University of Nebraska-Lincoln
Progress report

Proposal# 129426
Jul 30th, 2019

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.

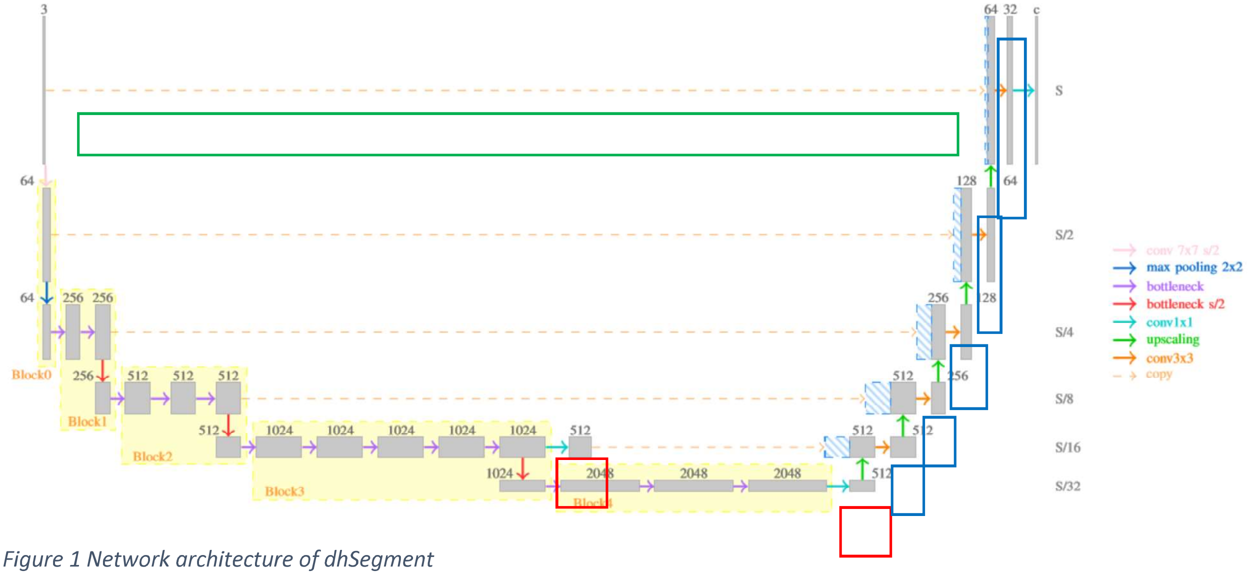


Figure 1 Network architecture of dhSegment

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.

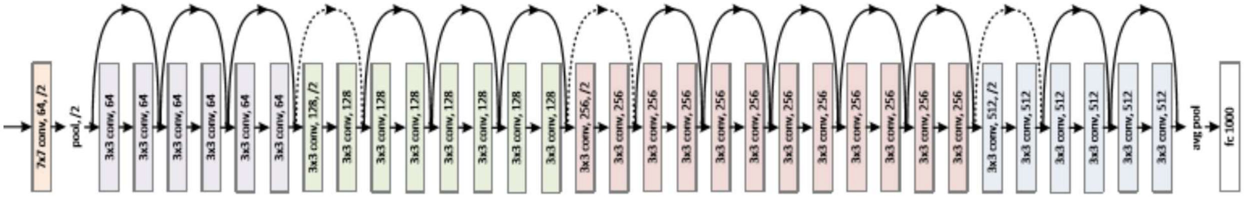


Figure 2 Network architecture of ResNet

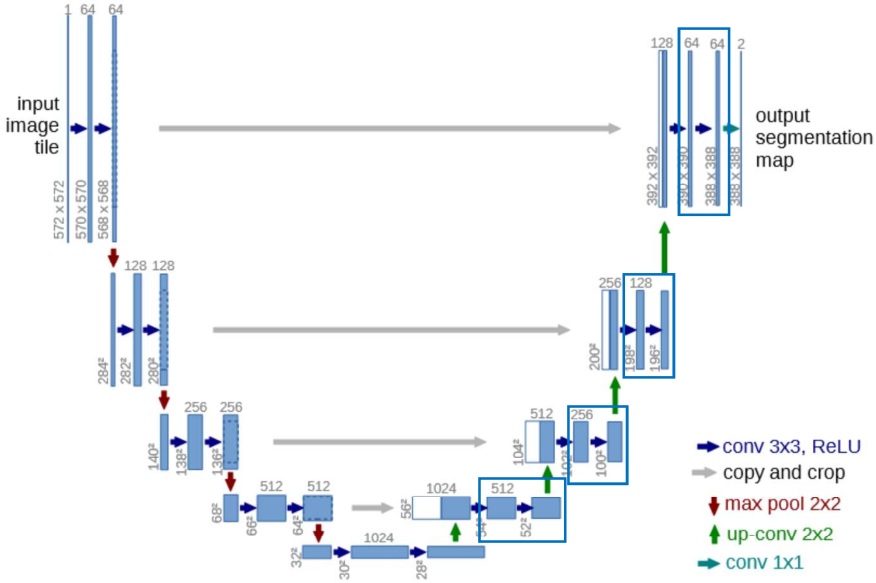


Figure 3 Network architecture of U-Net

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.

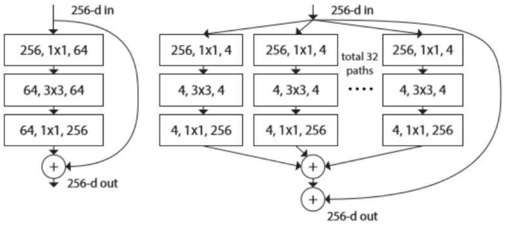


*Figure 4 **(Left)** Residual blocks of ResNet. **(Right)** Residual blocks of ResNext.*

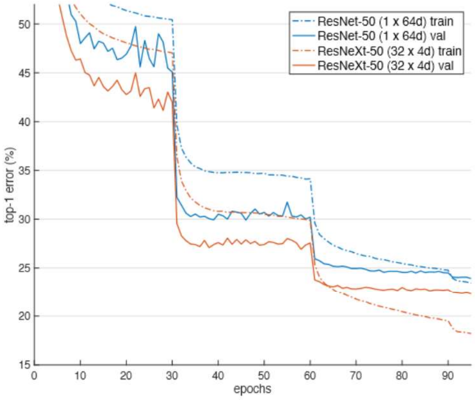


Figure 5 Comparative results between ResNet and ResNeXt on ImageNet-1K dataset.

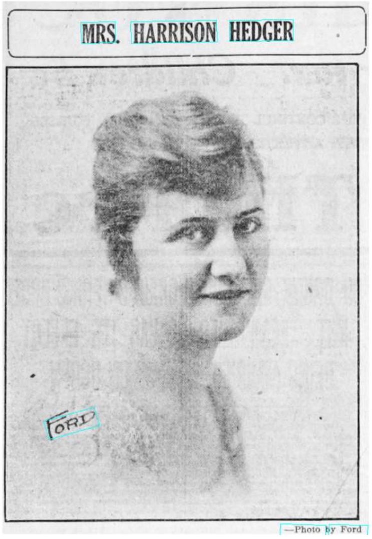


Figure 6 EAST text detection on Beyond Words snippet. Blue rectangles indicate detected text regions.

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).

