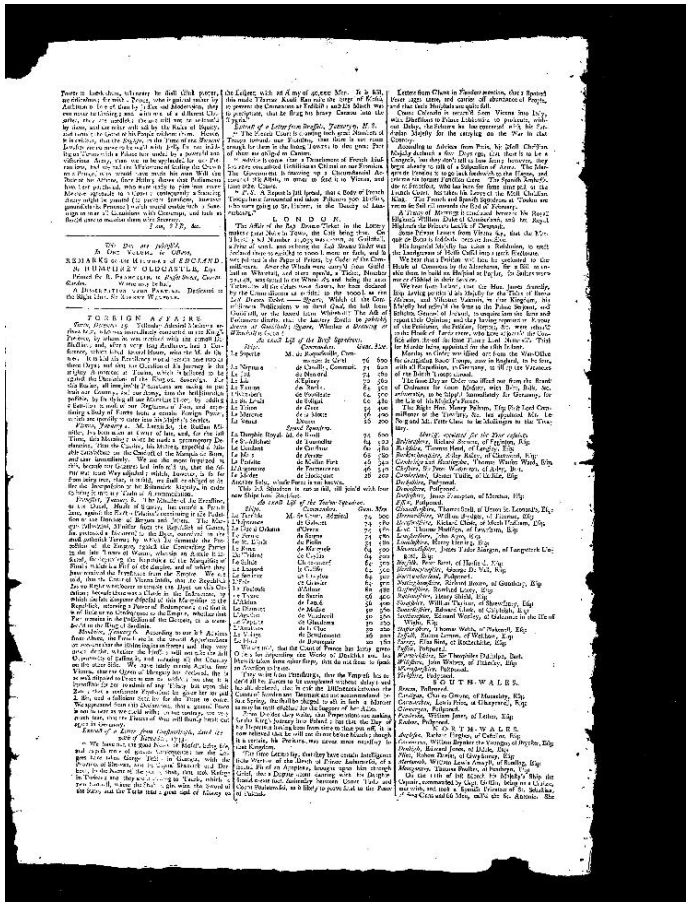


Prediction

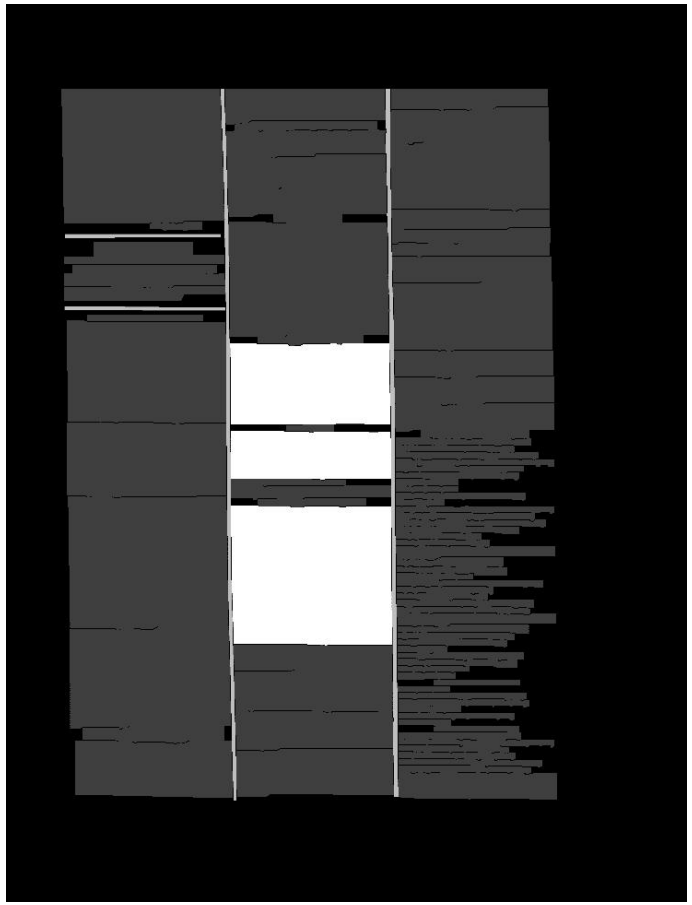
Figure/Graph Extraction from Document | Conclusions

- Promising preliminary results
- Potential applications
  - Segmentation based on content type to increase item-level accessibility
  - Retrieval of figures/graphs for further study
- Challenges
  - U-NeXt still needs more iterations of training
  - Preliminary training indicates that tables may be the hardest type to extract

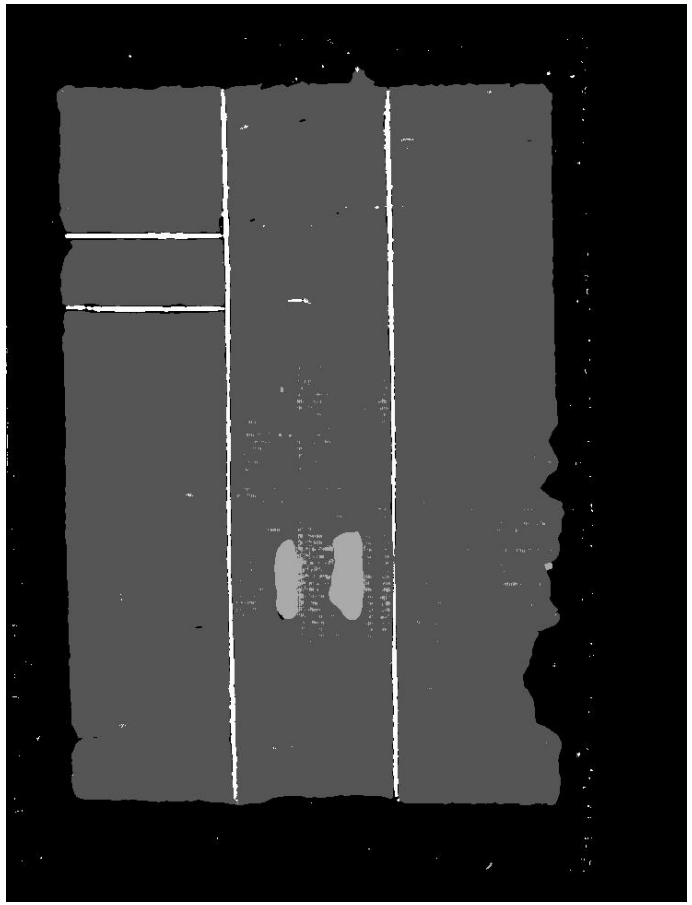
Figure/Graph Extraction from Document | Preliminary Results



Document Image



Ground truth



Prediction

Project 3.2: Text Extraction from Figure/Graph

Objectives | Extract texts from figure/graph

Applications | Metadata generation, OCR for figure/graph caption

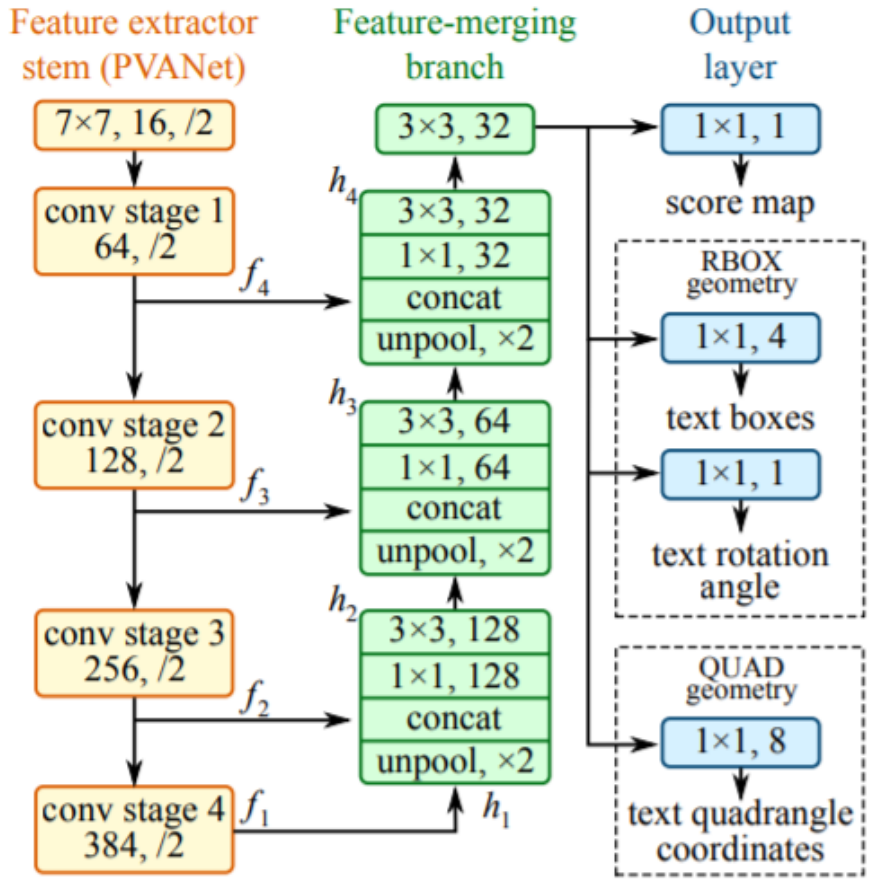
Text Extraction from Figure/Graph | Technical Details

EAST text detector

- EAST: Efficient and Accurate Scene Text detector
- HyperNet + U-Net
- Detect texts in graphic images in any direction

Why applicable?

- figures/illustrations are snippets of a graphic region



Text Extraction from Figure/Graph | Preliminary Results

Detected Texts

- Performance on detecting texts in newspaper figure/graph is good
- Texts location is recorded

Text Lines

- 6 text lines
- { "x0": 62, "y0": 608, "x1": 135, "y1": 588, "x2": 143
- { "x0": 188, "y0": 33, "x1": 312, "y1": 31, "x2": 313,
- { "x0": 331, "y0": 31, "x1": 423, "y1": 30, "x2": 423,
- { "x0": 116, "y0": 34, "x1": 166, "y1": 33, "x2": 166,
- { "x0": 405, "y0": 755, "x1": 470, "y1": 757, "x2": 47
- { "x0": 475, "y0": 756, "x1": 531, "y1": 757, "x2": 53

Text Extraction from Figure/Graph | Conclusions

- Promising preliminary results
- Potential application
  - Perform OCR on detected text regions for higher accuracy
  - Extract OCR-ed words in detected text regions as metadata

Project 4.1: Subjective Quality Assessment

Objectives | Access document images based on human perception

Applications | Providing metadata based on human visual perception

Subjective Quality Assessment | Proposal

- Adding an interface to allow users to classify the quality of document images
  - No need for **verbal annotation**
- A simple interface with
  - A drop box having five-level rating scores for **MOS** (i.e., **5-Excellent, 4-Good, 3-Fair, 2-Poor, and 1-Bad**)
  - Buttons, if detailed aspects such as contrast, range-effect, background-cleanness, and content density are needed

Subjective Quality Assessment | Benefits

- A human perception-based document image quality assessment (DIQA) database that can support ***further studies and experiments*** such as machine learning model training
- A ***publicly available*** database can draw attention to more research teams for research competition in academia
- Trained machine learning mode could ***enhance the filter or query search*** in the new UI of Beyond Word to sort images based on their quality

Project 4.2: Objective Quality Assessment

Objectives | Analyze image quality of the civil war collection By the People

Applications | Providing quality scores for machine reading on four criteria: (1) *skewness*, (2) *contrast*, (3) *range-effect*, and (4) *bleed-through*

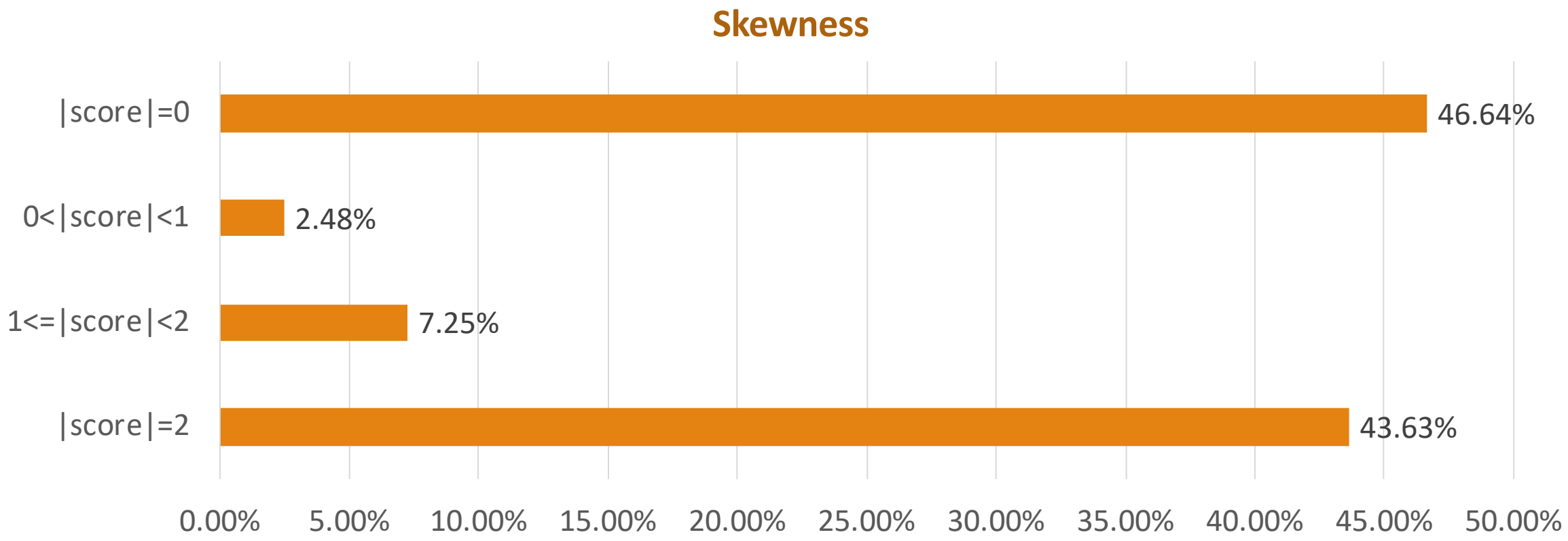
Objective Quality Assessment | Technical Details

- Objective quality assessment on four criteria
  - *Skewness, Contrast, Range-effect, Bleed-through*
  - *Based on the DIQA programs developed at Aida @ UNL (previously tested using Chronicling America's repository of archived newspaper pages*
  - *Not* directly machine learning related
- **Why**?
  - Help identify images that need pre-processing
  - Reduce unnecessary workload for pre-processing images
  - Indicate general qualities of the dataset

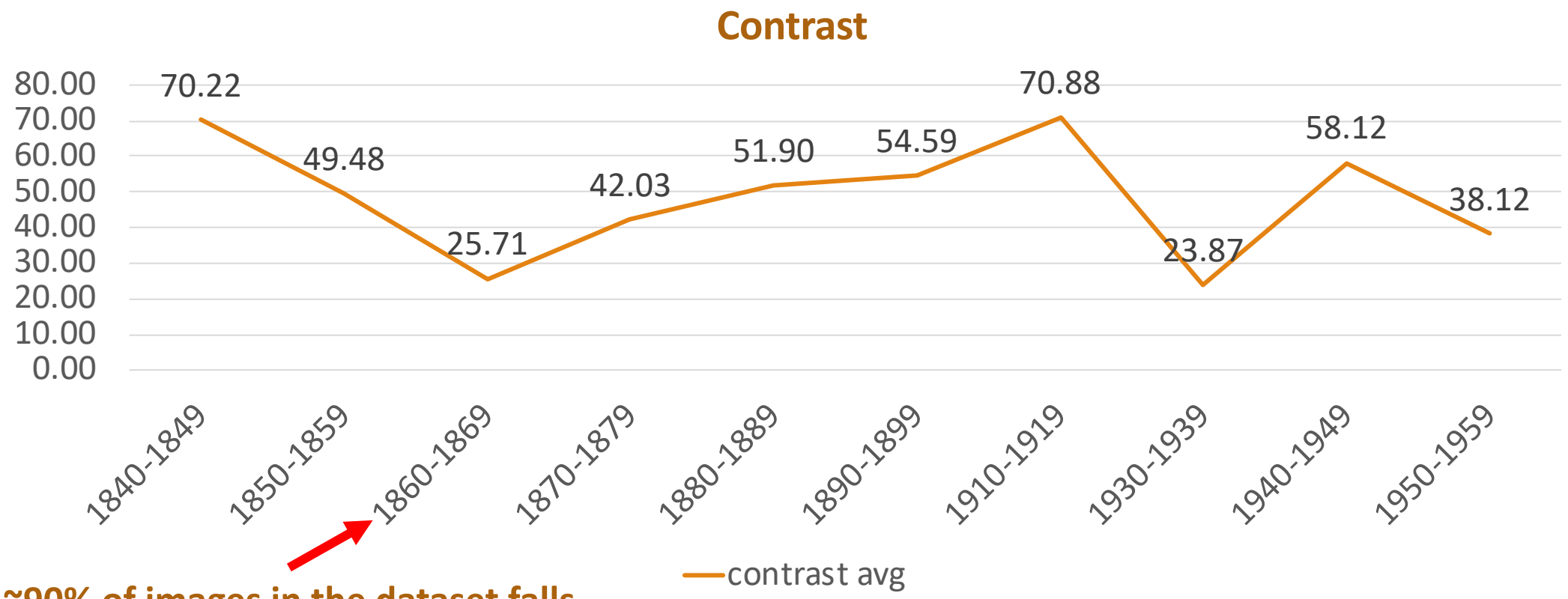
Objective Quality Assessment | Datasets

- The Civil War collection within By the People:
  - 36003 images were downloaded
  - 35990 images passed the DIQA program
    - *13 images failed as they barely had texts (see examples later)*