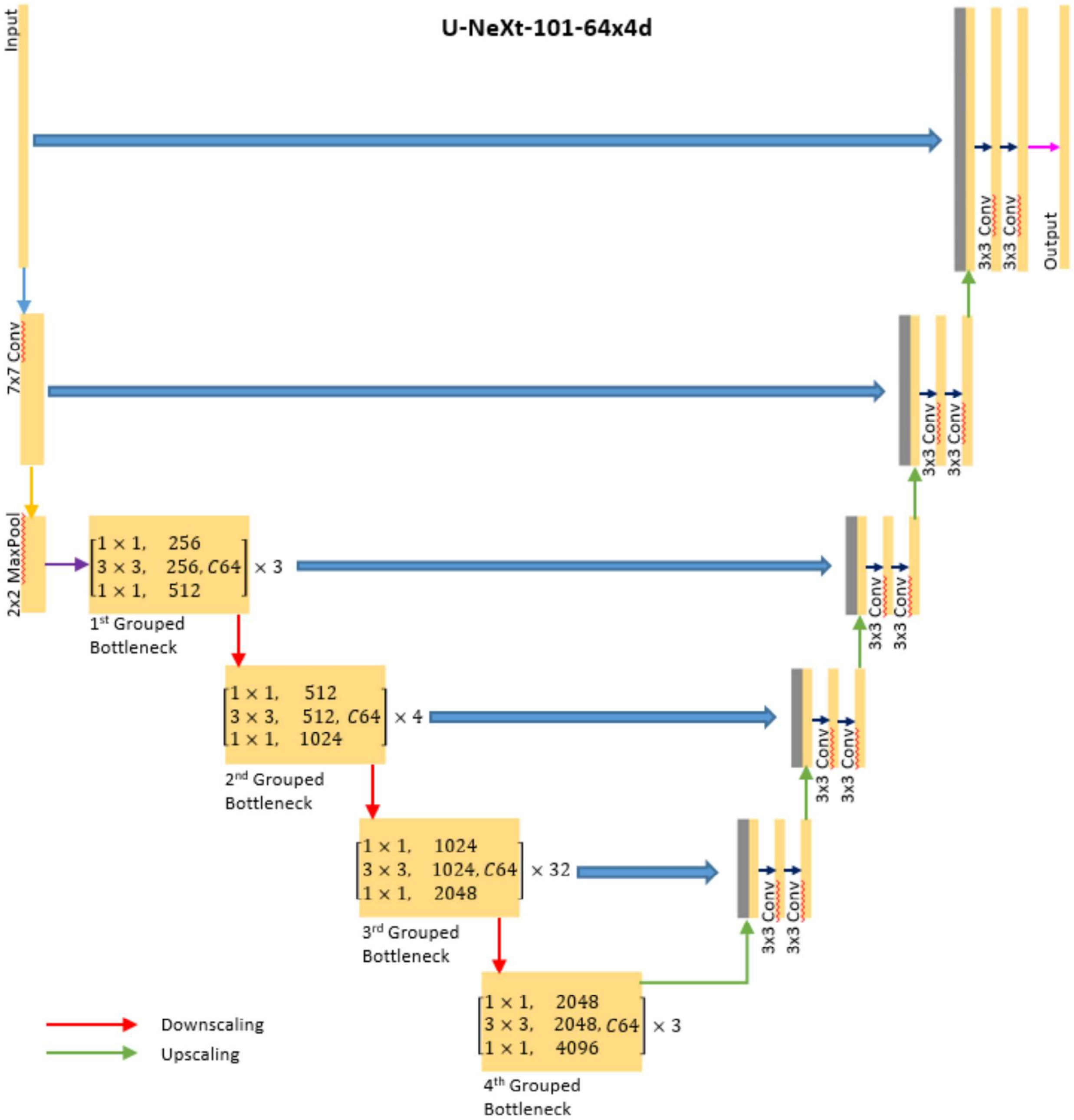
[4] Oliveira, S. A., Seguin, B., & Kaplan, F. (2018, August). dhSegment: A generic deep-learning approach for document segmentation. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)* (pp. 7-12). IEEE.

[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
3x3 Conv
3x3 Conv
3x3 Conv
3x3 Conv
3x3 Conv
3x3 Conv
3x3 Conv
3x3 Conv
3x3 Conv
3x3 Conv
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

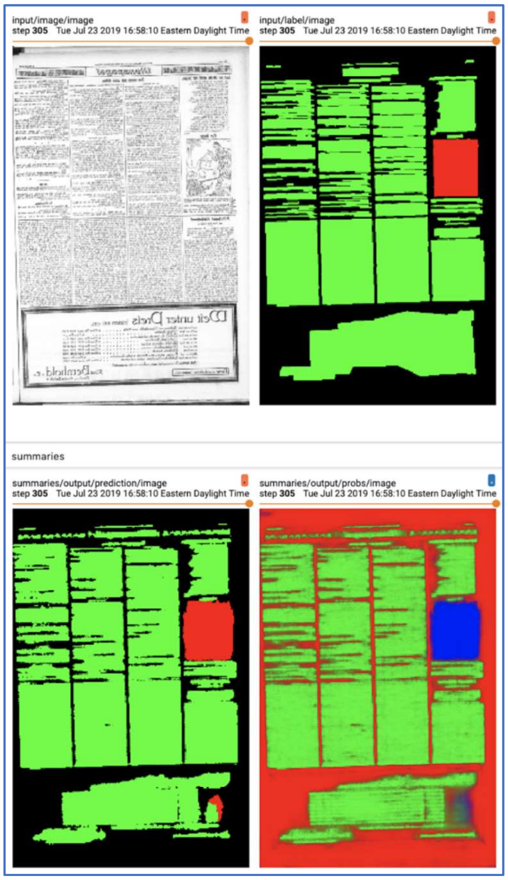

input/image/image
step 305 Tue Jul 23 2019 16:58:10 Eastern Daylight Time
input/label/image
step 305 Tue Jul 23 2019 16:58:10 Eastern Daylight Time
summaries
summaries/output/prediction/image
step 305 Tue Jul 23 2019 16:58:10 Eastern Daylight Time
summaries/output/probs/image
step 305 Tue Jul 23 2019 16:58:10 Eastern Daylight Time

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.

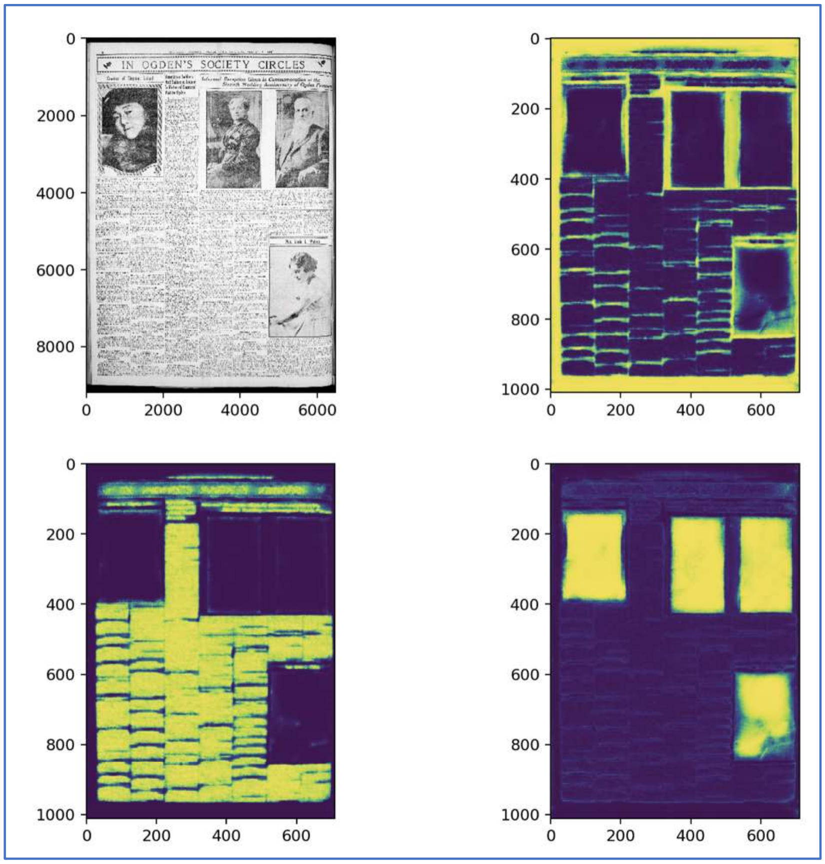


Figure 2. Visual inspection on the segmentation result of model trained on ENP dataset. Note the image shown here is from the Chronicling America corpus, which is never shown to the model during the training. Clockwise from top-left: (1) Input, (2) background-map, (3) image-map, and (4) text-map. In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

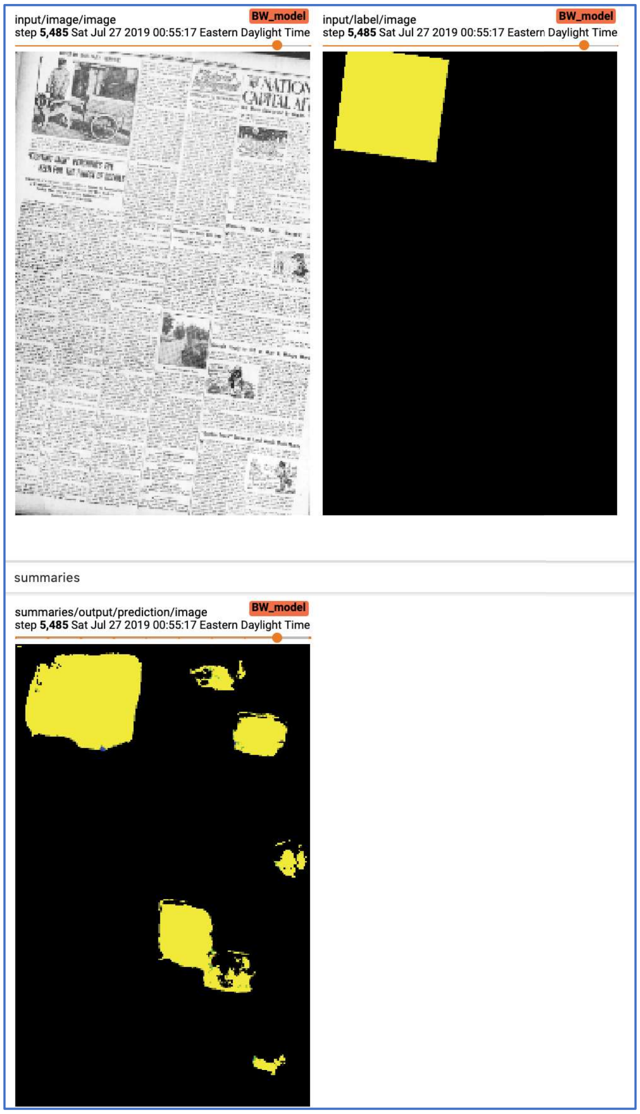
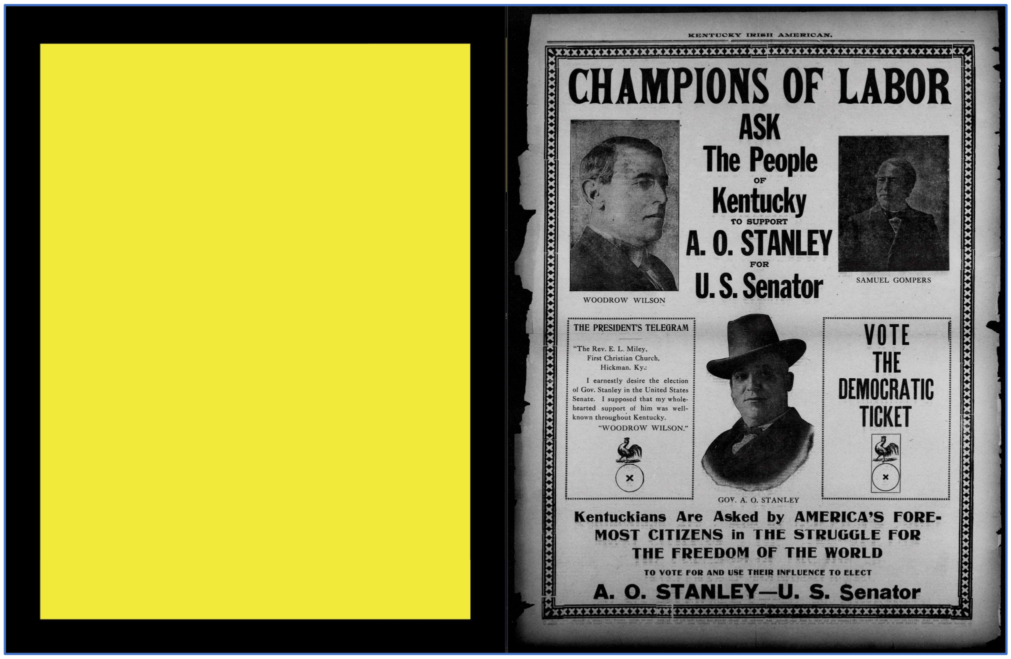

input/image/image
BW_model
step 5,485 Sat Jul 27 2019 00:55:17 Eastern Daylight Time
input/label/image
BW_model
step 5,485 Sat Jul 27 2019 00:55:17 Eastern Daylight Time
summaries
summaries/output/prediction/image
BW_model
step 5,485 Sat Jul 27 2019 00:55:17 Eastern Daylight Time

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.



KENTUCKY IRISH AMERICAN.

CHAMPIONS OF LABOR

ASK

The People

OF

Kentucky

TO SUPPORT

A. O. STANLEY

FOR

U. S. Senator

WOODROW WILSON

SAMUEL GOMPERS

THE PRESIDENT'S TELEGRAM

"The Rev. E. L. Miley,
First Christian Church,
Hickman, Ky.:

I earnestly desire the election of Gov. Stanley in the United States Senate. I supposed that my whole-hearted support of him was well-known throughout Kentucky.

"WOODROW WILSON."

GOV. A. O. STANLEY

VOTE THE DEMOCRATIC TICKET

Kentuckians Are Asked by AMERICA'S FOREMOST CITIZENS in THE STRUGGLE FOR THE FREEDOM OF THE WORLD

TO VOTE FOR AND USE THEIR INFLUENCE TO ELECT

A. O. STANLEY—U. S. Senator

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.

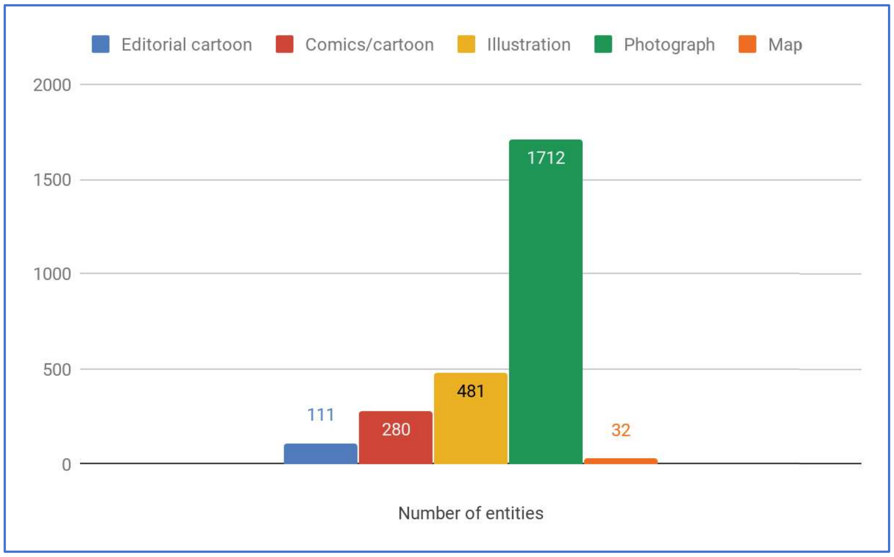


Figure 5. Number of entities in the Beyond Words JSON file. Note here that the dataset is overwhelmed with photograph class (4%, 11%, 18%, 65%, and 1%).

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

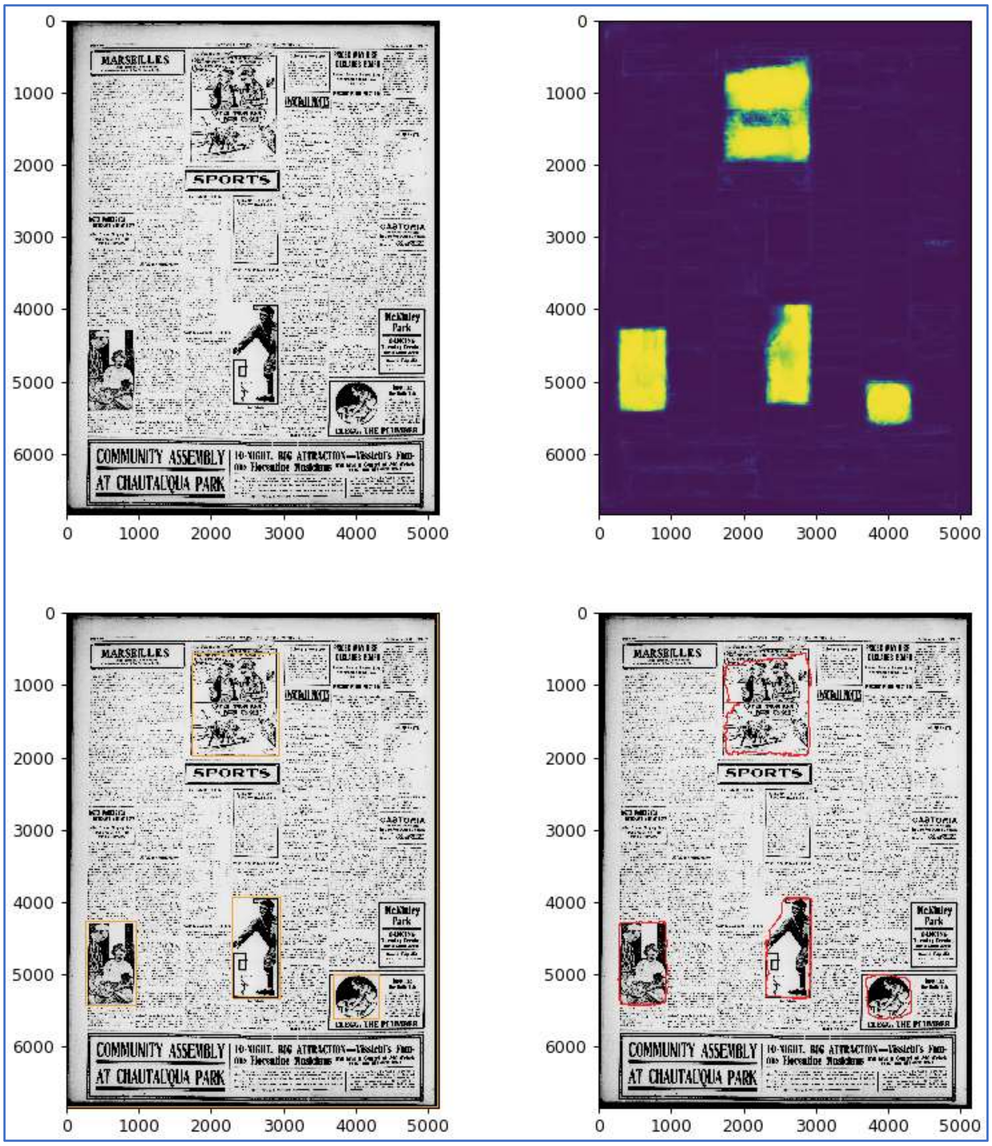


Figure 1. Segmentation result of ENP_500_v4 on Chronicling America image (sn92053240-19190805.jpg). Clockwise from top-left: (1) Input, (2) probability map for figure class, (3) detected figures in polygon, and (4) detected figures in bounding-box. In the probability map, pixels with higher probability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