Objective Quality Assessment | Experimental Results

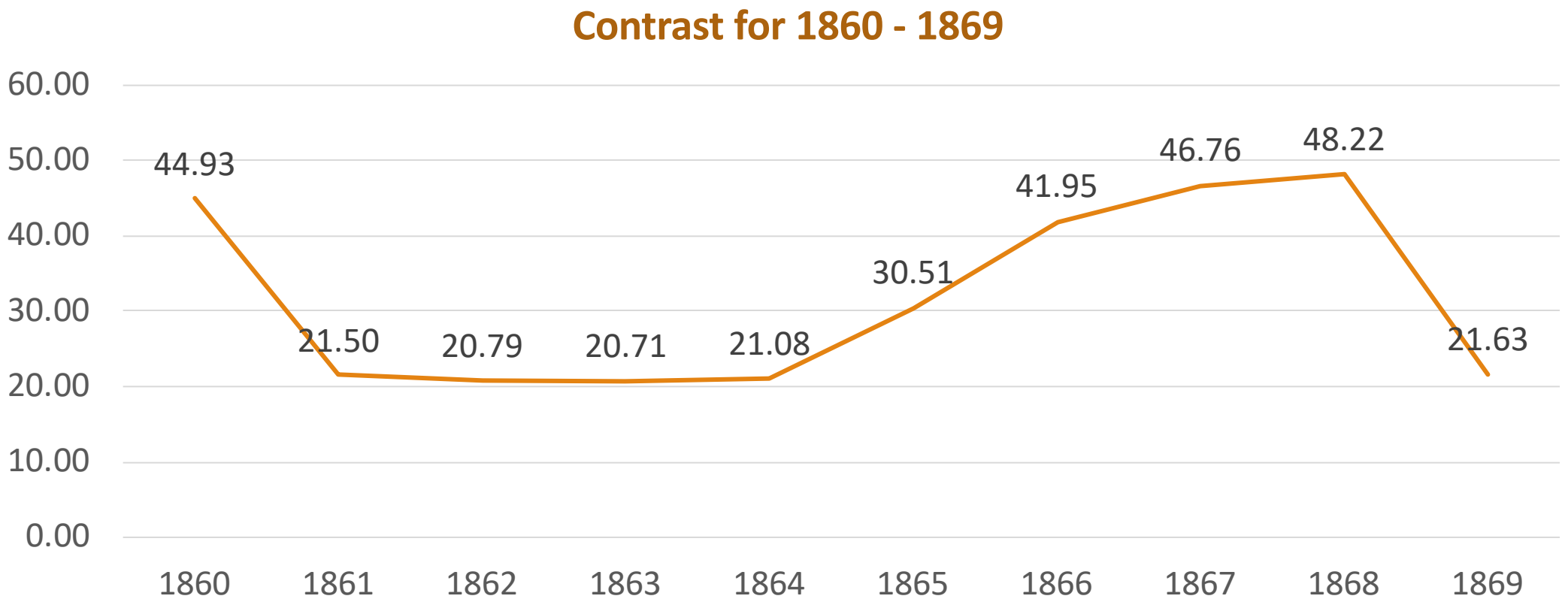


Objective Quality Assessment | Experimental Results

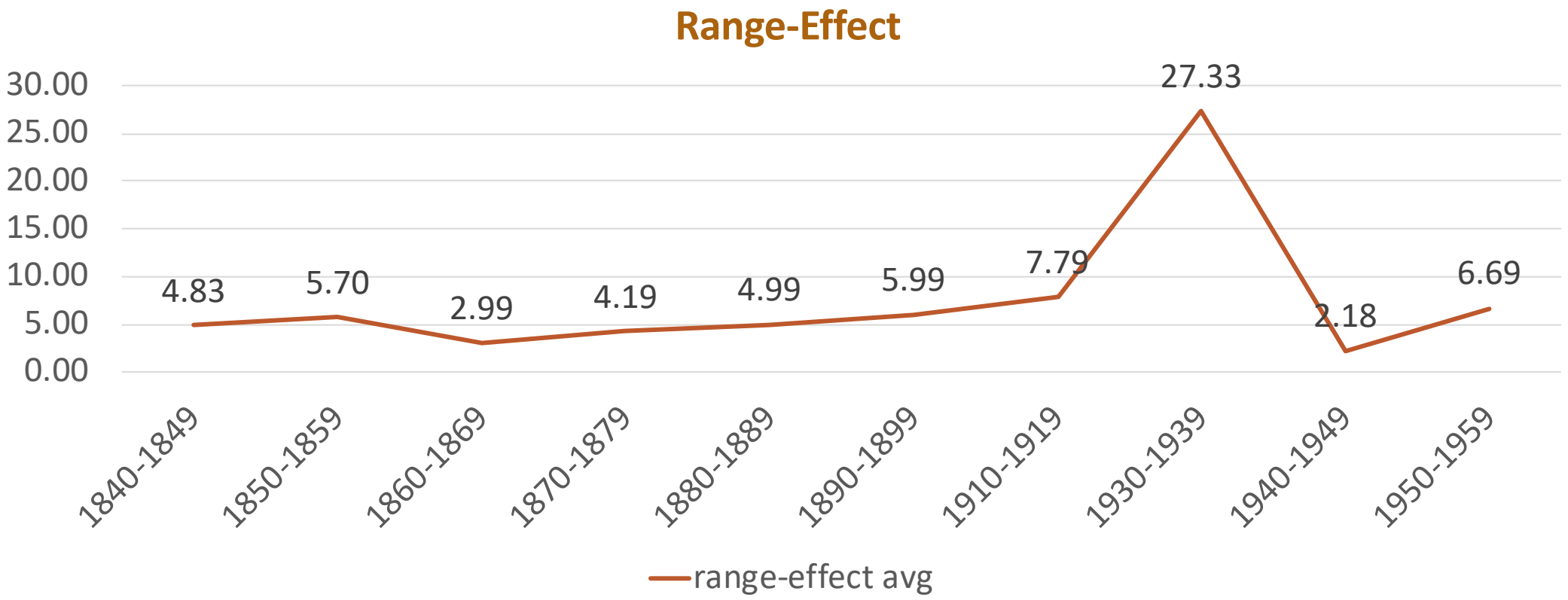


~90% of images in the dataset falls within this range

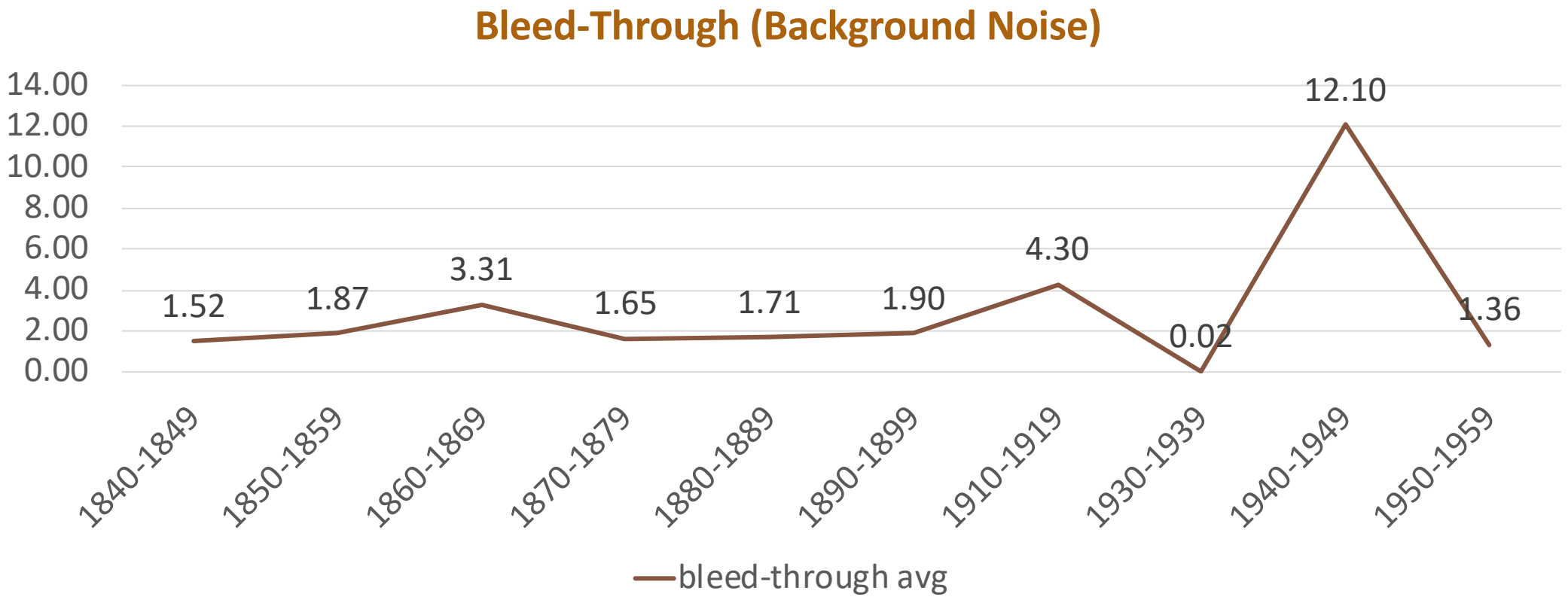
Objective Quality Assessment | Experimental Results



Objective Quality Assessment | Experimental Results



Objective Quality Assessment | Experimental Results

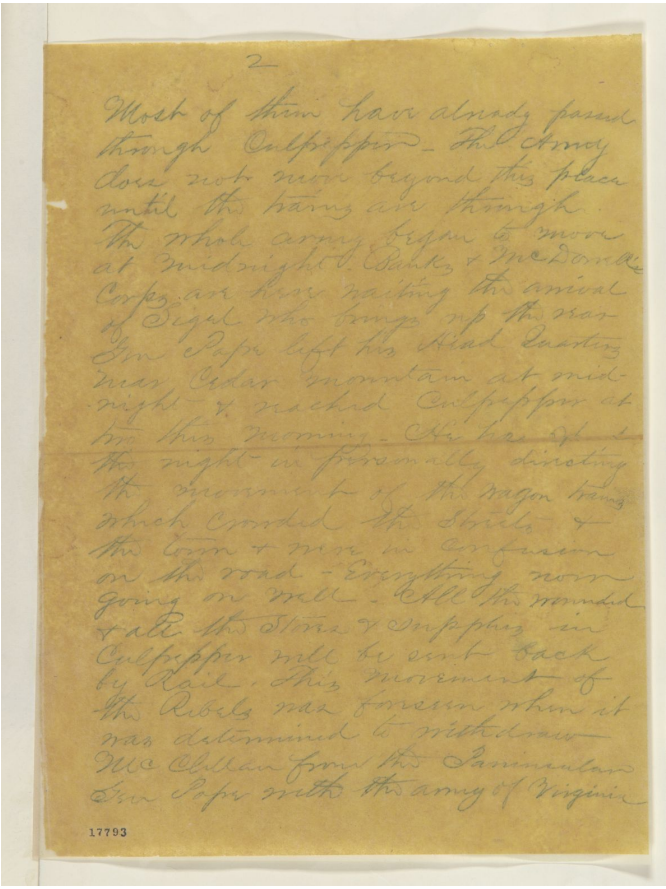


Objective Quality Assessment | Observations

- Must say something about your assessment. Good? Bad? What about the images?

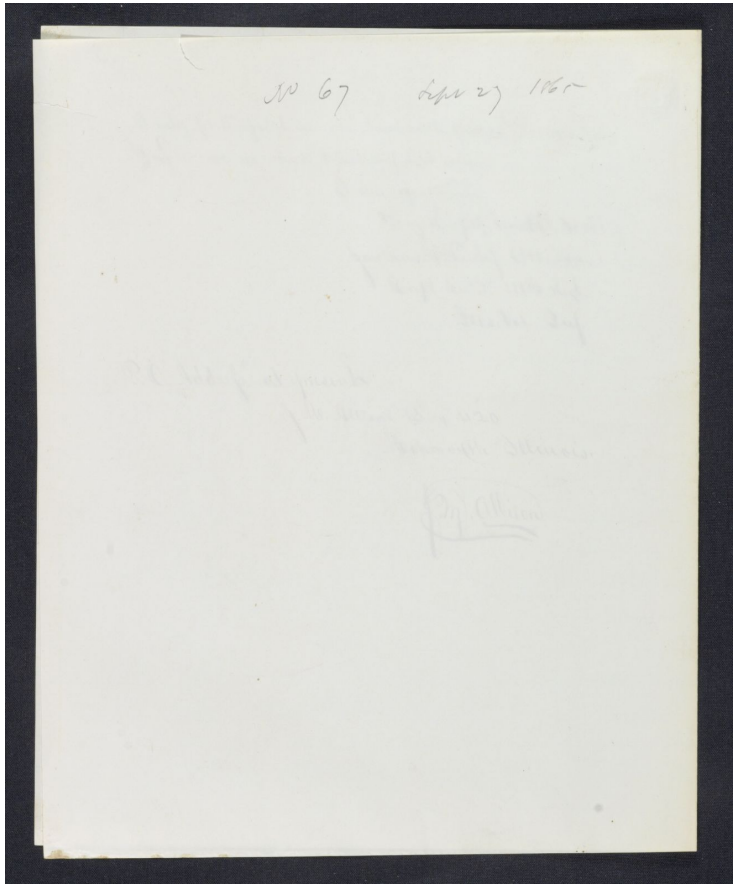
Objective Quality Assessment | Potential Issues

- Numerous images with yellowish background and faded inks
- They are hard to read even to human eye
  - Contrast could be lowered
  - Skewness could be almost impossible to compute



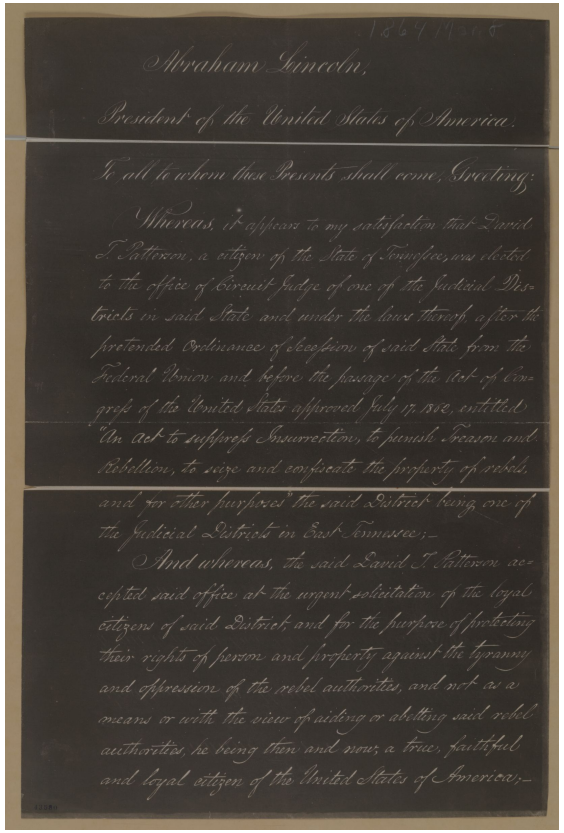
Objective Quality Assessment | Potential Issues

- Numerous images are covers or labels of a series
- These images are largely blank
  - Contrast is poor
  - Histogram equalization might be able to enhance the quality



Objective Quality Assessment | Potential Issues

- There are color-inverted images from microfilm
  - Renders bleed-through assessment useless



Abraham Lincoln,

President of the United States of America:

To all to whom these Presents shall come, Greeting:

Whereas, it appears to my satisfaction that David T. Patterson, a citizen of the State of Tennessee, was elected to the office of Circuit Judge of one of the Judicial Districts in said State, and under the laws thereof, after the pretended Ordinance of Secession of said State from the Federal Union and before the passage of the Act of Congress of the United States approved July 17, 1862, entitled "An Act to suppress Insurrection, to punish Treason and Rebellion, to seize and confiscate the property of rebels, and for other purposes" the said District being one of the Judicial Districts in East Tennessee;—

And whereas, the said David T. Patterson accepted said office at the urgent solicitation of the loyal citizens of said District, and for the purpose of protecting their rights of person and property against the tyranny and oppression of the rebel authorities, and not as a means or with the view of aiding or abetting said rebel authorities, he being then and now a true, faithful and legal citizen of the United States of America;—

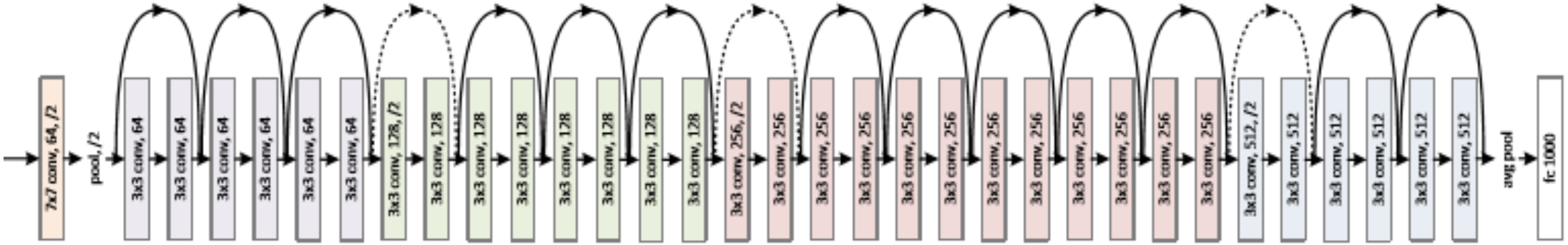
Project 5: Digitization Type Differentiation: Microfilm or Scanned

Objectives | Recognize if an image digitized from *Scanned* or *Microfilm*

Applications | Metadata generation, pre-processing policy selection

Digitization Type Differentiation | Technical Details

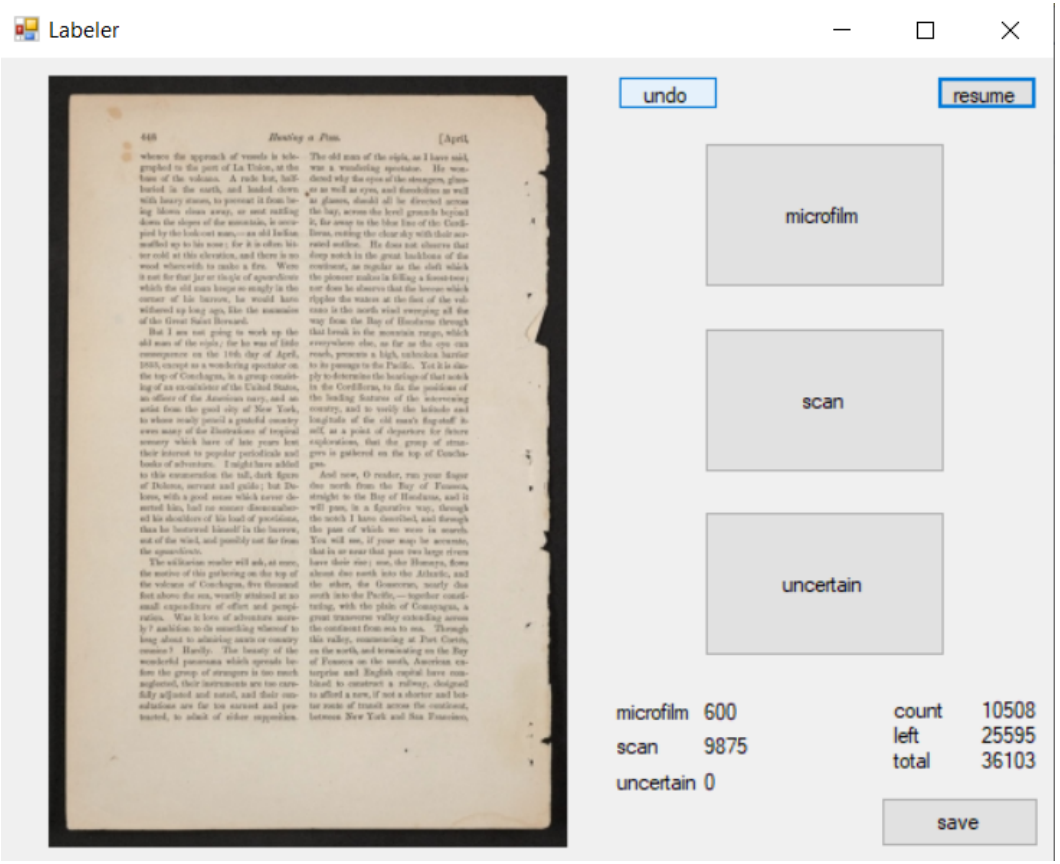
- Pre-trained ResNeXt is adopted
- Attached output layers are two dense layers with a 1D output vector
- **The pre-trained ResNeXt can classify images to 1000 different categories**
- The pre-trained ResNeXt is a good feature extractor
  - Number of parameters: 94.1 million → 12.6 million



Digitization Type Differentiation | Datasets

- Created from the Civil War collection within By the People
- A manually created database by ***randomly*** choosing 600 images on scanned materials and 600 images on microfilm materials
- The randomization was performed by shuffling the entire list of 36,003 images in the collection
- The randomization ensured that images in the collection have a fair chance to be chosen
- The randomization seed was fixed to ensure the experiments can be reproduced

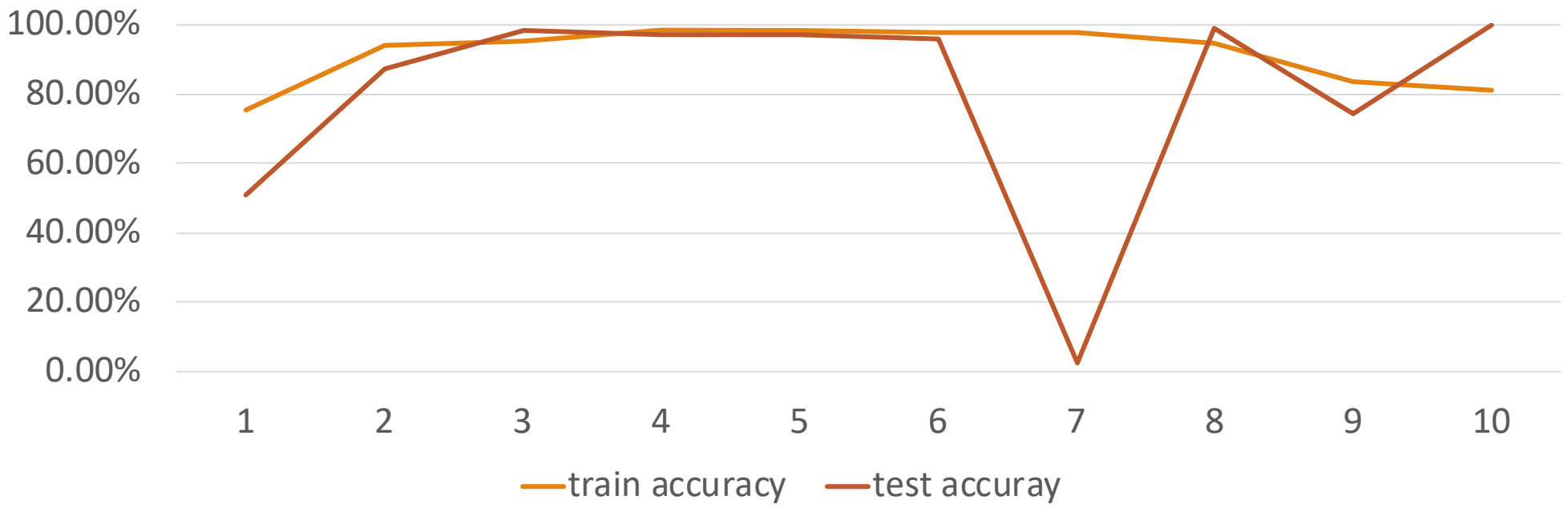
Digitization Type Differentiation | Datasets



- **Rough estimate**: Based on 10,508 images that was processed, ***ratio of images from microfilm to scanned materials is about 1:16***

Digitization Type Differentiation | Experimental Results

- With pre-trained ResNeXt,
  - It only took **one** iteration to reach more than 90% accuracy on training set, and
  - It only took **two** iterations to reach more than 90% accuracy on testing set



Digitization Type Differentiation | Experimental Results

- The best test iteration result was able to 100% correctly classify all images

		Ground Truth	
		Scanned	Microfilm
Prediction	Scanned	60	0
	Microfilm	0	60

Digitization Type Differentiation | Conclusions

- Existing pre-trained model can be easily extended to more designated tasks
- The extended model only need a small set of labeled data to reach near-perfect performance in this task
- Automated digitization type differentiation is ***readily*** achievable.