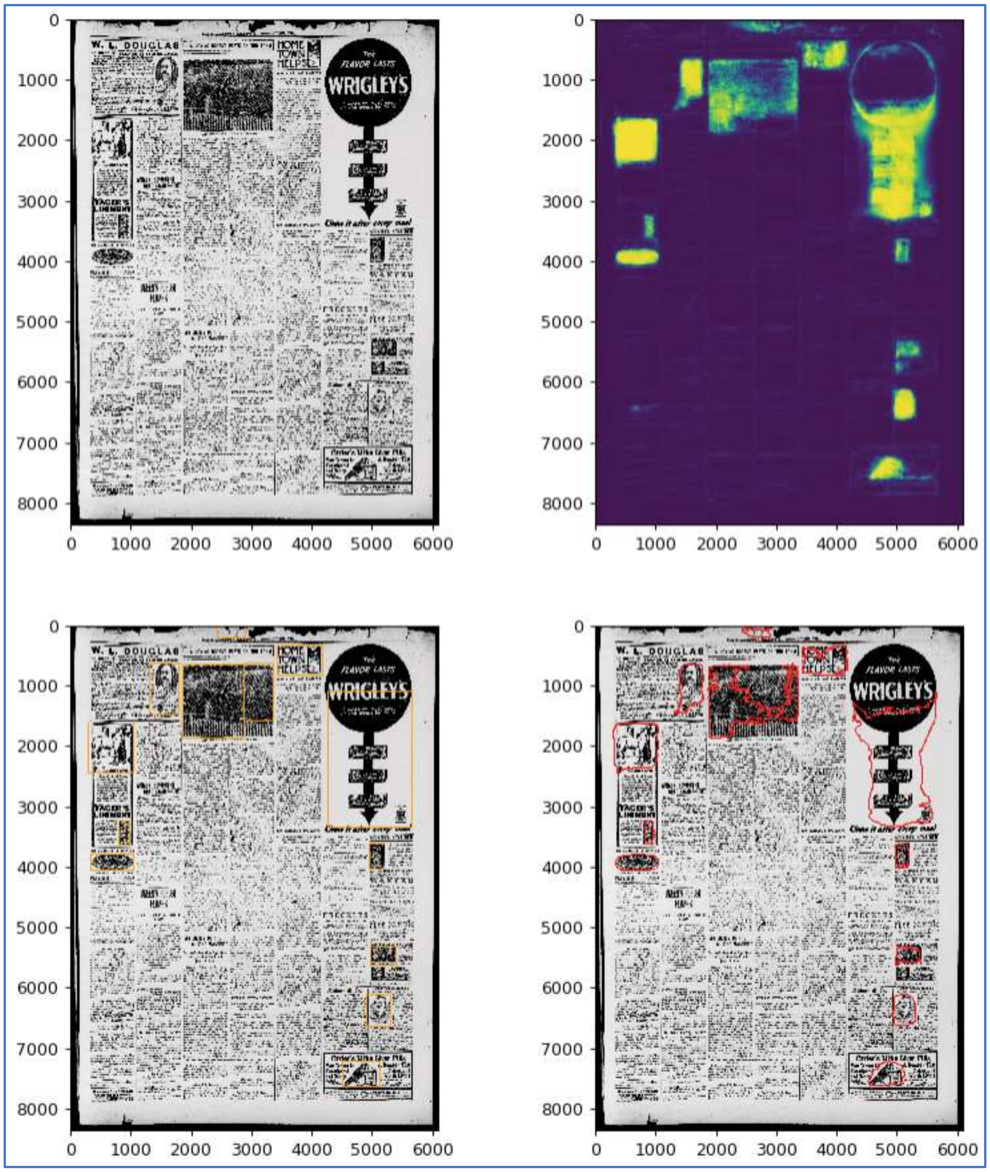


Figure 2. Segmentation result of ENP_500_v4 on Chronicling America image (ndnp-jpeg-surrogates_deu_descendo_ver01_data_sn84026820_00271765095_1917050501_0153.jpg). Clockwise from top-left: (1) Input, (2) probability map for figure class, (3) detected figures in polygon, and (4) detected figures in bounding-box.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

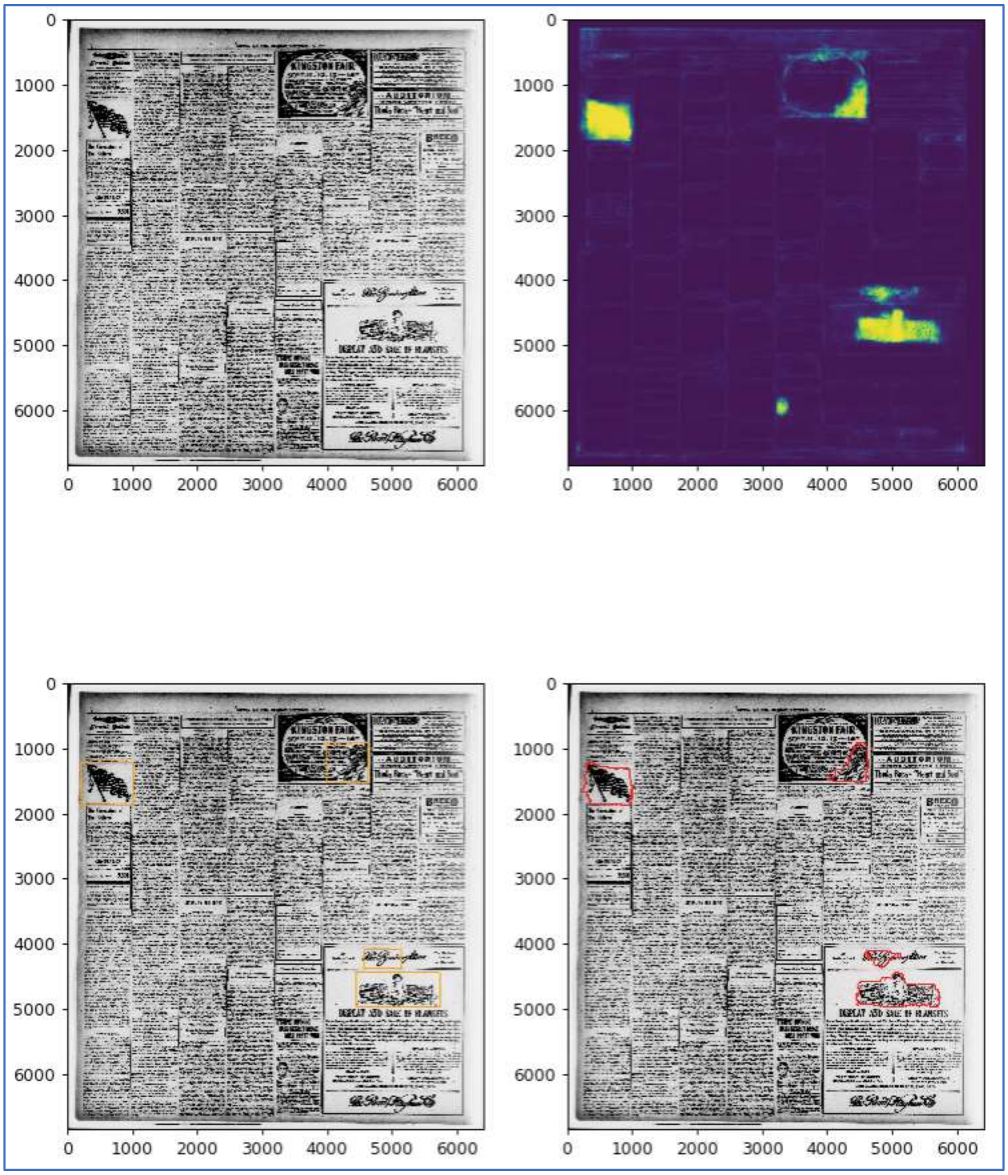


Figure 3. Segmentation result of ENP_500_v4 on Chronicling America image (ndnp-jpeg-surrogates_ct_berlin_ver01_data_sn82014086_00295866135_1917091301_0116.jpg). Clockwise from top-left: (1) Input, (2) probability map for figure class, (3) detected figures in polygon, and (4) detected figures in bounding-box.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