Digitization Type Differentiation | Tips on Choosing …

- **How** to choose pre-trained model from the “zoo” (or the ”kitchen”)?

Task Type	Model
Type differentiation/classification, with limited computing power	Mobile Net
Type differentiation/classification, with fair amount of computing power	ResNet, ResNeXt
Type differentiation/classification, with good amount of computing power	VGG Network, Inception
Task needs to locate or extract object/figure/graph, based on the amount of computing power	Combine a U-shaped network
Task needs to refine extracted location, and locations may be overlapped	HyperNet

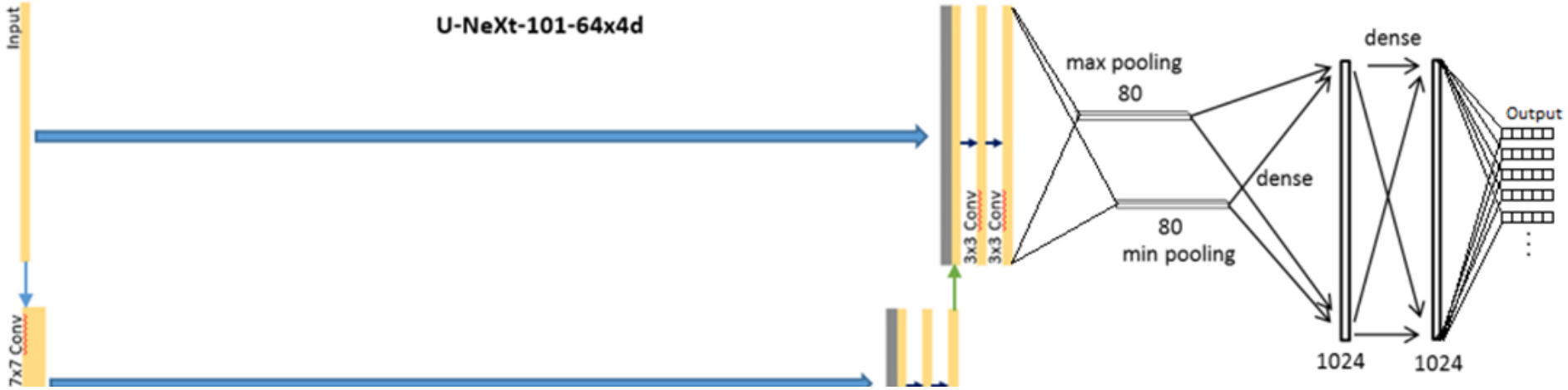
Questions ?

Thank you very much for your participation.

Thanks to Library of Congress + UNL Collaboratory

Subjective Quality Assessment | Technical Details

- Fine tuning pre-trained U-NeXt in Project 1
- **Difference:** DIQA need only high-level score on image quality
  - Instead of 2D matrix output, subjective quality assessment only need 1D vector
  - Elements of the 1D output are image quality scores, such as Mean Opinion Score



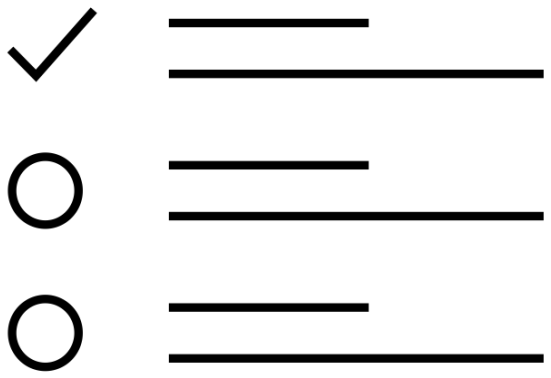
Subjective Quality Assessment | Datasets

- Machine Learning, especially for deep learning, requires large amounts of labeled data for training
- Current existing quality assessment databases contain only quality scores for machine perception
  - Previous Aida @ UNL work: Document Image Quality Assessment (DIQA) for Chronicling America newspapers
- **Challenge**
  - Lack of human perception-based DIQA database

A COLLABORATORY BETWEEN THE LIBRARY OF CONGRESS AND THE IMAGE ANALYSIS FOR ARCHIVAL DISCOVERY (AIDA) LAB AT THE UNIVERSITY OF NEBRASKA, LINCOLN, NE

Liz Lorang (faculty)
Leen-Kiat Soh (faculty)
Yi Liu (PhD student)
Chulwoo Pack (PhD student)

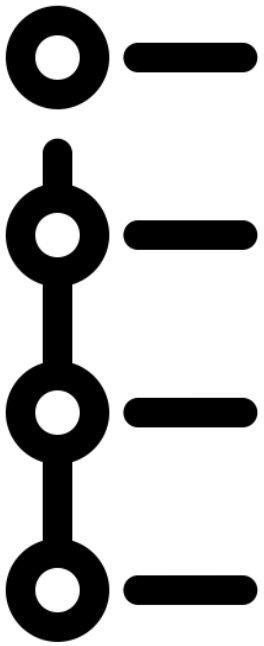
Digital Libraries, Intelligent Data Analytics, and Augmented Description: A Demonstration Project



Created by Alfredo @ IconsAlfredo.com
from Noun Project

framing

1. Aida research team and **background** for the demonstration project
2. **Broad outlines** of “Digital Libraries, Intelligent Data Analytics, and Augmented Description”
3. **What has changed** for us as a research team over the collaboration and why
4. **Deliverables** of our work
5. Some things we’re thinking about for “***What next***”?
6. **Deep-dive** into the explorations (Leen-Kiat, Yi, and Mike)

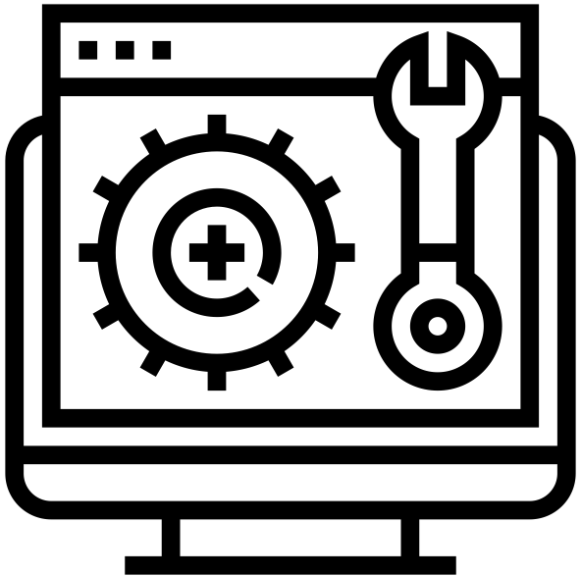


Created by bezier master
from Noun Project

background

- 2014: Our original goal was to find poetry in historic newspapers and create a corpus of poetry for further analysis.
- Since 2016: IMLS-funded project to extend Aida's software across a more diverse range of digitized newspapers and textual forms and assess the broader potential of image analysis as a methodology for information classification, identification, discovery, and retrieval in digital libraries.
- Reports, codes, and datasets are linked from projectaida.org

Aida

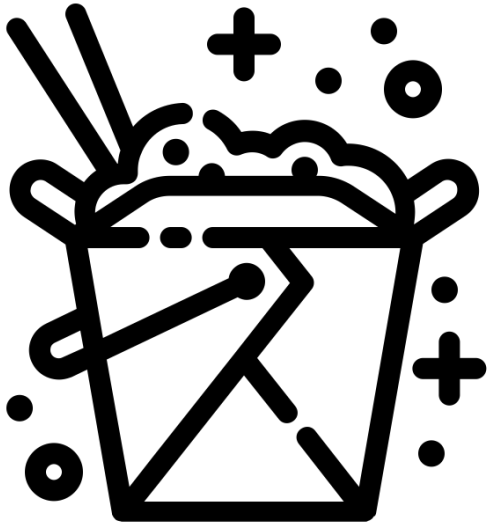


Created by Eucalyp
from Noun Project

background

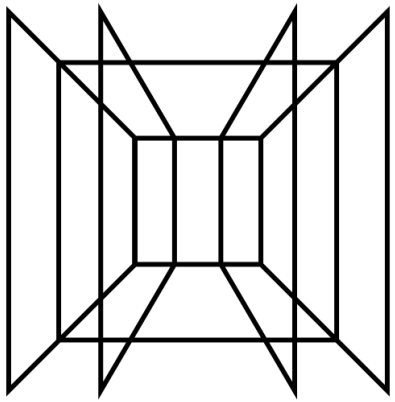
- In our proposal to IMLS, we anticipated that "image processing approaches and intelligent classifiers, such as those we are developing, are likely to become part of the **standard toolkit of archivists**."
- As part of the Oceanic Exchanges Digging into Data team, we consider how the models we're developing could become part of "**downstream" workflows.**

Takeaways



Created by Crystal Gordon
from Noun Project

- Even content of a single type (e.g., newspapers) can be incredibly diverse, with implications for generalizing approaches
- There may be trade-off between generalizability and effectiveness
- We need to better understand the materials themselves, as well as other intermediary processes through digitization, to inform our approaches.



Created by Pablo Rozenberg
from Noun Project

broad outlines

Digital Libraries, Intelligent Data Analytics, and Augmented Description

Three broad goals as framed in proposal:

1. Develop and investigate the viability and feasibility of textual and image-based data analytics approaches to support and facilitate discovery

2. Understand technical tools and requirements for the Library of Congress to improve access and discovery of its digital collections

3. Enable the Library of Congress to plan for future technical possibilities

Aida

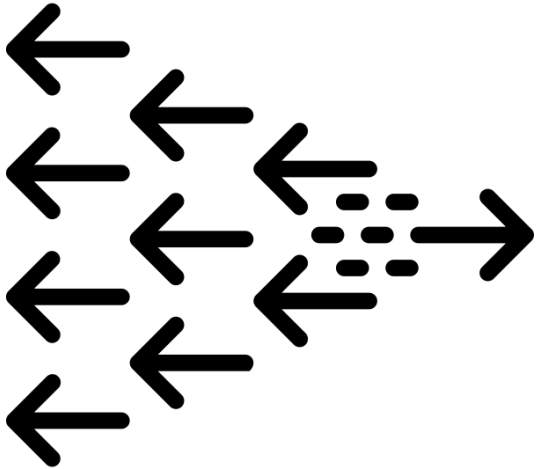
broad outlines

Created by mikicon
from Noun Project

We posed *a lot* of questions in our plan for collaborating with the Library of Congress. Just a few . . . .

- What are the various types of information, files, and structures that researchers want access to?
- How can more of this information be made available to researchers in ways that are meaningful to them, including in bulk formats?
- What types of accuracy are needed in digital collections to make them usable to researchers asking a range of questions? What levels of accuracy do different types of processes and research questions require?

what has changed



Created by priyanka
from Noun Project

July 2019 kick-off meeting as a bit of a turning point . . .

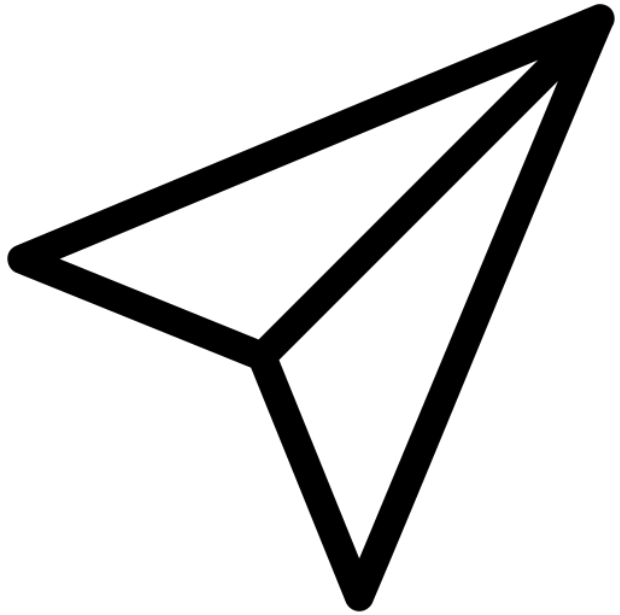
What would help the LC manage the material and serve it to users better?

What might we learn about size and condition of materials?

What are the inadvertent barriers we are creating? Where are we roping off next potential steps for researchers?

Can we identify printed materials in handwritten manuscript collections?

What are the internal needs and questions?



Created by Aya Sofya
from Noun Project

deliverables

- Real-time reports, completed throughout project
- Code and code documentation
- Curated datasets
- Final white paper

what's next

- usable, operationally-significant machine learning;
- engaging, research-facilitating machine learning; and
- socially, culturally responsible machine learning

in cultural heritage digital libraries and organizations.

Digital Libraries, Intelligent Data Analytics, And Augmented Description: A Demonstration Project

A COLLABORATORY BETWEEN THE LIBRARY OF CONGRESS AND THE IMAGE ANALYSIS FOR ARCHIVAL DISCOVERY (AIDA) LAB AT THE UNIVERSITY OF NEBRASKA, LINCOLN, NE

Liz Lorang (faculty)
Leen-Kiat Soh (faculty)
Yi Liu (PhD student)
Chulwoo Pack (PhD student)

Introduction

projectaida.org

Collaborative research project between the Library of Congress and the Aida digital libraries research team at the University of Nebraska

5-month applied research project with the following goals:

- develop and investigate the viability and feasibility of textual and image-based data analytics approaches to support and facilitate discovery
- understand technical tools and requirements for the Library of Congress to improve access and discovery of its digital collections
- enable the Library of Congress to plan for improved applications and technical capacity as well as future innovations

Where We Were

Efforts

- Focused on **textual** materials (newspapers and minimally processed manuscript collections)
  - Chronicling America and Chron-Am content in Beyond Words
  - Manuscript collections represented in By the People
- Explorative projects in
  - image segmentation;
  - visual content extraction from textual materials; text extraction from images;
  - document/corpus quality assessment;
  - differentiation among documents created via different means (e.g., microfilming and direct scanning);
  - differentiating among printed, handwritten, and mixed content;
  - metadata generation
- Have completed one round of iterative development for each project; 2nd round of development underway now

Preliminary Take-Aways

- **Expanded sense of how these projects may be useful**:
  - Still thinking of end users, including end users who are looking for a variety of data and information not only content *about* something.
  - Greater emphasis on internal use within the Library of Congress for understanding materials, decision-making processes, and more
- Considering how **crowd-sourced information can aid in machine learning**, as well as what may be well-suited to the crowd, to the machine, and to domain experts
- Not only immediate application of machine learning, but **analysis of the materials through a variety of strategies to inform machine learning**
- **Full range of resources**—computational, human, technical, and social—*necessary* to do this work

Aida

Text Extraction from Figure/Graph | Preliminary Results

Detected Texts

- Performance on detecting texts in newspaper figure/graph is good
- Texts location is recorded

Text Lines
- 6 text lines
- { "x0": 62, "y0": 608, "x1": 135, "y1": 588, "x2": 143
- { "x0": 188, "y0": 33, "x1": 312, "y1": 31, "x2": 313,
- { "x0": 331, "y0": 31, "x1": 423, "y1": 30, "x2": 423,
- { "x0": 116, "y0": 34, "x1": 166, "y1": 33, "x2": 166,
- { "x0": 405, "y0": 755, "x1": 470, "y1": 757, "x2": 47
- { "x0": 475, "y0": 756, "x1": 531, "y1": 757, "x2": 53

Document Type Classification | Datasets

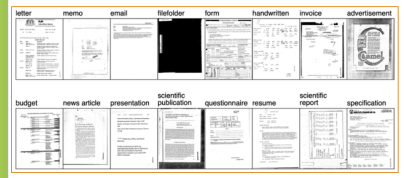


Figure 9. Example document images from each 16 different classes in *RVL_CDIP* dataset

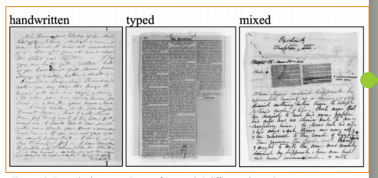
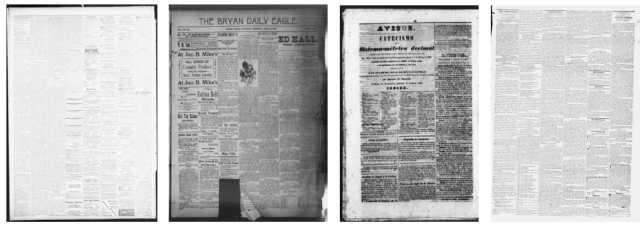


Figure 10. Example document images from each 3 different classes in *suffrage_1002* dataset

Objective Quality Assessment | Examples



Contrast | Range-effect | Bleed-through | Skewness

Project 1. Document Segmentation

Objectives | Find and localize *Figure/Illustration/Cartoon* presented in an image
Applications | metadata generation, discover-/search-ability, visualization, etc.

Project 2.1. Figure/Graph Extraction

Objectives | Find and localize *Figure*/Graph in a document image
Applications | Graph retrieval, document segmentation based on content type

Project 2.2. Text Extraction from Figure/Graph

Objectives | Extract texts from figure/graph
Applications | Metadata generation, OCR for figure/graph caption

Project 3. Document Type Classification

Objectives | (1) Classify a given image into one of *Handwritten/Typed/Mixed* type; (2) Classify a given image into one of *Scanned/Microfilmed*
Applications | metadata generation, discover-/search-ability, cataloging, etc.

Project 4. Quality Assessment

Objectives | Analyze image quality of the civil war collection By the People
Applications | Providing quality scores for machine reading on four criteria: (1) *skewness*, (2) *contrast*, (3) *range-effect*, and (4) *bleed-through*

Project 5. Digitization Type Differentiation: Microfilm or Scanned

Objectives | Recognize if an image digitized from *Scanned* or *Microfilm*
Applications | Metadata generation, pre-processing policy selection

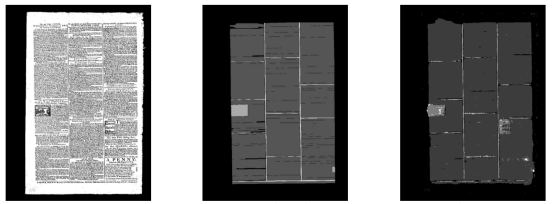
Document Segmentation | Dataset

European Historical Newspapers (ENP)
- Total of 57,339 image snippets in 500 pages
  - All pages have multiple snippets
- Issues
  - Data imbalance
    - Text: 43,780
    - Figure: 1,452
    - Line-separator: 11,896
    - Table: 221

Figure 4. Example of image (left) and ground-truth (right) from ENP dataset. In the ground-truth, each color represents the following components: (1) black: background, (2) red: text, (3) green: figure, (4) blue: line-separator, and (5) yellow: table.

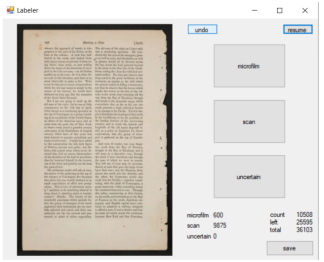
Figure/Graph Extraction | Preliminary Results

- Transfer parameters from pre-trained ResNeXt101 64x4d
- Trained on ENP dataset

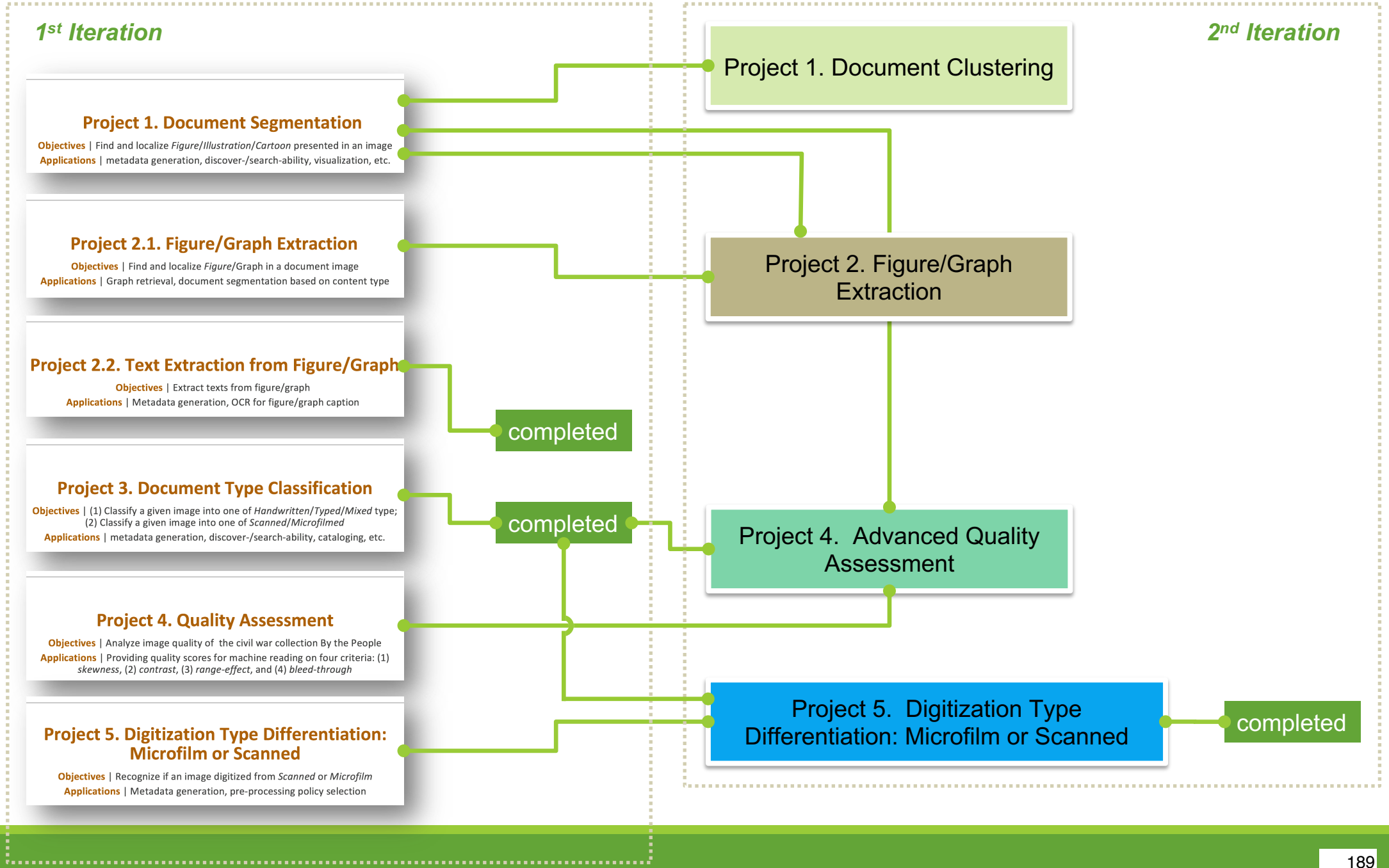


Document Image | Ground truth | Prediction

Digitization Type Differentiation | Datasets



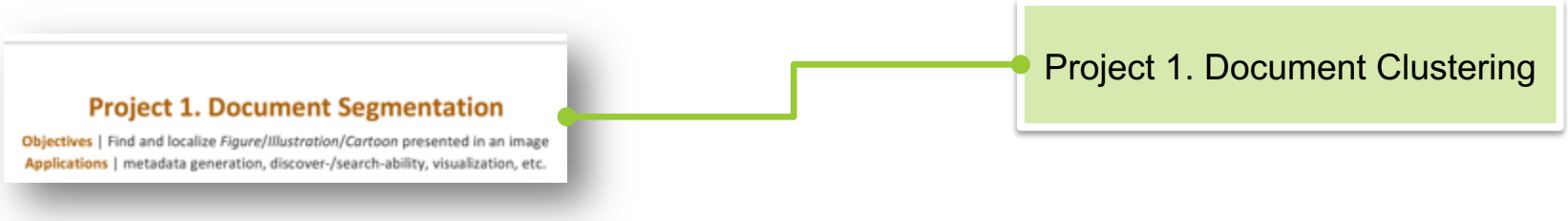
- **Rough estimate**: Based on 10,508 images that was processed, ***ratio of images from microfilm to scanned materials is about 1:16***


1st Iteration
2nd Iteration
Project 1. Document Segmentation
Objectives | Find and localize Figure/Illustration/Cartoon presented in an image
Applications | metadata generation, discover-/search-ability, visualization, etc.
Project 2.1. Figure/Graph Extraction
Objectives | Find and localize Figure/Graph in a document image
Applications | Graph retrieval, document segmentation based on content type
Project 2.2. Text Extraction from Figure/Graph
Objectives | Extract texts from figure/graph
Applications | Metadata generation, OCR for figure/graph caption
completed
Project 3. Document Type Classification
Objectives | (1) Classify a given image into one of Handwritten/Typed/Mixed type; (2) Classify a given image into one of Scanned/Microfilmed
Applications | metadata generation, discover-/search-ability, cataloging, etc.
completed
Project 4. Quality Assessment
Objectives | Analyze image quality of the civil war collection By the People
Applications | Providing quality scores for machine reading on four criteria: (1) skewness, (2) contrast, (3) range-effect, and (4) bleed-through
Project 5. Digitization Type Differentiation: Microfilm or Scanned
Objectives | Recognize if an image digitized from Scanned or Microfilm
Applications | Metadata generation, pre-processing policy selection
Project 1. Document Clustering
Project 2. Figure/Graph Extraction
Project 4. Advanced Quality Assessment
Project 5. Digitization Type Differentiation: Microfilm or Scanned
completed

Project 1: Document Segmentation

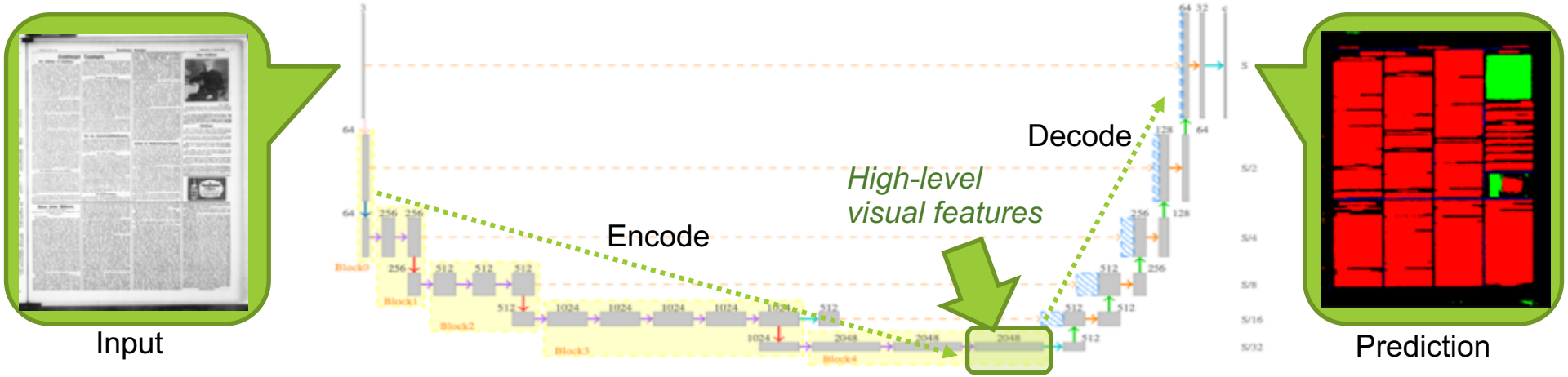
Objectives | Find and localize *Figure/Illustration/Cartoon* presented in an image

Applications | metadata generation, discover-/search-ability, visualization, etc.



Project 1 Document Segmentation | Where We Were

- **What we did**:
  - Used **dhSegment** [1] to localize texts, figures, layout borders, and tables
- **Datasets used**: 481 images from external dataset *European historical newspapers* [2]
- **Results**: pixel-wise accuracy = 91%; mIoU = 69%



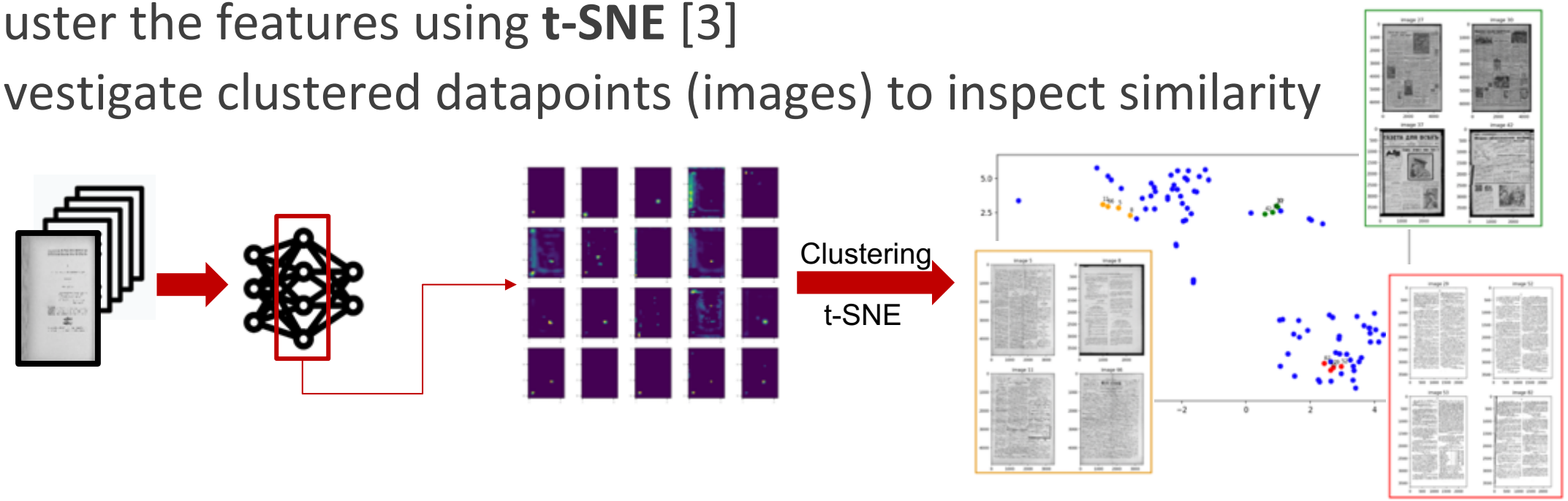
[1] Oliveira, Sofia Ares, Benoit Seguin, and Frederic Kaplan. "dhSegment: A generic deep-learning approach for document segmentation." *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2018. [2] https://www.primaresearch.org/datasets/ENP

Project 1 Document Segmentation | Where We Were

- **Implications**
  - **ResNet**, a feature extractor in **dhSegment**, is capable of encoding a whole image down to a set of *high-level visual features* effectively and efficiently
  - Visually similar document images likely have similar high-level visual features
  - **Question:** *In an effectively clustered space of high-level visual features, would document images in the same cluster share similar visual metadata?*

Project 1 Document Clustering | 2nd Iteration

- Use **dhSegment** to extract a set of high-level features
- Cluster the features using **t-SNE** [3]
- Investigate clustered datapoints (images) to inspect similarity



- **Results**: Same-cluster document images share similar visual cues (e.g., brightness, density, existence of figures, etc.) and those in different clusters do not

[3]. Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of machine learning research* 9.Nov (2008): 2579-2605.

Project 1 Document Clustering | Conclusions

- ***Limitation*** Solution model is *supervised*—based on the domain of training images (i.e., newspaper/typed), results might not transfer to other types of document images (e.g., letter/handwritten)
- ***Next Iteration*** Explore *unsupervised* deep-learning model to build a more *generic/universal* solution
- ***Future Directions*** A document cluster model can enrich high-level metadata and improve crowd-sourcing operations

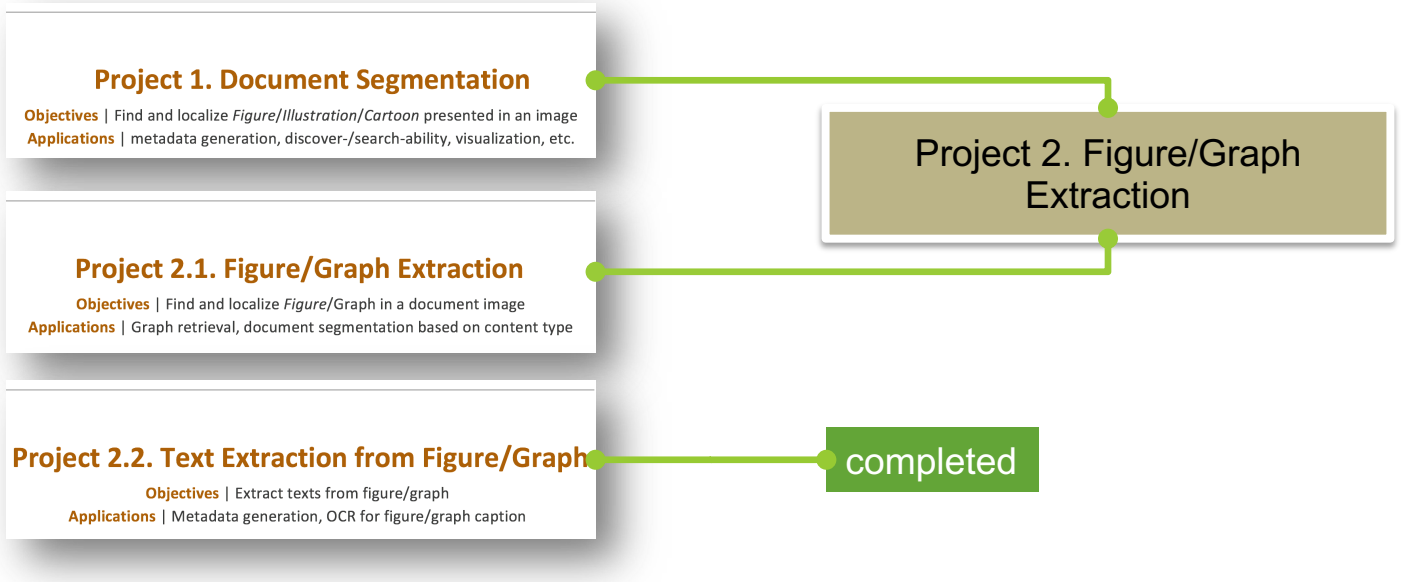
Idea 1 Suggest metadata tags for crowd-sourcing process and receive feedback

Idea 2 Suggest visually similar document images matching user interests and receive feedback

Project 2: Figure/Graph Extraction

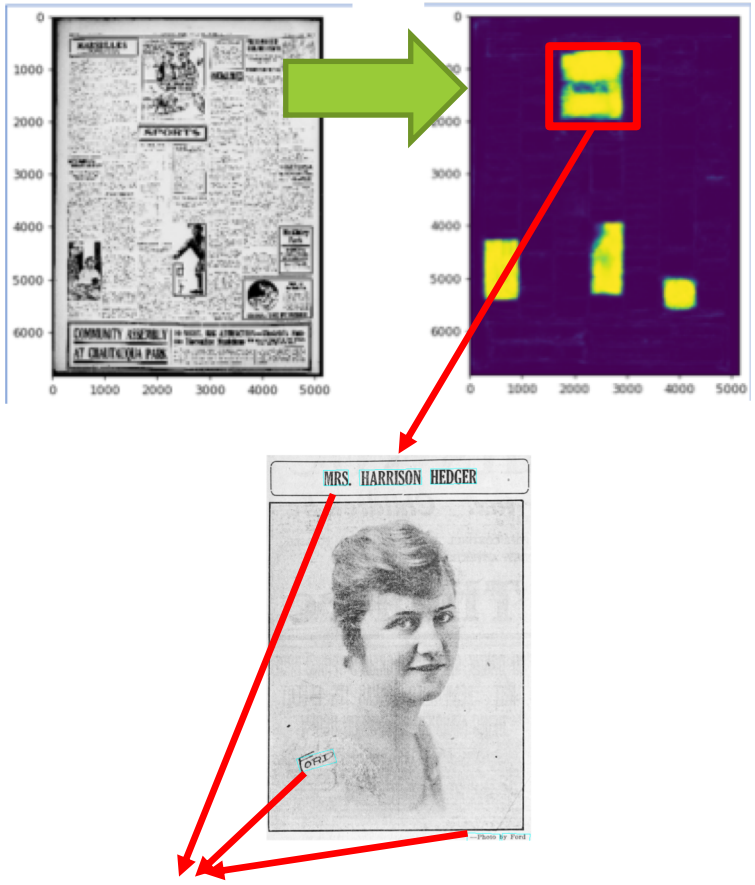
Objectives | Find and localize *Figure/Illustration/Cartoon* presented in an image

Applications | metadata generation, discover-/search-ability, visualization, etc.