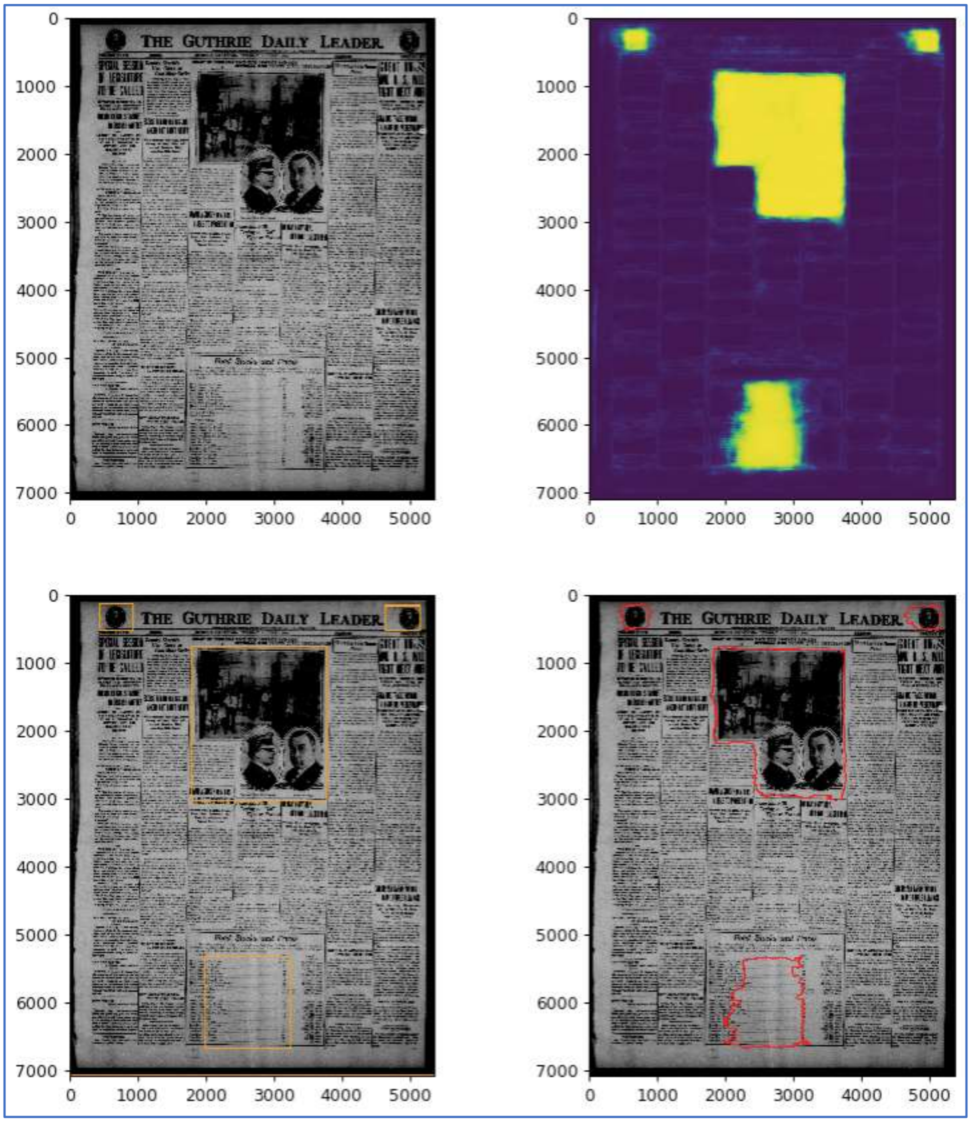
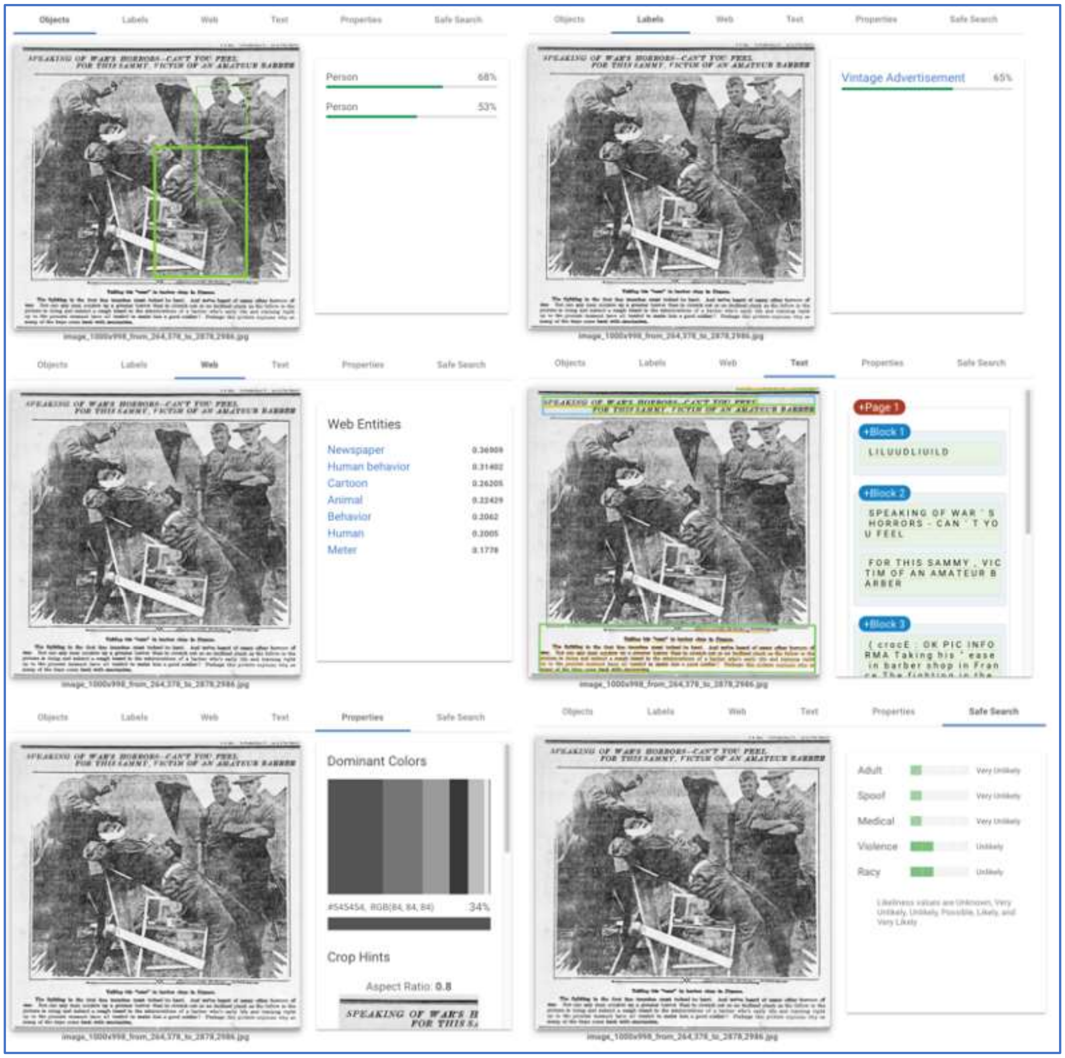


Figure 4. Segmentation result of ENP_500_v4 on Chronicling America image (sn86063952-19190805.jpg). Clockwise from top-left: (1) Input, (2) probability map for figure class, (3) detected figures in polygon, and (4) detected figures in bounding-box. In the probability map, pixels with higher probability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


Objects
Labels
Web
Text
Properties
Safe Search
SPEAKING OF WAR'S HORRORS—CAN'T YOU FEEL FOR THIS SAMMY, VICTIM OF AN AMATEUR BARBER
image_1000x998_from_264,378_to_2878,2986.jpg

Person 68%

Person 53%

Vintage Advertisement 65%

Web Entities

Newspaper	0.36909
Human behavior	0.31402
Cartoon	0.26205
Animal	0.22429
Behavior	0.2062
Human	0.2005
Meter	0.1778

+Page 1

+Block 1

LILUUDLIUILD

+Block 2

SPEAKING OF WAR ' S HORRORS - CAN ' T YO U FEEL

FOR THIS SAMMY , VIC TIM OF AN AMATEUR B ARBER

+Block 3

(crocE : OK PIC INFO RMA Taking his ' ease in barber shop in Fran ce The fighting in the

Dominant Colors

#545454, RGB(84, 84, 84) 34%

Crop Hints

Aspect Ratio: 0.8

Adult	Very Unlikely
Spoof	Very Unlikely
Medical	Very Unlikely
Violence	Unlikely
Racy	Unlikely

Likeliness values are Unknown, Very Unlikely, Unlikely, Possible, Likely, and Very Likely

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.

University of Nebraska-Lincoln
Progress report

Proposal# 129426
Aug 9th, 2019

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.

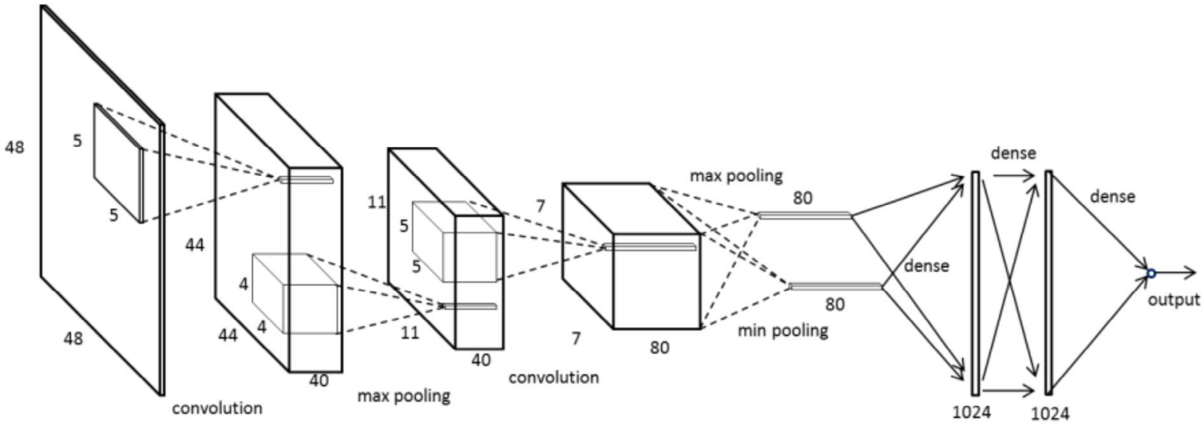


Figure 1 CNN model in [Kang et al. 2014] for OCR accuracy prediction.

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.

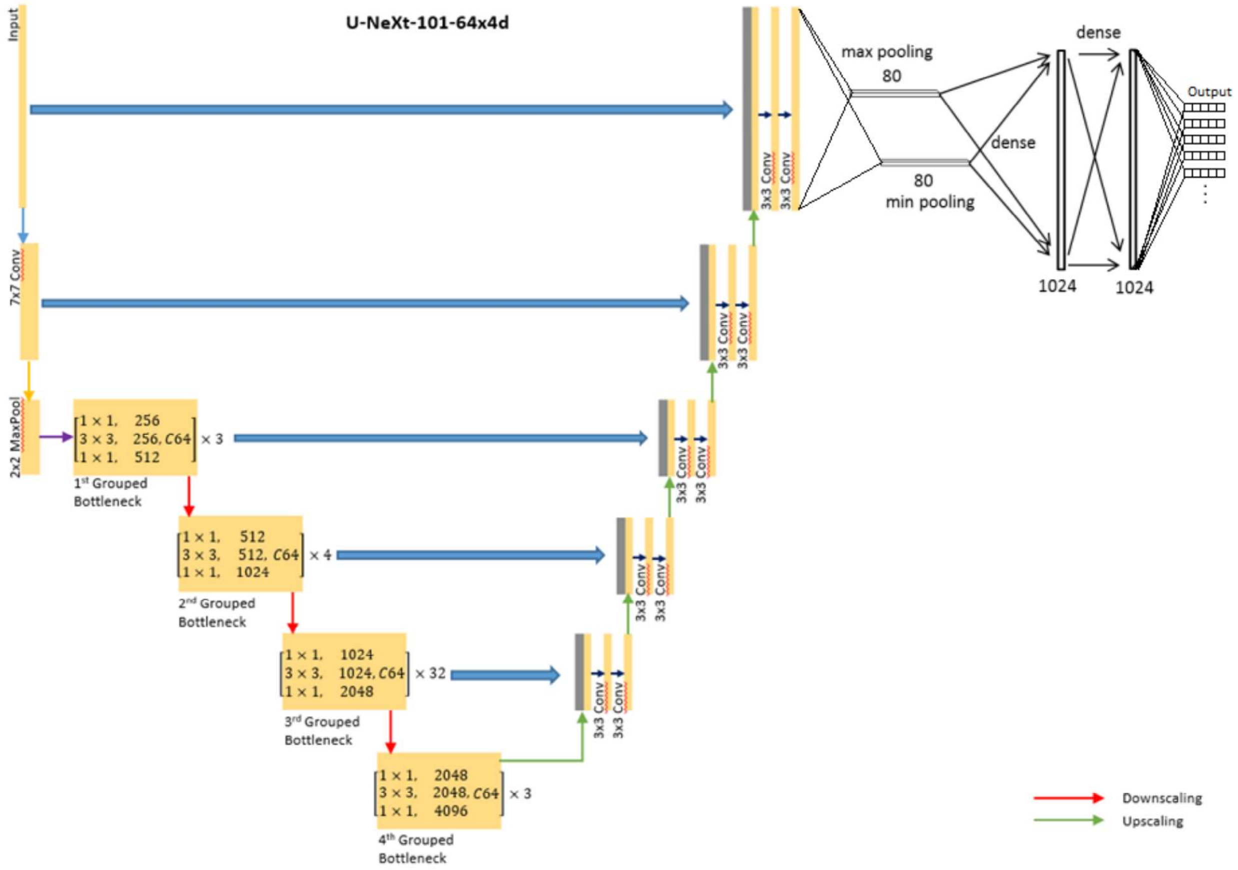


Figure 2 Inference multi-output U-NeXt model for subject DIQA.

Appendix

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

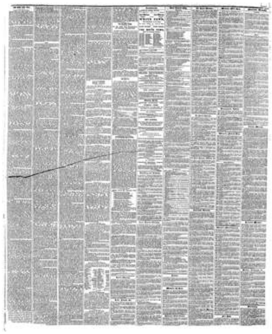
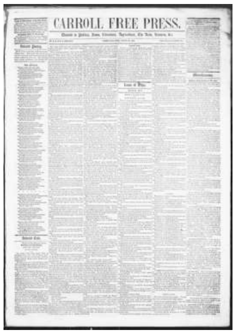
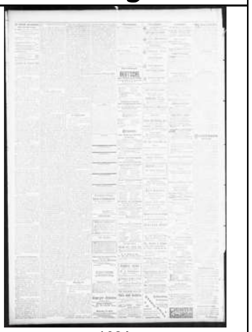
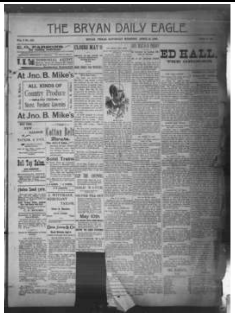
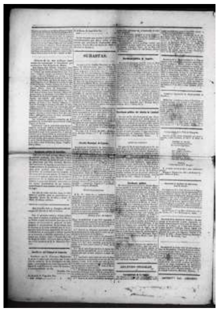
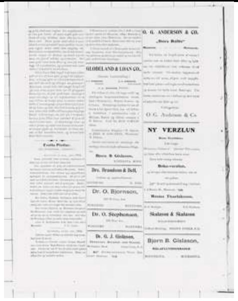
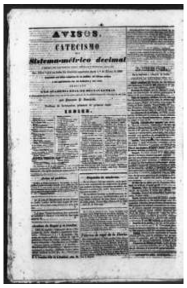
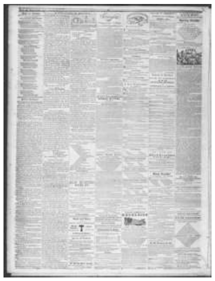
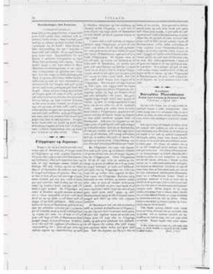
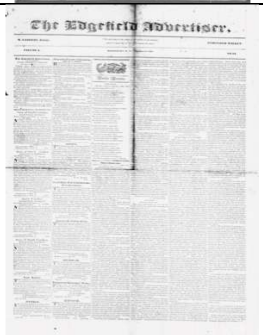
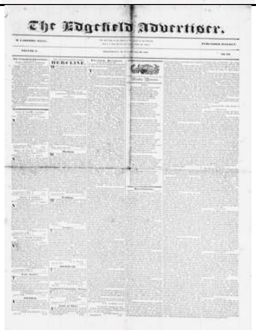
	High/Severe		Some		Low/None	
	Value	Image	Value	Image	Value	Image
Contrast	146.08	 1834-1922FullPages_20PerYr/1868_English/sn82014064/1868-07-18/ed-1/seq-4.jp2	55.2	 1834-1922FullPages_20PerYr/1848_English/sn83035366/1848-03-16/ed-1/seq-1.jp2	3.11	 1834-1922FullPages_20PerYr/1898_German/sn83045081/1898-12-29/ed-1/seq-2.jp2
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.