Project 2 Figure/Graph Extraction | Where We Were

- **What we did**: Used **dhSegment** [4] to extract figure/graph regions
  - dhSegment is based on **ResNet** [5]
- **Datasets used**: 1,532 images from the *Beyond Words* collection
- **Results**: pixel-wise accuracy = 88%; mIoU = 26%



Text Lines

- 6 text lines
- { "x0": 62, "y0": 608, "x1": 135, "y1": 588, "x2": 143
- { "x0": 188, "y0": 33, "x1": 312, "y1": 31, "x2": 313,
- { "x0": 331, "y0": 31, "x1": 423, "y1": 30, "x2": 423,
- { "x0": 116, "y0": 34, "x1": 166, "y1": 33, "x2": 166,
- { "x0": 405, "y0": 755, "x1": 470, "y1": 757, "x2": 47
- { "x0": 475, "y0": 756, "x1": 531, "y1": 757, "x2": 53

[4] S. Ares Oliveira, B. Seguin and F. Kaplan, "dhSegment: A Generic Deep-Learning Approach for Document Segmentation," *2018 16th ICFHR*, 2018, pp. 7-12.
[5] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conf. on CVPR, 2016, pp. 770-778.

Project 2 Figure/Graph Extraction | Where We Were

- **Implications**
  - Due to poor performance on mIoU–possibly due to poor feature extraction in ResNet–we proposed to use a more advanced version of dhSegment called **U-NeXt** that has a stronger feature extractor
  - ***Automated solution to extract figures/graphs is promising***
    - Enrich page-level metadata by cataloging the types of visual components
    - Enrich item-level metadata by extracting texts in figure/graph regions

[4] S. Ares Oliveira, B. Seguin and F. Kaplan, "dhSegment: A Generic Deep-Learning Approach for Document Segmentation," *2018 16th ICFHR*, 2018, pp. 7-12.
[5] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conf. on CVPR, 2016, pp. 770-778.

Project 2 Figure/Graph Extraction | 2nd Iteration

- Used U-NeXt to extract figure/graph regions
  - U-NeXt uses **ResNeXt** [6]
- **Datasets used**: 1,532 images from Beyond Words collection
- **Results**: pixel-wise accuracy = 87%; mIoU = 38%
- **Implications**
  - 1st vs. 2nd iterations: U-NeXt showed noticeably higher mIoU while maintaining almost the same pixel-wise accuracy
  - *U-NeXt's ResNeXt is a better feature extractor than dhSegment's ResNet*

	dhSegment	U-NeXt
Pixel-wise Accuracy	88%	87%
mIoU	26%	38%

[6] S. Xie, R. Girshick, P. Dollár, Z. Tu and K. He, "Aggregated Residual Transformations for Deep Neural Networks," 2017 IEEE Conf. on CVPR, 2017, pp. 5987-5995.

Project 2 Figure/Graph Extraction | Another Implication

Table 1 Performance Comparison
1st vs 2nd Iterations

	dhSegment	U-NeXt
Pixel-wise Accuracy	88%	87%
mIoU	26%	38%

Why are these values so much lower than pixel-wise accuracy?

- **Implications**
  - Low mIoU indicates low regional coincidence rate, meaning extracted regions only partially matching the ground truth regions
    - … Visual inspection to investigate …

Project 2 Figure/Graph Extraction | Visual Inspection

Findings: Issues with ground truth

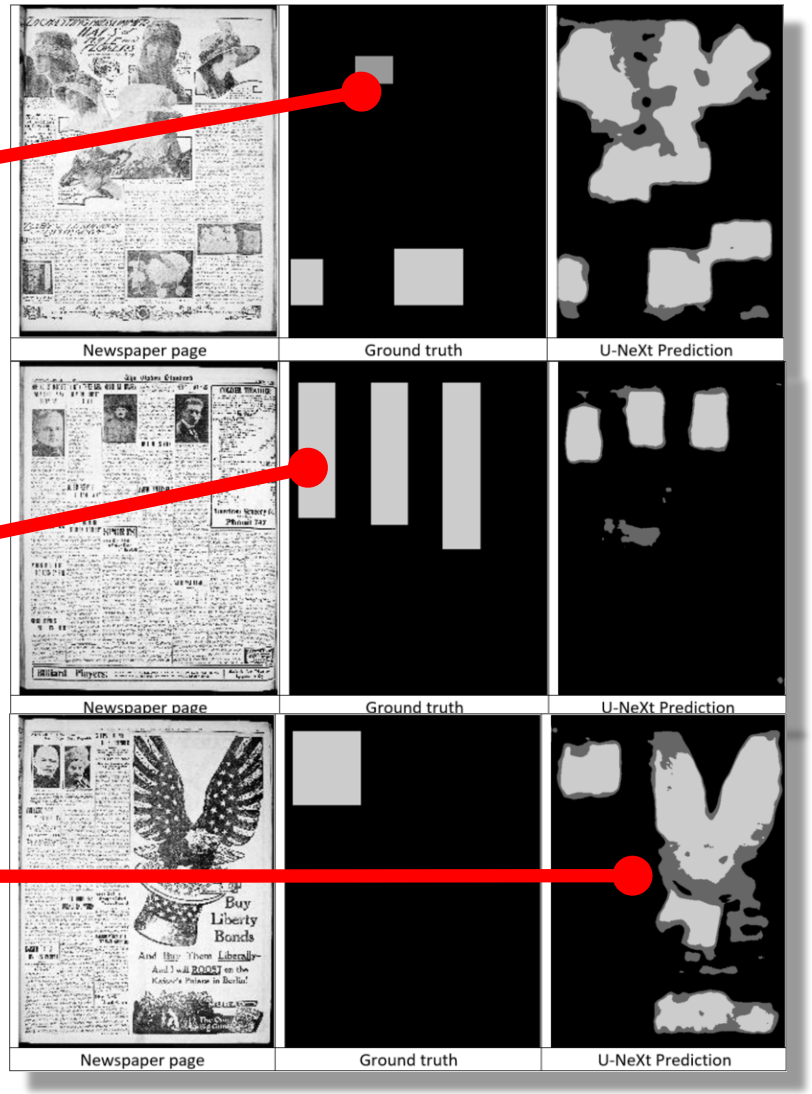
- **Ground truth missing regions**
  - Figure/graph in the document ***missing*** from ground truth but our approach captured missing regions

Ground truth including incorrect regions

- Ground truth includes ***incorrect*** textual content and our approach did not include such textual content

Ground truth not fitting form

- Our approach tried to fit the ***exact shape*** of the figure/graph region



Project 2 Figure/Graph Extraction | Conclusions

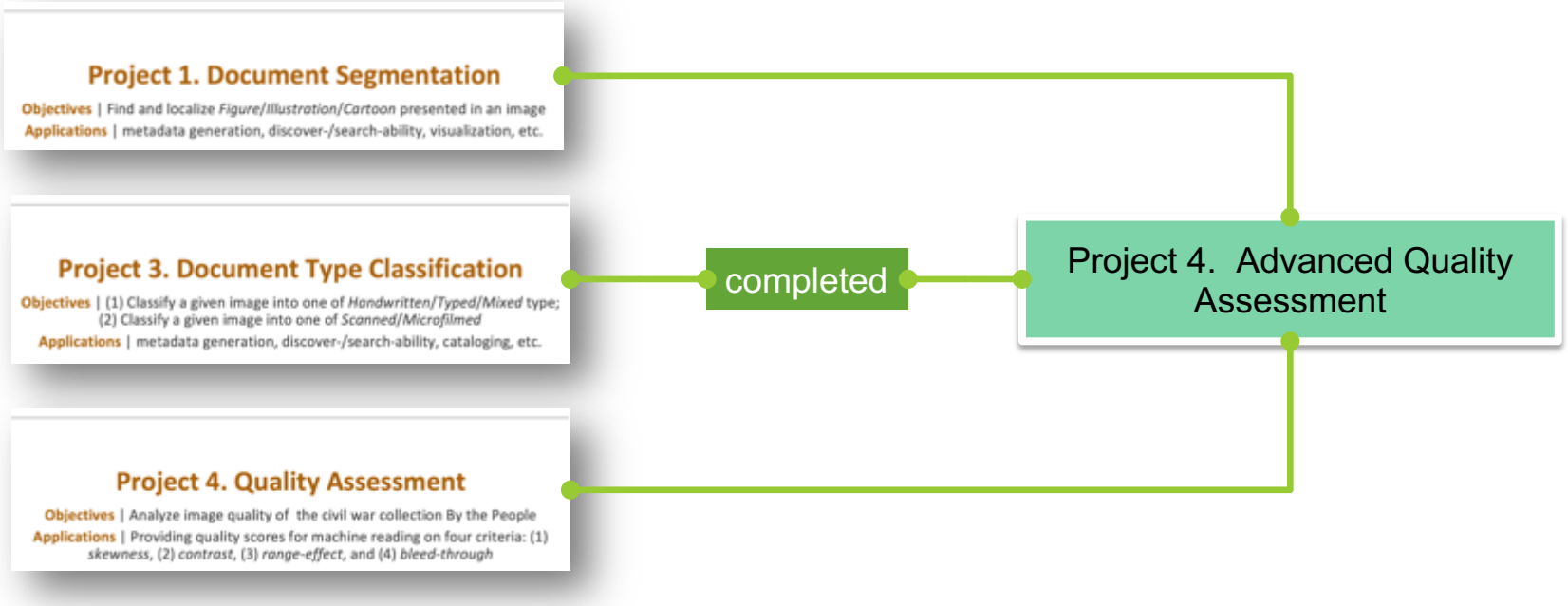
- ***Results*** Actual performance could be better than mIoU indicated, because of issues in the accuracy of ground truth
- ***Challenges*** Ground truth (in)accuracy causes performance loss but ground truth with high accuracy is expensive
- ***Next Iteration*** Build better ground truth only for a small subset for testing and benchmarking purposes
- ***Future Directions*** Combining crowd-sourcing with the U-NeXt model, a *loop-based* system could be built where

  Idea 1 Crowd-sourcing operations receive labeled data from the U-NeXt model, users revise labels, the U-NeXt model improves its predictions based on revision, and repeats

Project 4: Quality Assessment

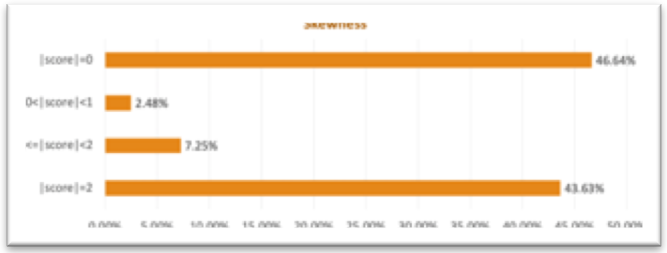
Objectives | Analyze image quality of the civil war collection By the People

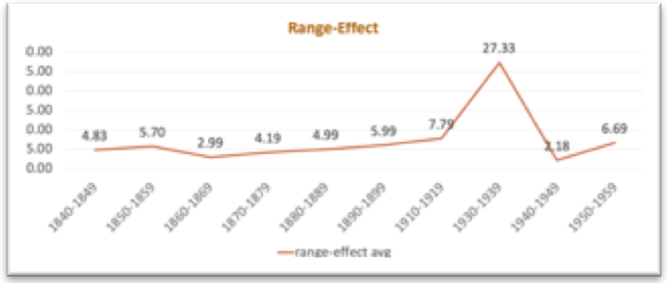
Applications | metadata generation, discover-/search-ability, visualization, etc.

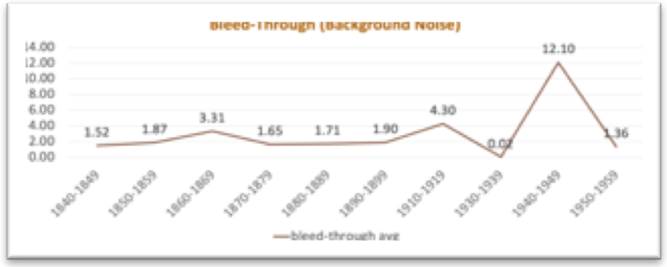


Project 4 Quality Assessment | Where We Were

- **What we did**: Computationally analyzed images for overall image properties
- **Datasets used**: 36,003 images in the civil war collection within *By the People*
- **Implications**
  - A set of image properties (e.g., contrast, range-effect, and bleed-through) can be assessed computationally to understand overall visual attributes of dataset
  - Further investigating relationship (if any) between various visual features and human perception of difficulty (i.e., *difficulty score* collected by the Library of Congress)

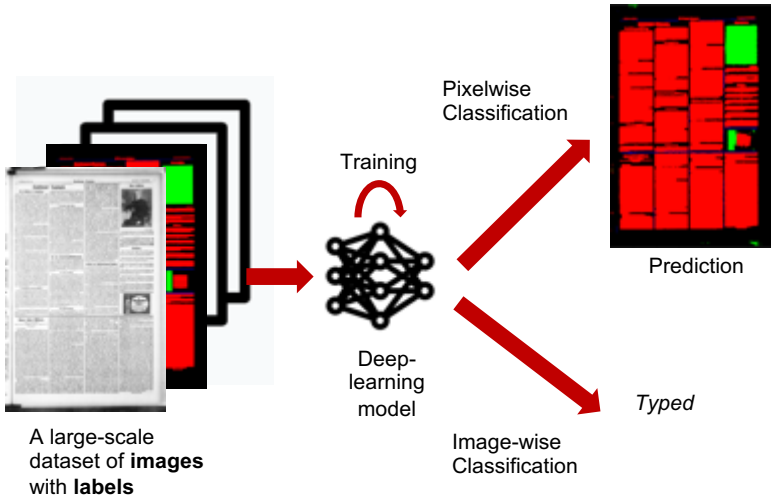


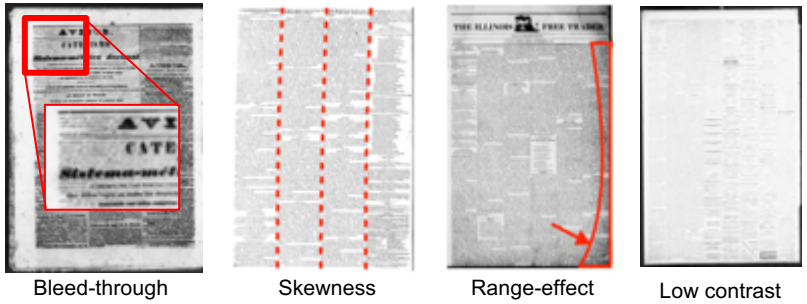




Project 4 Advanced Quality Assessment | 2nd Iteration

- Collect a set of visual quality and visual structural features
  - **1st Iteration Project 1:** Segmented components (e.g., number of zones)
  - **1st Iteration Project 2**: Type of document image (e.g., handwritten, typed, mixed)
  - **1st Iteration Project 4**: Quality of image (e.g., contrast, range-effect, bleed-through)
  - New: Image processing-based connected component analysis (e.g., the number of characters)
  - New: density measures image's compactness
  - New: zone size abnormality measures zone size irregularity
- Compute Pearson's correlation between each visual feature and the difficulty score





Project 4 Advanced Quality Assessment | Results

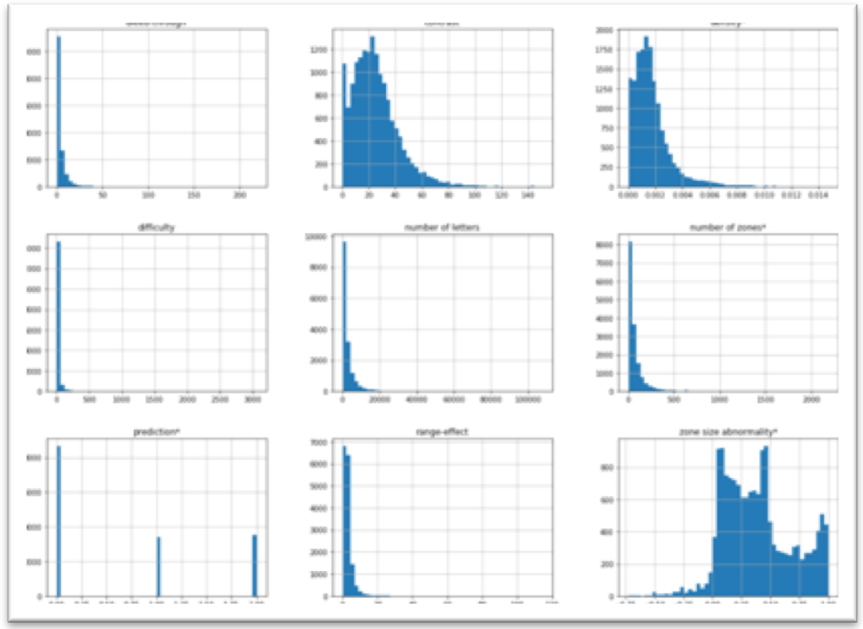
- **Results**
  - **Correlation in general is low, < 0.2**
    - No features investigated has linear correlation with the difficulty score
  - Density, contrast, the number of characters show the highest correlation with the difficulty score

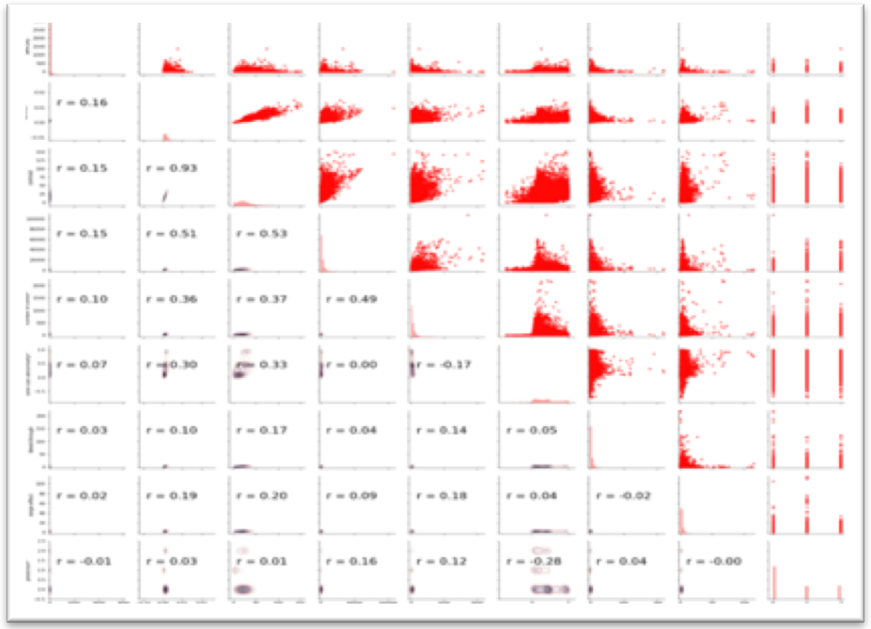
Visual features	Correlation
Density*	0.17
Contrast	0.15
Number of Characters	0.15
Number of Zones	0.10
Zone Size Abnormality*	0.07
Bleed-through	0.03
Range-effect	0.02
Type Prediction	0.01

Correlation between each visual feature and the difficulty score. Note that visual features with asterisks (*) denote *composite visual features* using several visual structural features

Project 4 Advanced Quality Assessment | Another Thought

- **Observations**
  - (Left) Only few visual features are normally distributed
  - (Right) Most relationships are not linear





- **Implication**
  - Explore more complex, non-linear models to capture relationships between visual features and the difficulty score

Project 4 Advanced Quality Assessment | Conclusions

- ***Results*** Using Pearson's correlation was *not* sufficient to capture relationship between visual features and difficulty score
- ***Challenges*** Difficulty score itself may not match human perception of difficulty as complex non-linear relationships exist among visual features
- ***Next Iteration***
  - Analyze correlation between *deep visual feature* (Project 1, latent space) and the difficulty score
  - Explore non-linear models (e.g., polynomial regression, SVM, neural network, etc.)
- ***Future Directions***