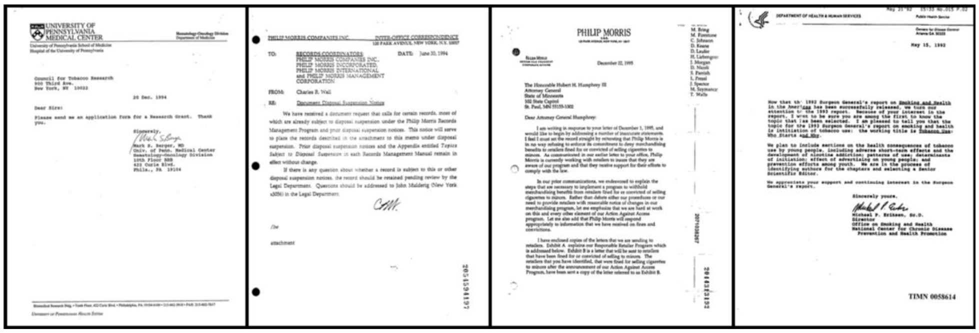


Figure 1. Examples of document images that show a wide range of visual variability within the same type (i.e., a *letter* type in this particular example). Note that no two documents show the exact same spatial arrangement of header, date, address, body, and signature; some of the documents even omit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

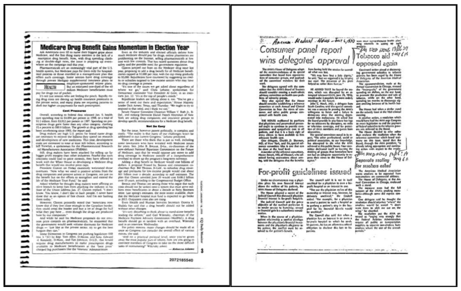


Figure 2. Examples of document images that show visual similarities across different types (Left: *form*, right: *scientific publication*) Note that two different types of documents share similar spatial arrangement of title and body. Even the amount of contents, so-called density, is also similar.

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

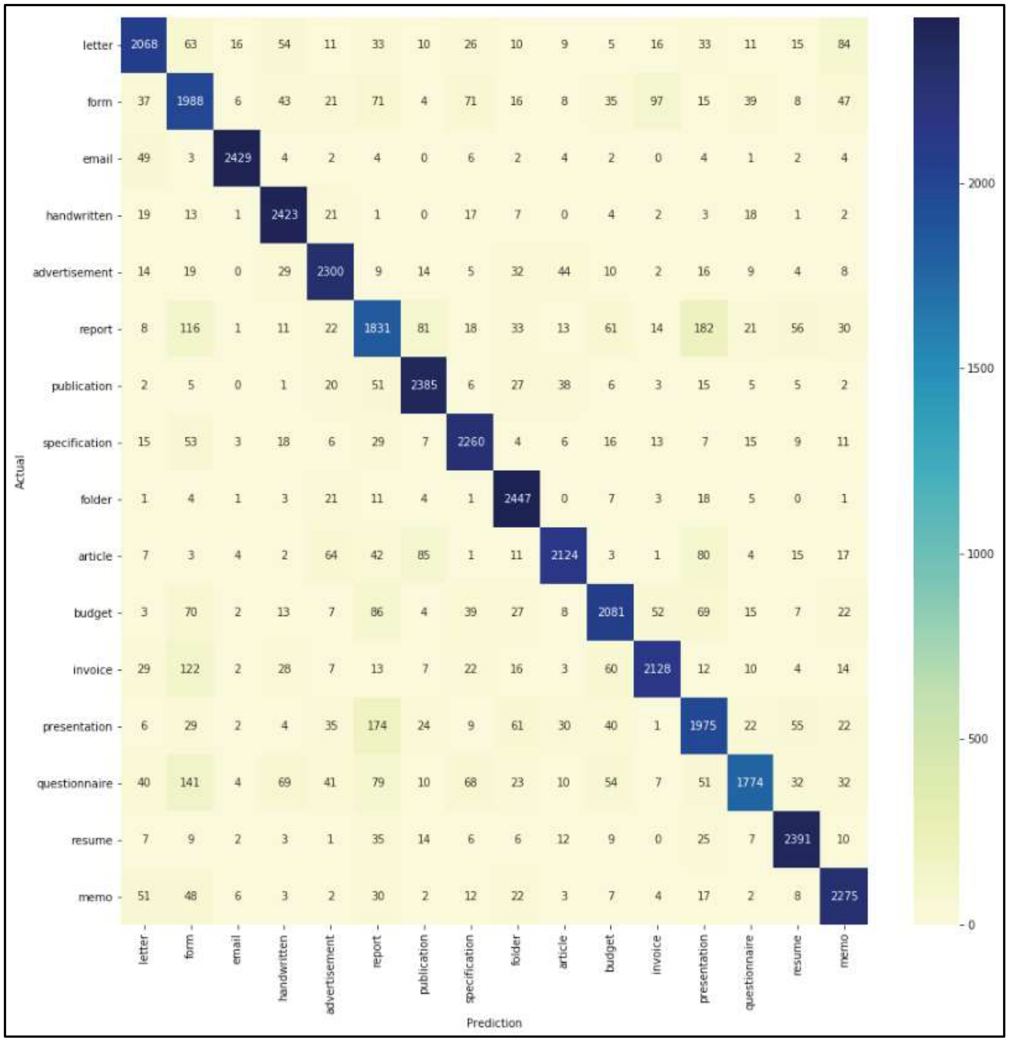
As a preprocessing step, in order to make the shape of our input to match with that of *VGG-16*, we convert grayscale images to three-channel images by simply copying the pixel values of the single-channel to three channels. Also, each image is resized to 224 by 224, and normalized to range from 0 to 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



	letter	form	email	handwritten	advertisement	report	publication	specification	folder	article	budget	invoice	presentation	questionnaire	resume	memo
letter	2068	63	16	54	11	33	10	26	10	9	5	16	33	11	15	84
form	37	1988	6	43	21	71	4	71	16	8	35	97	15	39	8	47
email	49	3	2429	4	2	4	0	6	2	4	2	0	4	1	2	4
handwritten	19	13	1	2423	21	1	0	17	7	0	4	2	3	18	1	2
advertisement	14	19	0	29	2300	9	14	5	32	44	10	2	16	9	4	8
report	8	116	1	11	22	1831	81	18	33	13	61	14	182	21	56	30
publication	2	5	0	1	20	51	2385	6	27	38	6	3	15	5	5	2
specification	15	53	3	18	6	29	7	2260	4	6	16	13	7	15	9	11
folder	1	4	1	3	21	11	4	1	2447	0	7	3	18	5	0	1
article	7	3	4	2	64	42	85	1	11	2124	3	1	80	4	15	17
budget	3	70	2	13	7	86	4	39	27	8	2081	52	69	15	7	22
invoice	29	122	2	28	7	13	7	22	16	3	60	2128	12	10	4	14
presentation	6	29	2	4	35	174	24	9	61	30	40	1	1975	22	55	22
questionnaire	40	141	4	69	41	79	10	68	23	10	54	7	51	1774	32	32
resume	7	9	2	3	1	35	14	6	6	12	9	0	25	7	2391	10
memo	51	48	6	3	2	30	2	12	22	3	7	4	17	2	8	2275

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.

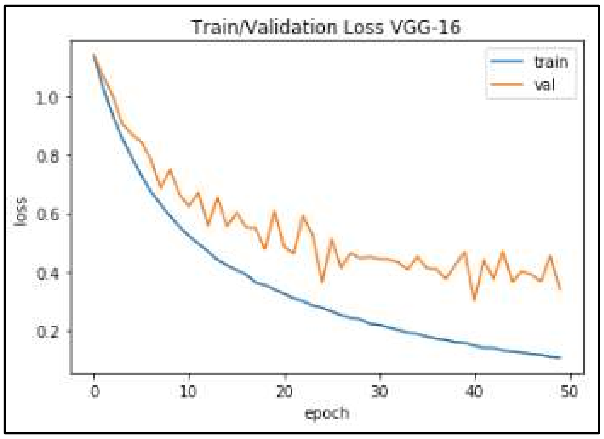


Figure 2. Training and validation loss of *VGG-16* with *suffrage_1002* training and validation set. In spite of some fluctuations, the overall trend of validation loss goes down.

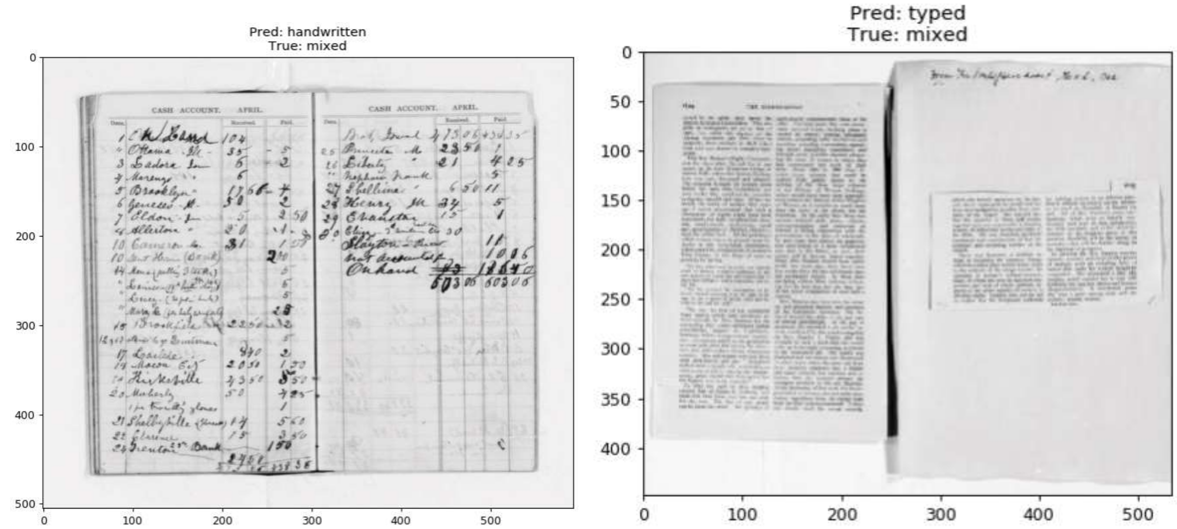


Figure 3. Failure prediction cases. On the left example, a typed region is relatively smaller than that of handwriting. On the right example, a handwriting region is relatively smaller than that of typing.