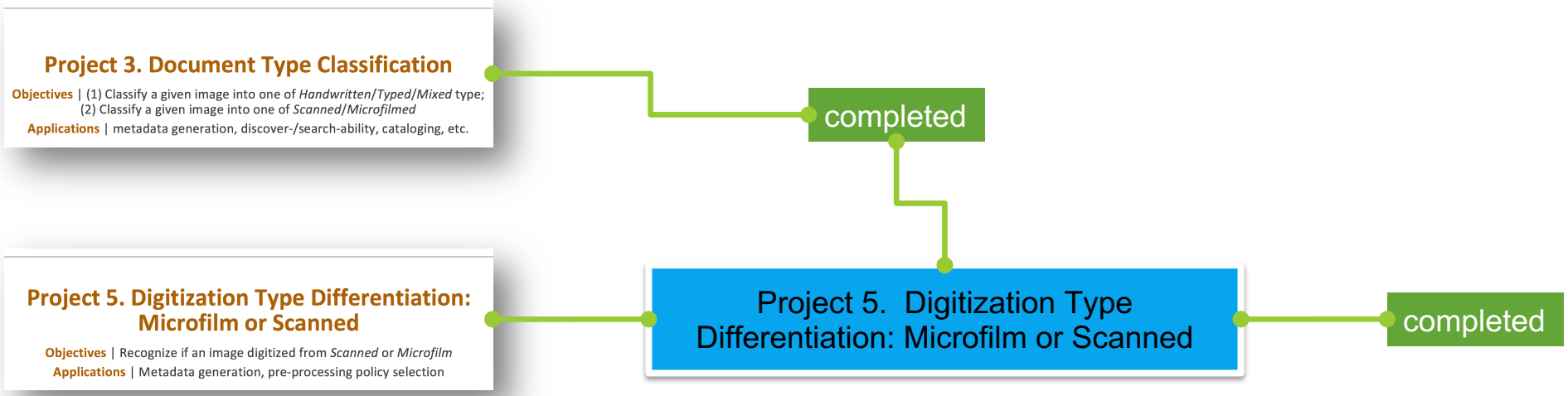
Idea 2 Develop a difficulty assessment model to inform crowd-sourcing operations, e.g., prioritizing images to be transcribed

Idea 3 Assess more accurate and detailed difficulty score (e.g., elicit from users)

Project 5: Digitization Type Differentiation Microfilm or Scanned

Objectives | Recognize if an image digitized from *Scanned* or *Microfilm*

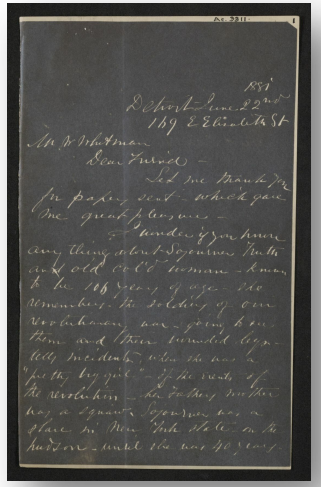
Applications | Metadata generation, pre-processing policy selection



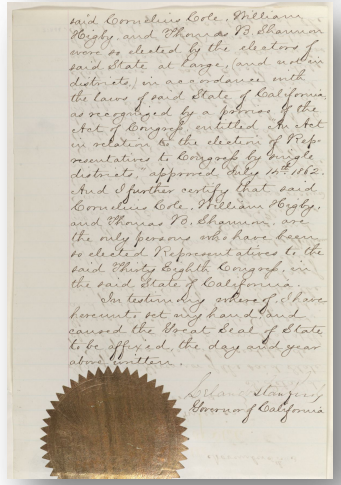
Project 5: Digitization Type Differentiation Microfilm or Scanned |

Where We Were

- **What we did**: Used ResNeXt [6] (a deep learning method) to differentiate document images digitized from microfilm or scanning processes
- **Datasets used**: 1,200 images from the Civil War collection on By The People repository
  - Dataset was manually built by randomly choosing 600 scanned materials and 600 microfilm materials
  - Randomization ensured each image a fair chance to be chosen
- **Results**: 100% accurate for 120 test samples



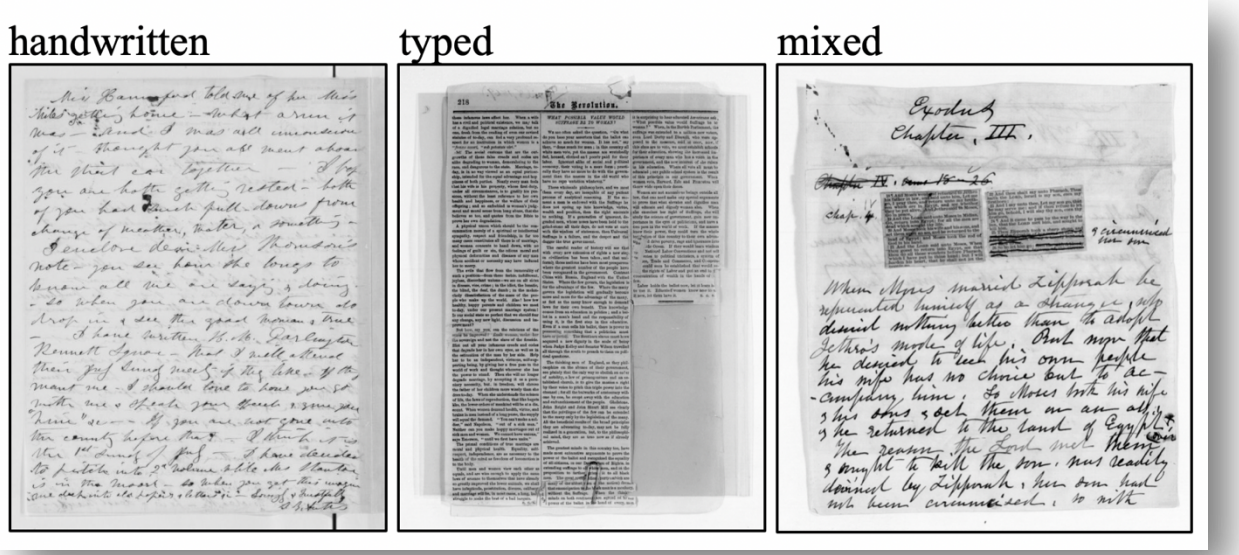
microfilm



scanned

Project 5: Digitization Type Differentiation Microfilm or Scanned | Where We Were | Project 3

- **What we did**: Used VGG-16 [12] (a deep learning method) to differentiate document images to handwritten, typed or mixed images
- **Datasets used**: 1,002 images from the Suffrage campaign collection from the By The People repository
- **Results**: 89.90% accurate for 99 test samples



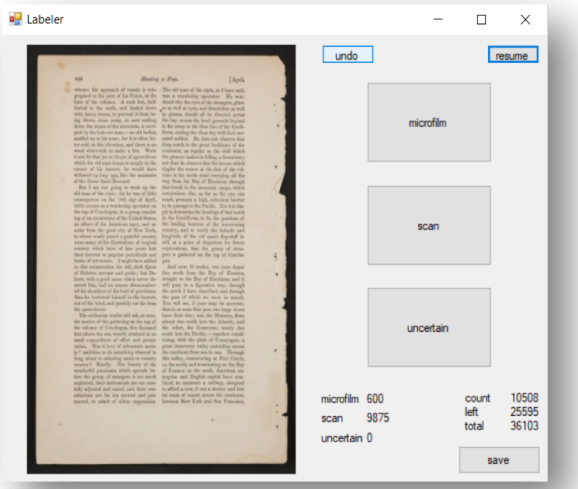
[12] Afzal, M. Z., Kölsch, A., Ahmed, S., & Liwicki, M. (2017, November). Cutting the error by half: Investigation of very deep CNN and advanced training strategies for document image classification. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*(Vol. 1, pp. 883-888). IEEE.

Project 5: Digitization Type Differentiation Microfilm or Scanned | Where We Were

- **Implications**
  - Almost perfect accuracy might be *too good to be true*—possibly due to the bias in the small sample dataset—we proposed to compare the *classified* ratio of microfilm to scanned materials with the *ground truth* ratio of microfilm to scanned materials in the entire Civil War collection to evaluate our approach comprehensively
  - Digitization type differentiation using deep learning ***is promising***
    - Enrich metadata tagging by recognizing digitization type automatically
  - **Note**: Based on 10,508 visually inspected images, ***"ground truth" ratio of microfilm to scanned materials is about 1:16***

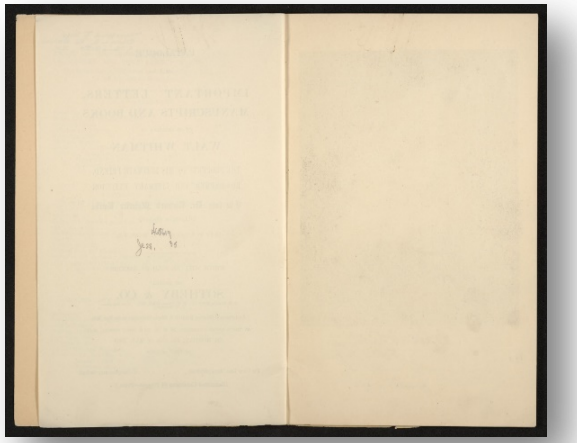
Project 5: Digitization Type Differentiation Microfilm or Scanned | 2nd Iteration

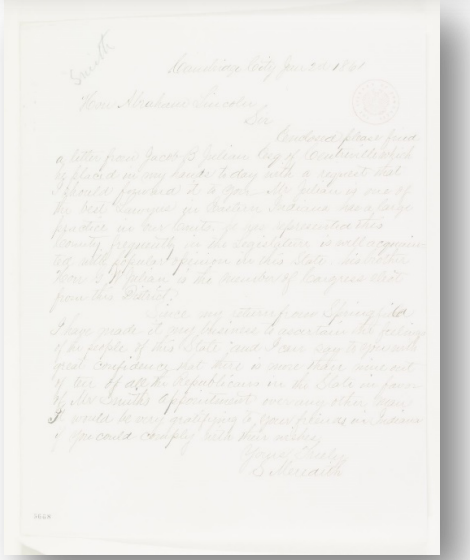
- **Processing additional document images**
  - The entire Civil War collection on By The People repository (36,103 images)
  - 3% images involved in training
  - Ground truth Microfilm : Scanned ratio = 1 : 16
  - Classified Microfilm : Scanned ratio = 1 : 11.74
- **Implication**
  - Classifier labeled more images as microfilm
    - **Why?**



Project 5: Digitization Type Differentiation Microfilm or Scanned | 2nd Iteration

- **Findings: Issues with ground truth**
- **Missing scenarios in training data**
  - Largely “blank” document images
  - Poor contrast document images
  - Not a document image





Project 5: Digitization Type Differentiation Microfilm or Scanned | Conclusions

microfilm

- ***Results*** Automated type differentiation is viable and computationally cheap, and our approach can be applied to a large collection
  - Our approach had never seen microfilm photos in the training set, but it made the correct classification
- ***Challenges*** Small training sets may not contain some scenarios
- ***Next Iteration***
  - Sample dataset comprehensively to identify all possible scenarios and add to the training dataset
  - Perform color inversion for microfilm images

Project 5: Digitization Type Differentiation Microfilm or Scanned | Conclusions

- ***Future Direction***
  - Add search by "digitization type" for query parameters

Idea 4

Explore data programming tool (e.g., Snorkel [7]) that uses heuristics rules to generate pseudo-ground truth
  - Could reduce human effort for ground truth construction

[7] A. J. Ratner, C. M. De Sa, S. Wu, D. Selsam, and C. Re, "Data ´ programming: Creating large training sets, quickly," in Advances in neural information processing systems, 2016, p 3567–3575.

Aida

Q&A

2nd Iteration

Project 1: Document Segmentation

Objectives | Find and localize *Figure/Illustration/Cartoon* presented in an image
Applications | metadata generation, discover-/search-ability, visualization, etc.

Project 2: Figure/Graph Extraction

Objectives | Find and localize *Figure/Illustration/Cartoon* presented in an image
Applications | metadata generation, discover-/search-ability, visualization, etc.

Project 4: Quality Assessment

Objectives | Analyze image quality of the civil war collection By the People
Applications | metadata generation, discover-/search-ability, visualization, etc.

Project 5: Digitization Type Differentiation Microfilm or Scanned

Objectives | Recognize if an image digitized from *Scanned* or *Microfilm*
Applications | Metadata generation, pre-processing policy selection

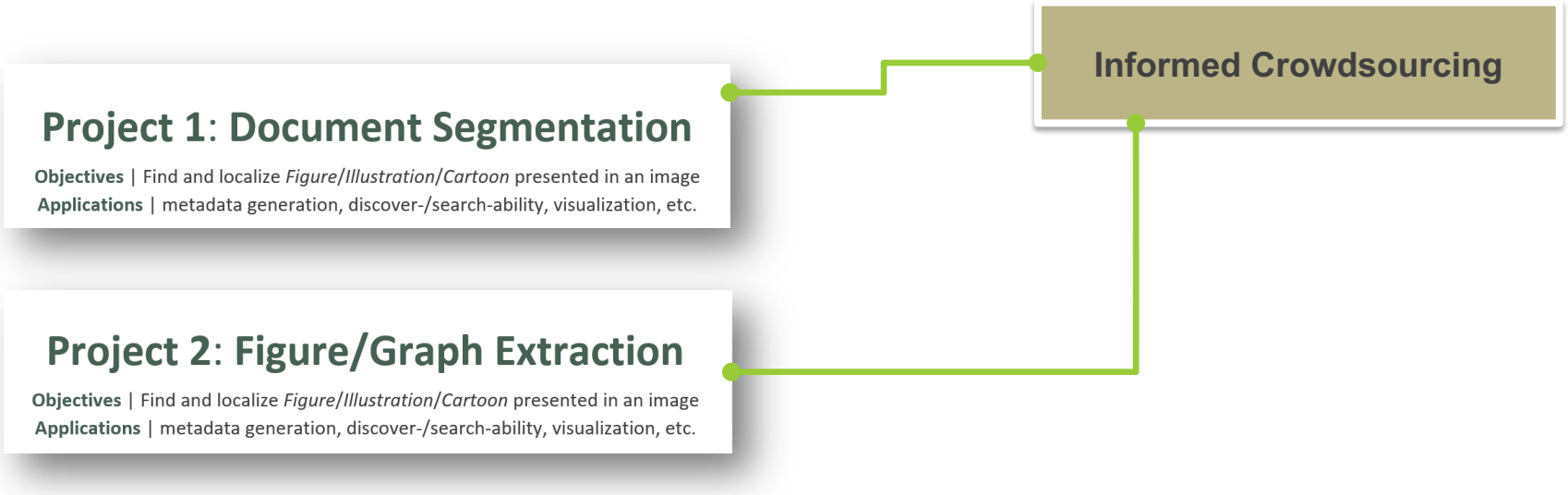
Future Direction

- Informed Crowdsourcing — Idea 1
- Enriched Metadata — Idea 2
- Benchmarking — Idea 3
- Low-Cost Groundtruthing — Idea 4
- Deep Learning — Idea 5

Informed Crowdsourcing

Objectives | Allow machine learning models to cumulatively improve their performance

Motivations | The need for an effective ground truthing approach for hard tasks



Idea 1: Informed Crowdsourcing | Advantages

- Crowdsourcing involves a large pool of users who provide inputs, which improve:
  - Quality
  - Flexibility
  - Scalability
  - Diversity

  → The same properties a good training dataset should have

- Above benefits make crowdsourcing a good approach to collect training data for machine learning models

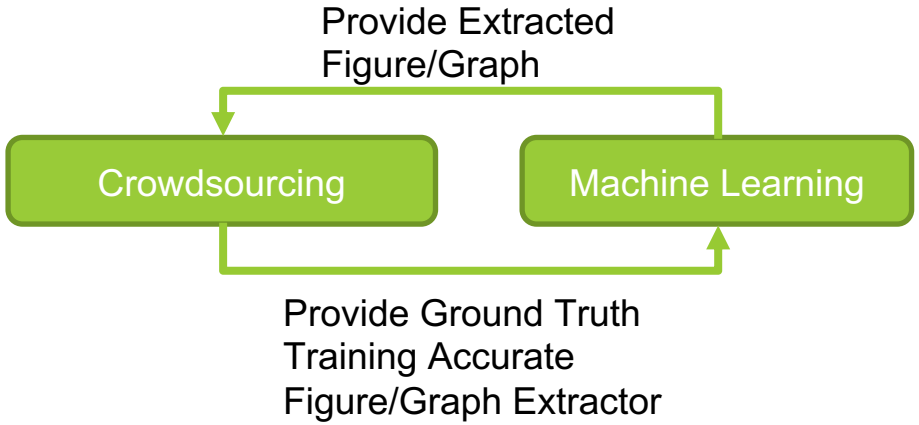
Idea 1: Informed Crowdsourcing | Motivation

- **Widely used**
  - It is not a secret that companies collect user data to improve their product
    - E.g., Google's search engine
- **Not limited** to task with deterministic heuristic rules
  - i.e., crowdsourcing can deal with problems involving high-level understanding, such as *metadata defining* and *document layout analysis*
- Can **support ground truth construction** without deterministic rules
  - Ground truth with deterministic rules, meaning that, the labeling process can be described as a yes or no question

Idea 1: Informed Crowdsourcing | Potential Application

- Recall **Project 2**
  - The three ground truth issues for figure/graph extraction
    - (1) Missing regions; (2) incorrect regions; (3) inaccurate region shapes
  - By taking advantage of the large pool of people, crowdsourcing could refine the ground truth with much less labor-cost, at the sometime, with good quality

- **With informed crowdsourcing**, a loop-based system could be built to improve our U-NeXt models
  - Crowd-sourcing operations receive labeled data from the U-NeXt model, users revise labels, the U-NeXt model improves its predictions based on revision, and repeats



Enriched Metadata

Objectives | Improve accessibility and searchability of digital libraries

Motivations | The need for enriched any-level searchability

Enriched Metadata

Project 4: Quality Assessment

Objectives | Analyze image quality of the civil war collection By the People

Applications | metadata generation, discover-/search-ability, visualization, etc.

Idea 2: Enriched Metadata | Different Needs

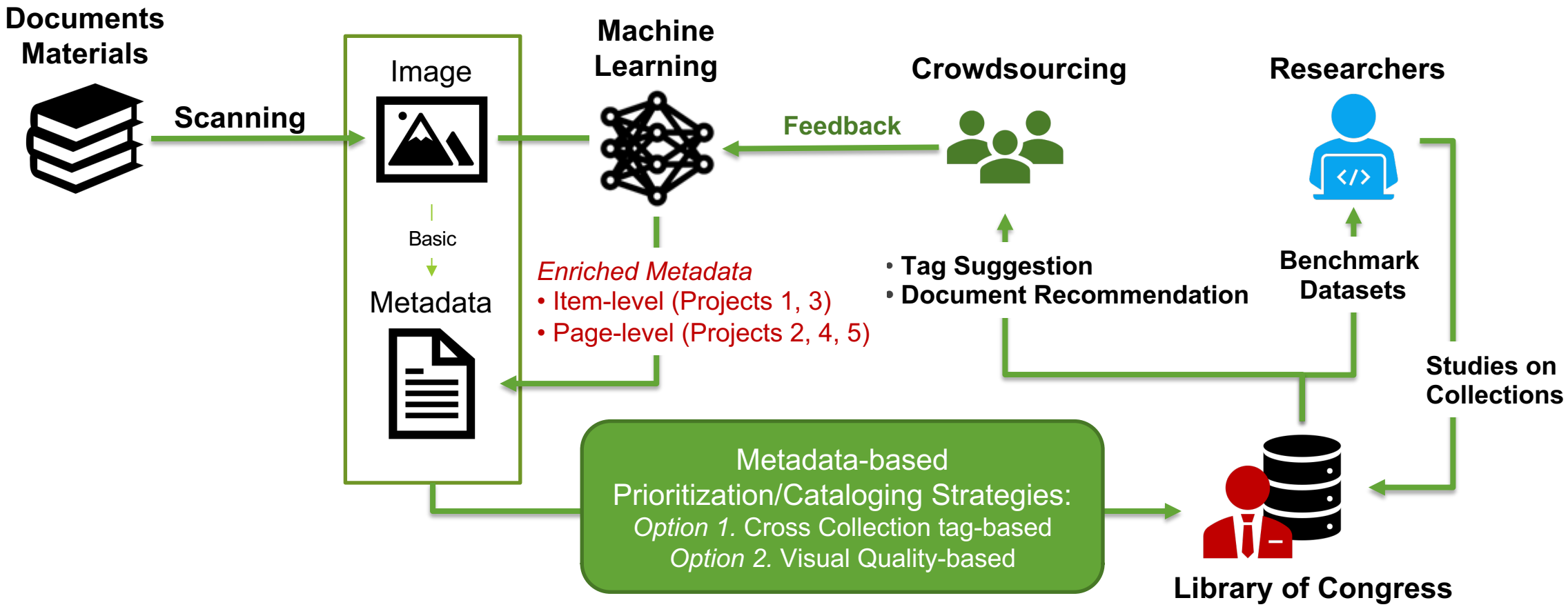
- *Metadata is key for improving **Discoverability** and **Accessibility**!*
  - ***From internal (Upstream):*** "How to maintain items effectively and efficiently?"
    - Focus of ***data administrators***: Item cataloging/prioritizing strategy
  - ***From external (Downstream):*** "Can I find the right items conveniently that I am looking for?"
    - Focus of ***general users***: page-level documents; basic metadata (e.g., "retrieve all document images from newspapers published 1836-1840")
    - Focus of ***researchers***: batch-level documents (benchmark datasets); enriched metadata is essential (e.g., "retrieve all document images that are of low contrast")

Idea 2: Enriched Metadata | Types of Metadata

- ***Basic metadata*** *(Easily obtainable at low cost)*
  - Image resolution
  - Generated data/time
  - Poor quality OCR
- ***Enriched metadata*** *(Not easily obtainable automatically; requires high cost)*
  - Keywords tagged by crowdsourcing
  - High quality OCR
  - (Project 1) Structural information (e.g., location of articles)
  - Logical relationships between substructures (e.g., reading-order)
  - (Project 4) Objective/subjective visual quality (e.g., contrast, noise, range effects)

Bailer W. et al. (2008) Multimedia Metadata Standards. In: Furht B. (eds) Encyclopedia of Multimedia. Springer, Boston, MA

Idea 2: Enriched Metadata | Ecosystem



Benchmark Datasets

Objectives | Create standard databases to evaluate approaches

Motivations | A shared database can encourage systematic rigorous research towards finding better approaches

Benchmarking

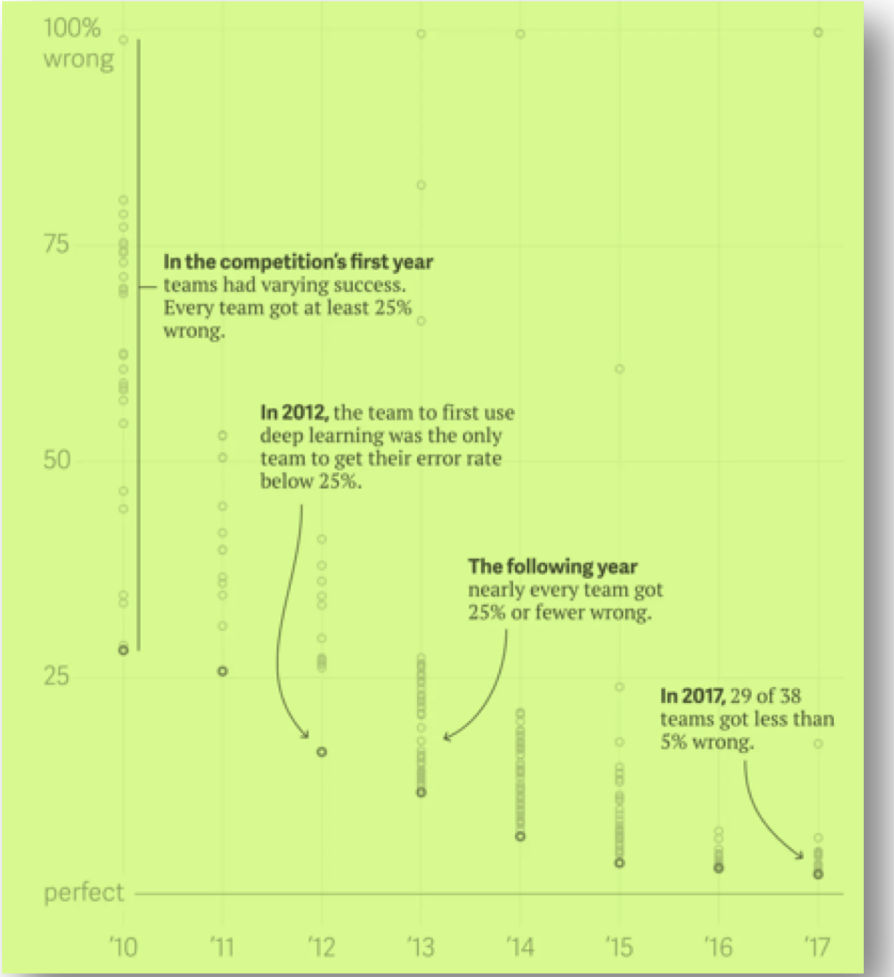
Project 4: Quality Assessment

Objectives | Analyze image quality of the civil war collection By the People
Applications | metadata generation, discover-/search-ability, visualization, etc.

Idea 3: Benchmark Datasets | Data, Data, Data!

"We don't have better algorithms. We just have more data."
– Peter Norvig

- ***ImageNet?***
  - ImageNet is a large-scale ***natural scene*** image dataset
  - ImageNet Challenge boosts image and vision research field vastly



Halevy, Alon, Peter Norvig, and Fernando Pereira. "The unreasonable effectiveness of data." (2009).

Idea 3: Benchmark Datasets | "DocuNet"

- ***Why not "DocuNet"?***
  - Existing document image datasets have following constraints
    - Datasets are ***NOT*** publicly available
    - Volume is relatively small (e.g., less than 1,000 images)
  - ***One effective way to encourage image/vision/ML researchers to delve into document analysis field***
- ***What do we need?***

Idea 4

  - Ground-truth (e.g., document types, coordinates of article regions, etc.)
  - Openness
  - Diversity & Balance (e.g., different document types should be comprehensively covered and equally distributed)
  - Clear objectives (e.g., segmentation, classification, clustering, etc.)

Idea 4

Low-Cost Groundtruthing

Objectives | Build ground truth for machine learning models in a low-cost fashion

Motivations | Having subject matter experts' hand-label data is expensive

Project 5: Digitization Type Differentiation Microfilm or Scanned

Objectives | Recognize if an image digitized from *Scanned* or *Microfilm*

Applications | Metadata generation, pre-processing policy selection

Low-Cost Groundtruthing

Idea 4: Low-Cost Groundtruthing | Traditional Approach

- Groundtruthing is a process of collecting labeled data for machine learning training based on observation
  - E.g., the process of collecting the 1,200 images for Project 5
- **Traditional approach**: Having subject matter experts' hand-label data
  - Advantages: Accurate
  - Disadvantages: Time and labor intensive

Idea 4: Low-Cost Groundtruthing | Motivation

- The training process of machine learning approaches typically needs **lots of** labeled data
  - E.g., One of the most comprehensive benchmarking dataset, ImageNet[11], contains *1,431,167* labeled images
- Both the hand-labeling processes and finding subject matter experts are expensive
  - E.g., when building the ground truth for our Project 5
    - Consulting experts
    - More than 4 hours
    - Only a 1,200-image groundtruth was built
- Need to get labeled data **cheaper** and **faster**

[11] J. Deng, W. Dong, R. Socher, L. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, 2009, pp. 248-255.

Idea 4: Low-Cost Groundtruthing | A Modern Approach

- **Weak** supervision [12]
  - Computers label data using heuristic rules, constraints, distributions, or/and invariances of the dataset
- Instead of having experts to hand-label data, only need to consult an expert on *how* to label data
- Advantages: Computers take care of the heavy-lifting of data-labeling
- Disadvantages: Less accurate groundtruth *(but could still produce very accurate classifier!)*

[12] Ratner, A., Bach, S., Varma, P. and Ré, C., 2017. Weak Supervision: The New Programming Paradigm for Machine Learning.

Idea 4: Low-Cost Groundtruthing | An Example

- Snorkel [7]
  - A system for programmatically building and managing training datasets
- **Cheap**
  - Develop a labeling program based on heuristic rules
- **Fast**
  - Develop training datasets in hours or days rather than hand-labeling them over weeks or months

Idea 4: Low-Cost Groundtruthing | Potential Application

- **Recall our <u>Project 5</u>**
  - Performance was reduced due to *missing scenarios* in the groundtruth
    - Only a little more than 10,000 images were inspected for the groundtruth construction → likely more than just three scenario types that were not included for training
- **With Snorkel**
  - By converting heuristic rules into a program, one could reduce a human visual labeling task to less than a second
    - Example heuristic rule: any images with colors are scanned materials
  - Resultant "groundtruth" data could be noisy …
  - **… but contain *significantly more* document images with *significantly more different* scenarios, especially needed to train deep learning models**

Idea 5

Applying Deep Learning

Objectives | Apply deep learning models to analyze documents in digital library

Motivations | Different deep learning models have different properties appropriate for different tasks

Applying Deep Learning

Aida

Idea 5: Applying Deep Learning | Choose Deep Learn Model By Task

Task Type	Task Properties	Suitable Models	Examples
Document layout analysis	Need pixel-level understanding	U-shaped models e.g., dhSegment, U-NeXt	Project 2
Document categorization	Need page-level recognition	Convolutional neural networks e.g., ResNet, ResNeXt	Projects 3 and 5
Audio/video understanding	Sequential data understanding	Recurrent neural networks	

Idea 5: Applying Deep Learning |
Choose Learning Scheme By Existence of Labeled data

Is There Labeled Data?	Learning Scheme	Examples
Yes	Supervised Learning	Projects 2, 3 and 5
No	Unsupervised Learning	Projects 1 and 4

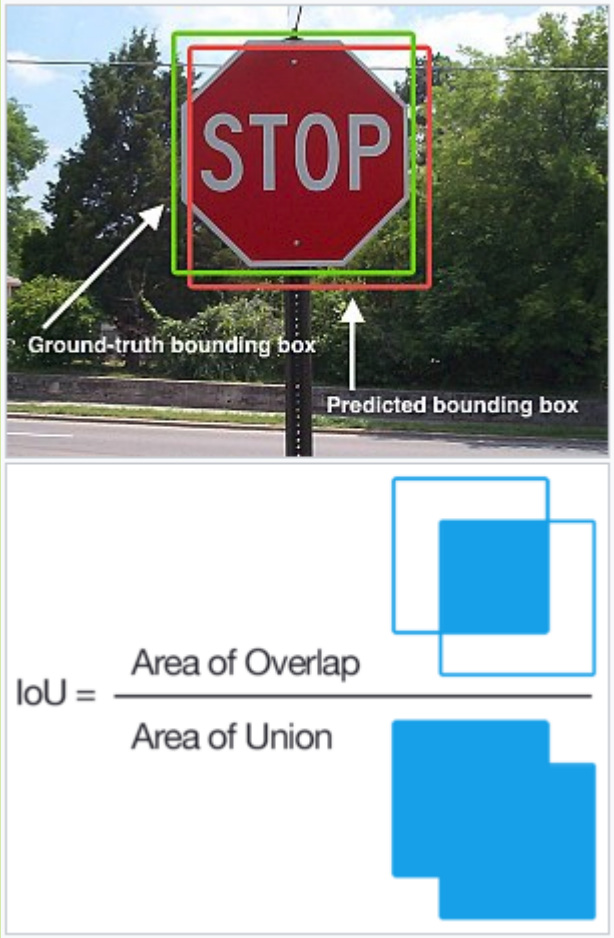
Q&A

Appendix

Back

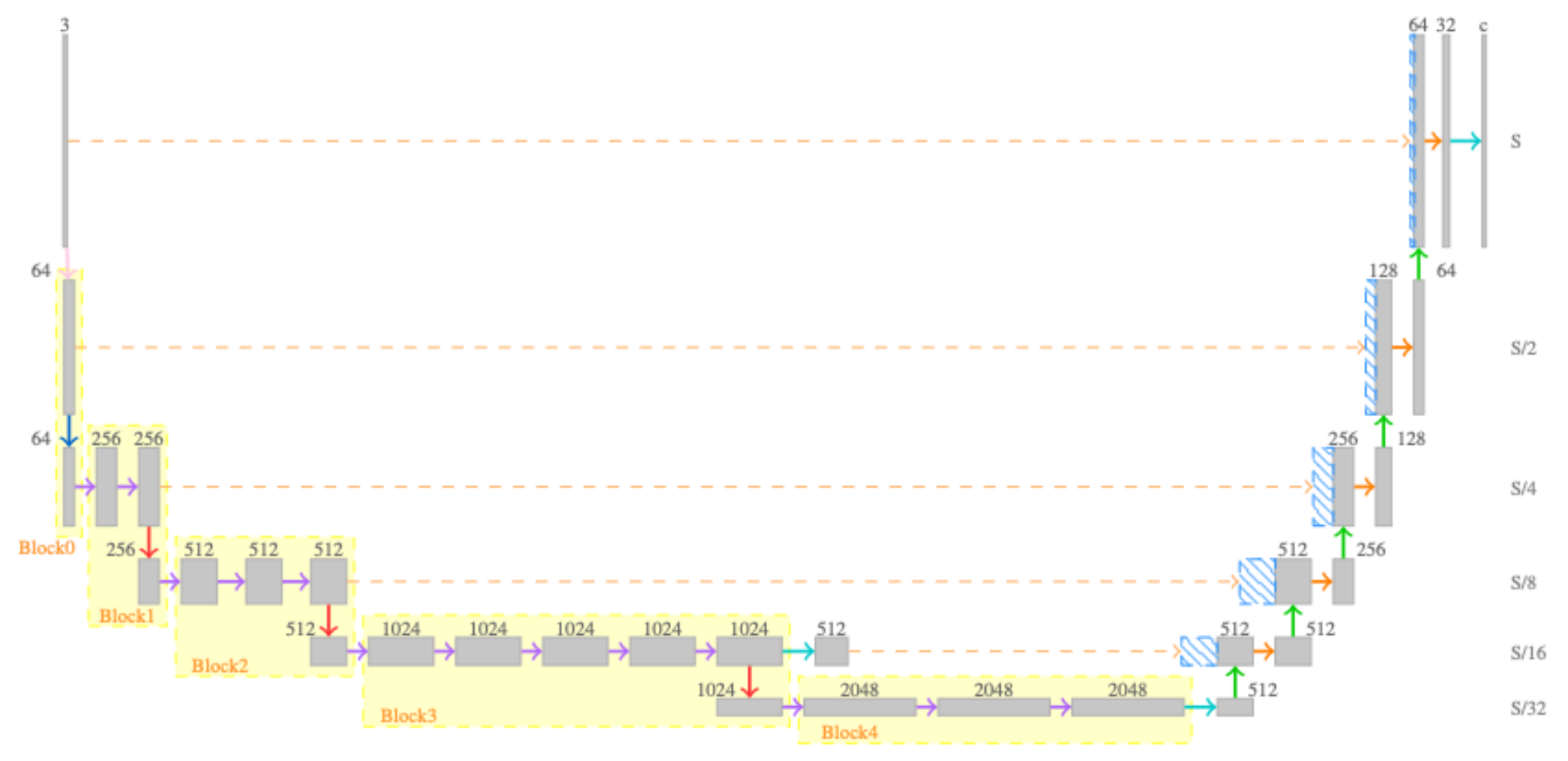
Mean Intersection over Union - mIoU

mIoU is a statistic used for gauging the similarity and diversity of extracted region to the ground truth



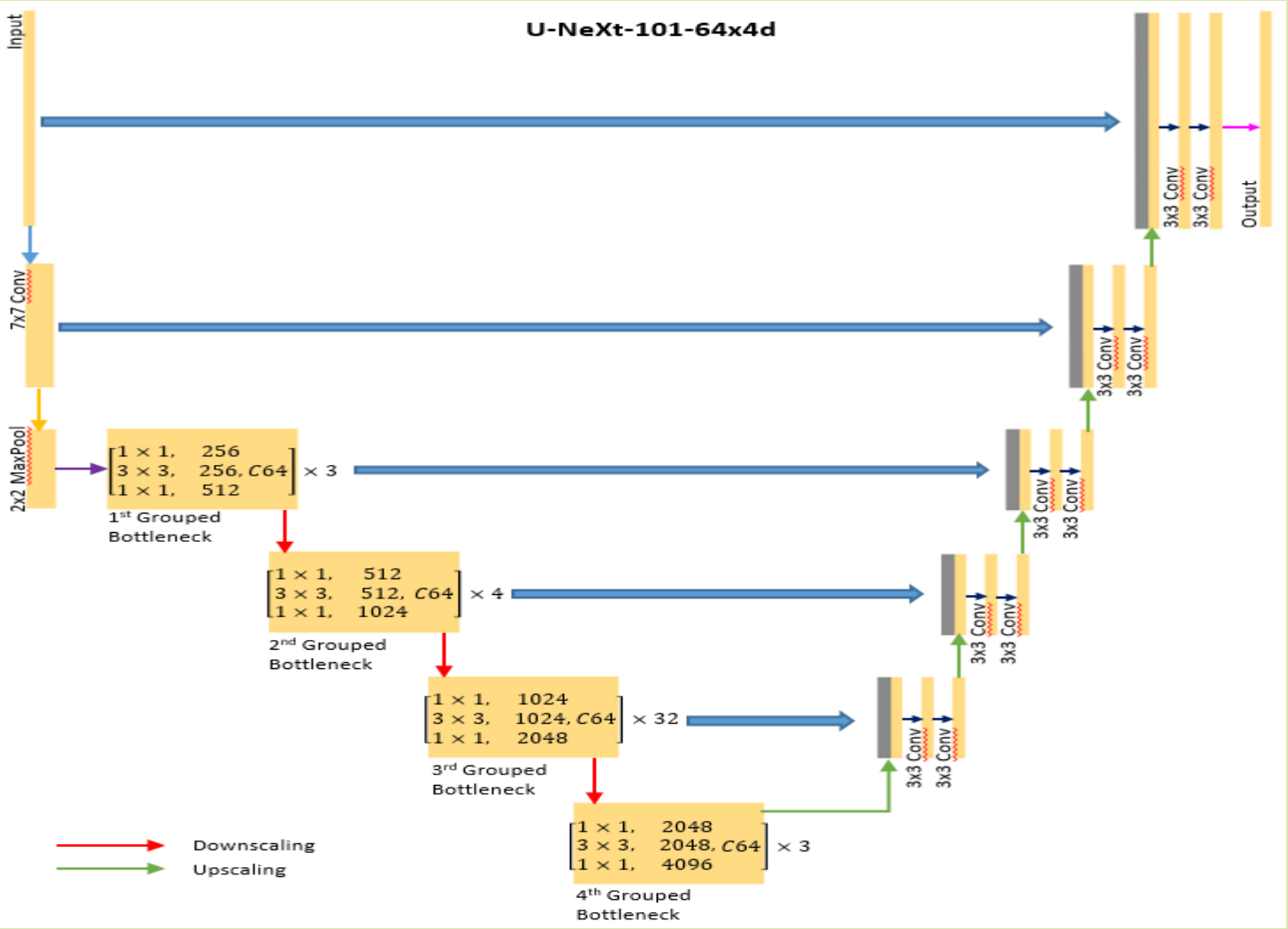
Back

The dhSegment Model



[1] Oliveira, Sofia Ares, Benoit Seguin, and Frederic Kaplan. "dhSegment: A generic deep-learning approach for document segmentation." *2018 16th International Conference on Frontier Handwriting Recognition (ICFHR)*. IEEE, 2018.

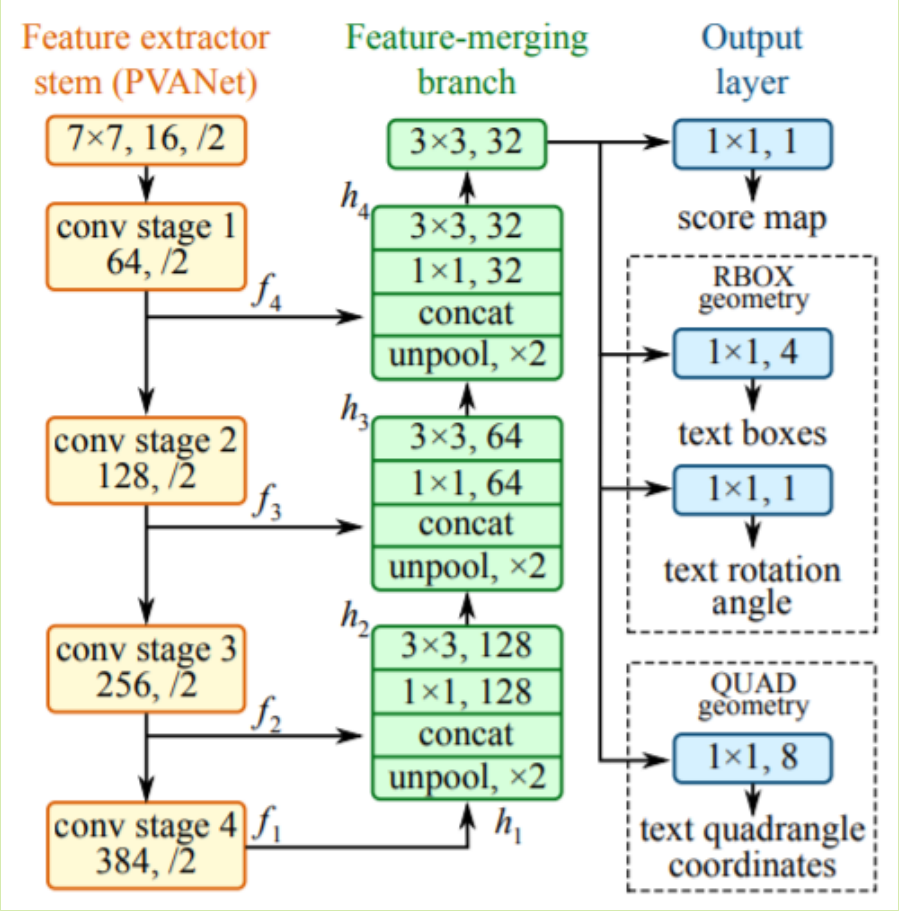
U-NeXt Model



[6] S. Xie, R. Girshick, P. Dollár, Z. Tu and K. He, "Aggregated Residual Transformations for Deep Neural Networks," 2017 IEEE Conf. on CVPR, 2017, pp. 5987-5995.

Efficient and Accurate Scene (EAST) Text Detector | Project 2.2 (1st Iteration)

- **HyperNet [8] + U-Net [9]**
  - Detect texts in graphic images in any direction
- Why applicable?
  - Figures/illustrations are snippets of a graphic region



[8] T. Kong, A. Yao, Y. Chen and F. Sun, "HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 845-853.
[9] O. Ronneberger, P.Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Medical Image Computing and Computer-Assisted Intervention (MICCAI), ser. LNCS, vol. 9351. Springer, 2015, pp. 234–241,

Back

Efficient and Accurate Scene (EAST) Text Detector | Project 2.2 (1st Iteration)

Detected Texts

- Performance on detecting texts in newspaper figure/graph is good
- Text location is recorded

Text Lines

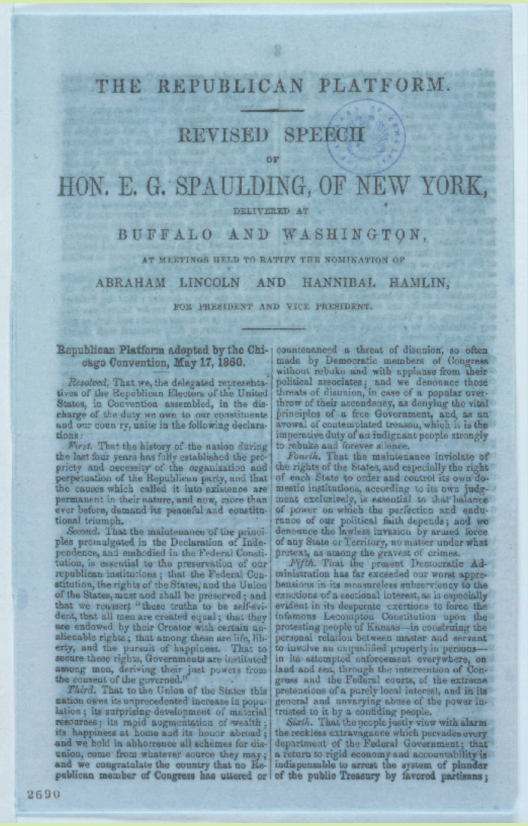
- 6 text lines
- { "x0": 62, "y0": 608, "x1": 135, "y1": 588, "x2": 143
- { "x0": 188, "y0": 33, "x1": 312, "y1": 31, "x2": 313,
- { "x0": 331, "y0": 31, "x1": 423, "y1": 30, "x2": 423,
- { "x0": 116, "y0": 34, "x1": 166, "y1": 33, "x2": 166,
- { "x0": 405, "y0": 755, "x1": 470, "y1": 757, "x2": 47
- { "x0": 475, "y0": 756, "x1": 531, "y1": 757, "x2": 53

Back

Visual Features (Density) | Project 4 (2nd Iteration)

- Density
  - A way of representing image's compactness

$$\frac{Area(pixels \in foreground)}{resolution}$$



THE REPUBLICAN PLATFORM.

REVISED SPEECH

OF

HON. E. G. SPAULDING, OF NEW YORK,

DELIVERED AT

BUFFALO AND WASHINGTON,

AT MEETINGS HELD TO RATIFY THE NOMINATION OF

ABRAHAM LINCOLN AND HANNIBAL HAMLIN,

FOR PRESIDENT AND VICE PRESIDENT.

Republican Platform adopted by the Chicago Convention, May 17, 1860.

Resolved, That we, the delegated representatives of the Republican Electors of the United States, in Convention assembled, in the discharge of the duty we owe to our constituents and our country, unite in the following declarations:

First. That the history of the nation during the last four years has fully established the propriety and necessity of the organization and perpetuation of the Republican party, and that the causes which called it into existence are permanent in their nature, and now, more than ever before, demand its peaceful and constitutional triumph.

Second. That the maintenance of the principles promulgated in the Declaration of Independence, and embodied in the Federal Constitution, is essential to the preservation of our republican institutions; that the Federal Constitution, the rights of the States, and the Union of the States, must and shall be preserved; and that we reaffirm "these truths to be self-evident, that all men are created equal; that they are endowed by their Creator with certain unalienable rights; that among these are life, liberty, and the pursuit of happiness. That to secure these rights, Governments are instituted among men, deriving their just powers from the consent of the governed."

Third. That to the Union of the States this nation owes its unprecedented increase in population; its surprising development of material resources; its rapid augmentation of wealth; its happiness at home and its honor abroad; and we hold in abhorrence all schemes for disunion, come from whatever source they may; and we congratulate the country that no Republican member of Congress has uttered or countenanced a threat of disunion, so often made by Democratic members of Congress without rebuke and with applause from their political associates; and we denounce these threats of disunion, in case of a popular overthrow of their ascendancy, as denying the vital principles of a free Government, and, as an avowal of contemplated treason, which it is the imperative duty of an indignant people strongly to rebuke and forever silence.

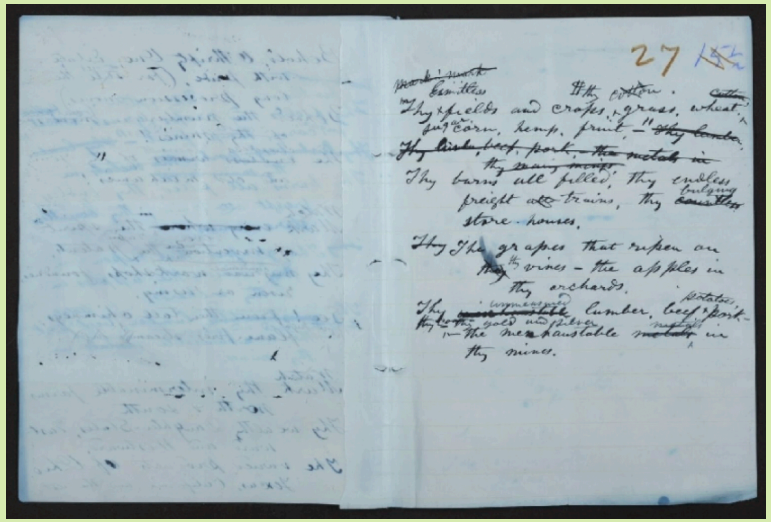
Fourth. That the maintenance inviolate of the rights of the States, and especially the right of each State to order and control its own domestic institutions, according to its own judgment exclusively, is essential to that balance of power on which the perfection and endurance of our political faith depends; and we denounce the lawless invasion by armed force of any State or Territory, no matter under what pretext, as among the gravest of crimes.

Fifth. That the present Democratic Administration has far exceeded our worst apprehensions in its measureless subserviency to the exactions of a sectional interest, as is especially evident in its desperate exertions to force the infamous Lecompton Constitution upon the protesting people of Kansas—in construing the personal relation between master and servant to involve an unparalleled property in persons—in its attempted enforcement everywhere, on land and sea, through the intervention of Congress and the Federal courts, of the extreme pretensions of a purely local interest, and in its general and unvarying abuse of the power intrusted to it by a confiding people.

Sixth. That the people justly view with alarm the reckless extravagance which pervades every department of the Federal Government; that a return to rigid economy and accountability is indispensable to arrest the system of plunder of the public Treasury by favored partisans;

2690

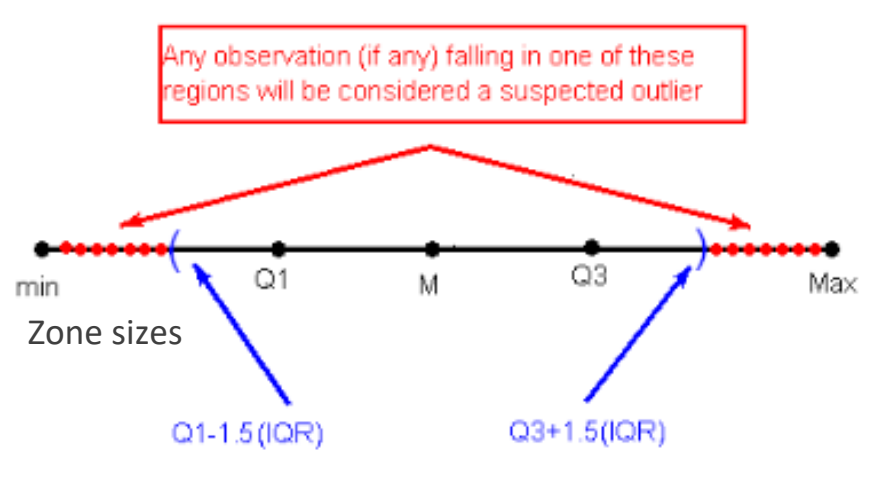
0.45



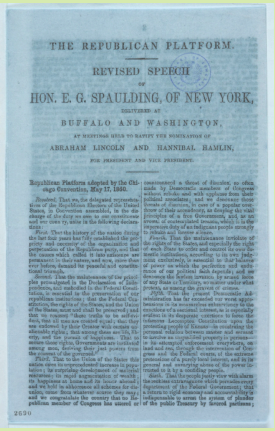
0.27

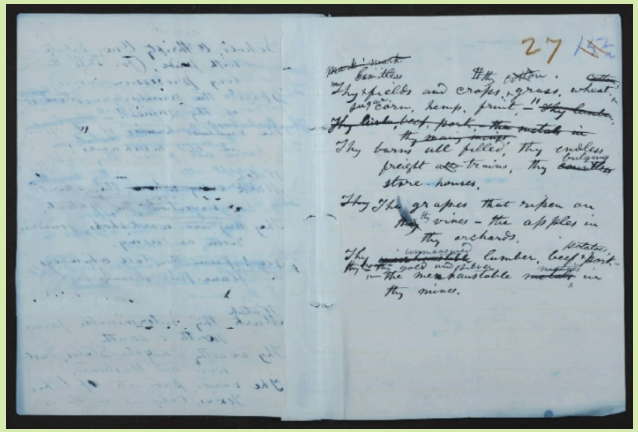
Visual Features (Zone Size Abnormality) | Project 4 (2nd Iteration)

- Zone Size Abnormality
  - A way of representing irregularity of the size of zone



$$1 - \frac{(Q3 + 1.5 * IQR) - (Q1 - 1.5 * IQR)}{Max - Min}$$





0.88

0.95

Back

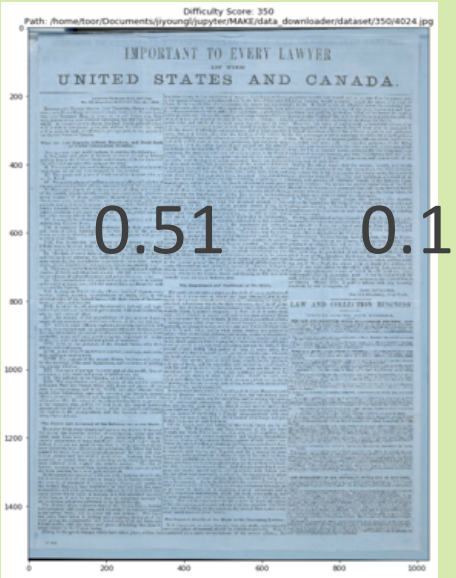
Visual Features (Zone Size Abnormality) | Project 4 (2nd Iteration)

Density

$$\frac{Area(pixels \in foreground)}{resolution}$$

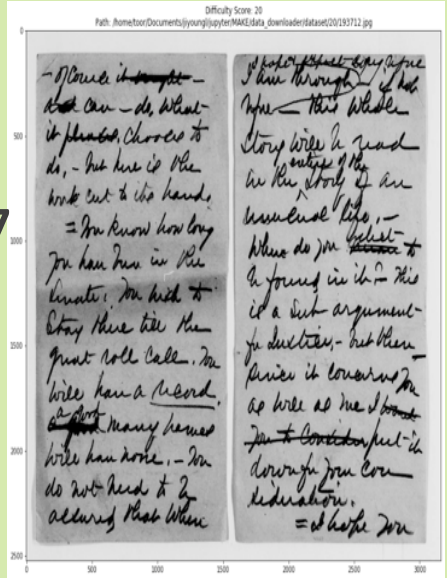
Zone size abnormality

$$\frac{||Zone \in I||}{I_{resolution}} \; where \; Area(Zone) \in [Zone_{Q_1} - 1.5 * Zone_{IQR}, Q_3 + 1.5 * Zone_{IQR}]$$



0.51

0.17



Text Lines

- 6 text lines
- { "x0": 62, "y0": 608, "x1": 135, "y1": 588, "x2": 143
- { "x0": 188, "y0": 33, "x1": 312, "y1": 31, "x2": 313,
- { "x0": 331, "y0": 31, "x1": 423, "y1": 30, "x2": 423,
- { "x0": 116, "y0": 34, "x1": 166, "y1": 33, "x2": 166,
- { "x0": 405, "y0": 755, "x1": 470, "y1": 757, "x2": 47
- { "x0": 475, "y0": 756, "x1": 531, "y1": 757, "x2": 53

Back

ResNet [5] vs. ResNeXt [6]

- The ResNeXt and ResNet models have the same layer architecture
- However, ResNeXt split one building block into multiple branches

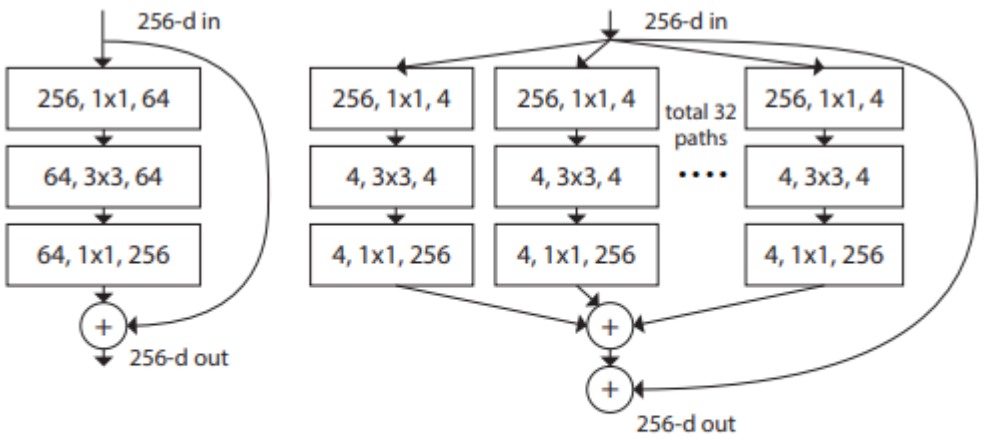
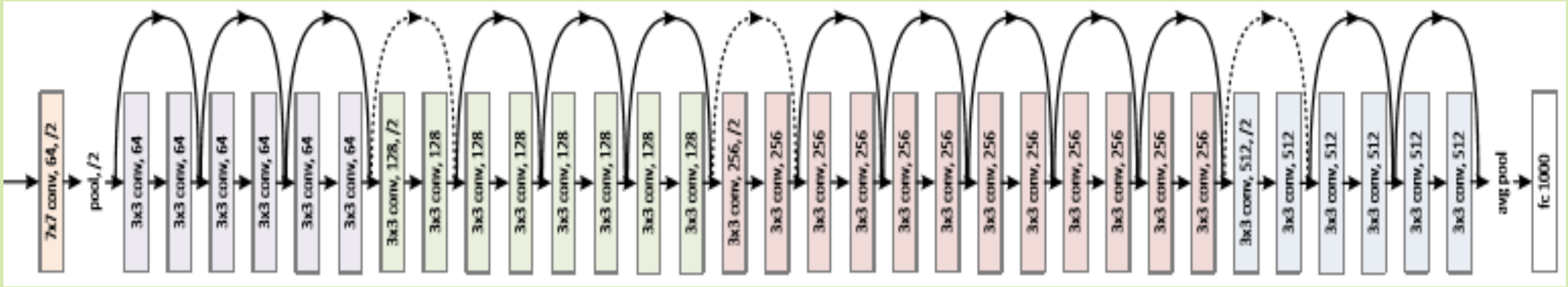


Figure 1. **Left**: A block of ResNet [14]. **Right**: A block of ResNeXt with cardinality = 32, with roughly the same complexity. A layer is shown as (# in channels, filter size, # out channels).



Back

Project 2 Figure/Graph Extraction | Various U-NeXt Approaches

- ***Model used*** U-NeXt
- ***Datasets used*** 481 images from external dataset European historical newspapers (ENP) [2] and 1,532 images from the Beyond Words collection

1 ***Approach 1 Configuration***
(1) using deconvolutional layer for upscaling
(2) training and testing on the BW directly

2 ***Approach 2 Configuration***
(1) using deconvolutional layer for upscaling
(2) pre-training on the ENP dataset
(3) fine-tuning and testing on the BW

3 ***Approach 3 Configuration***
(1) using resizing layer for upscaling
(2) training and testing on the BW directly

4 ***Approach 4 Configuration***
(1) using deconvolutional layer for upscaling
(2) combining all five non-background classes into one class
(3) training and testing on the BW directly

Performance of Four Various U-NeXt Approaches

	Without Pre-trained ENP Classifier 1		Using Pre-trained ENP Classifier 2		Using Resizing Layers 3		Combined Two-class Segmentation 4	
	Train	Test	Train	Test	Train	Test	Train	Test
Accuracy	89.08%	80.11%	89.42%	85.53%	88.90%	86.69%	91.76%	88.89%
mIoU	50.43%	38.00%	41.21%	38.57%	51.31%	37.84%	71.44%	64.97%