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

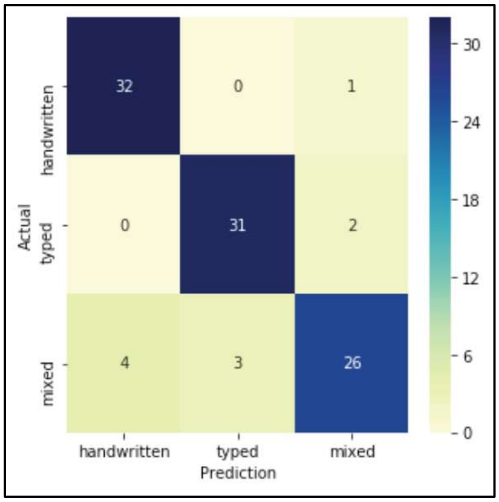


Figure 4. Heatmap of confusion matrices for classification performance of *VGG-16* trained on *suffrage_1002*. Note that diagonal elements contain most of datapoints, which indicates that most of our model's predictions are correct over all three classes.

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.

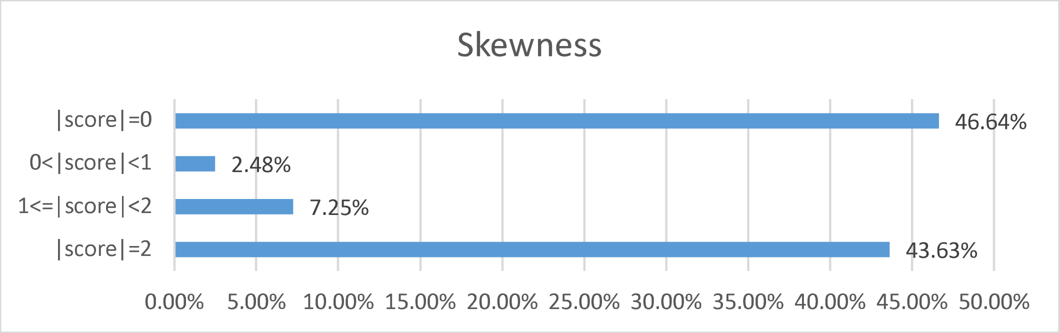


Figure 1 Skewness analysis

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.

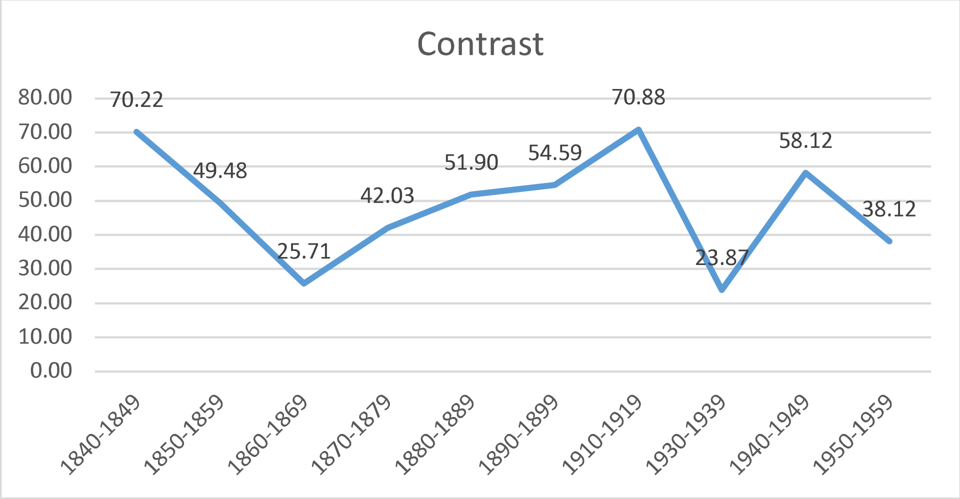


Figure 2 Contrast Score Analysis

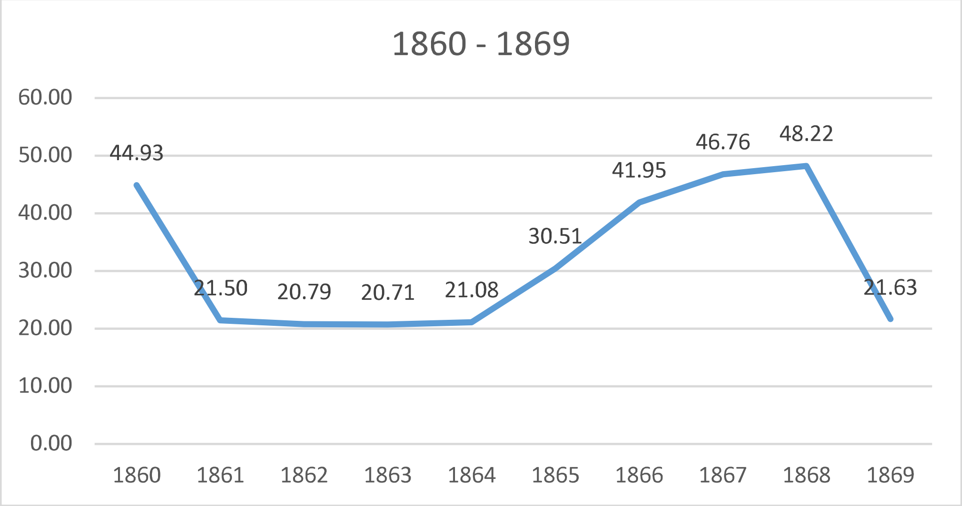
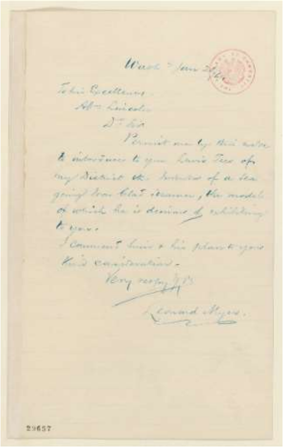


Figure 3 Contrast Score Analysis from 1860 to 1869

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.



Wash. 7 Jan [illegible]

To his Excellency
Abm Lincoln
Dr Sir
Permit me by this note
to introduce to you Lewis Tees of
my District the Inventor of a sea
going iron clad steamer, the model
of which he is desirous of exhibiting
to you.
I commend him & his plans to your
kind consideration.
Very respy Yrs
Leonard Myers.

29657

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.

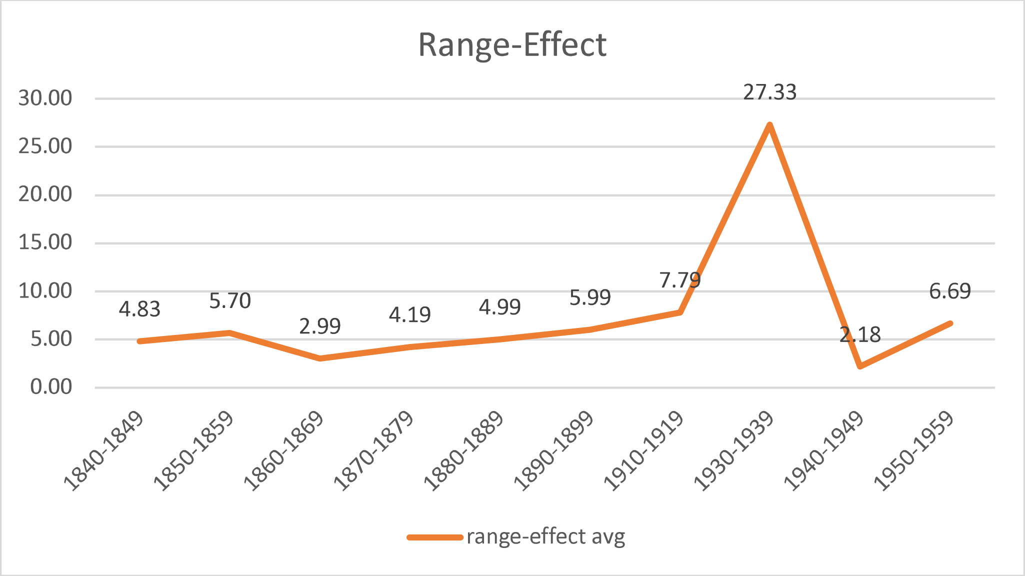


Figure 5 Range-effect Score Analysis

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.

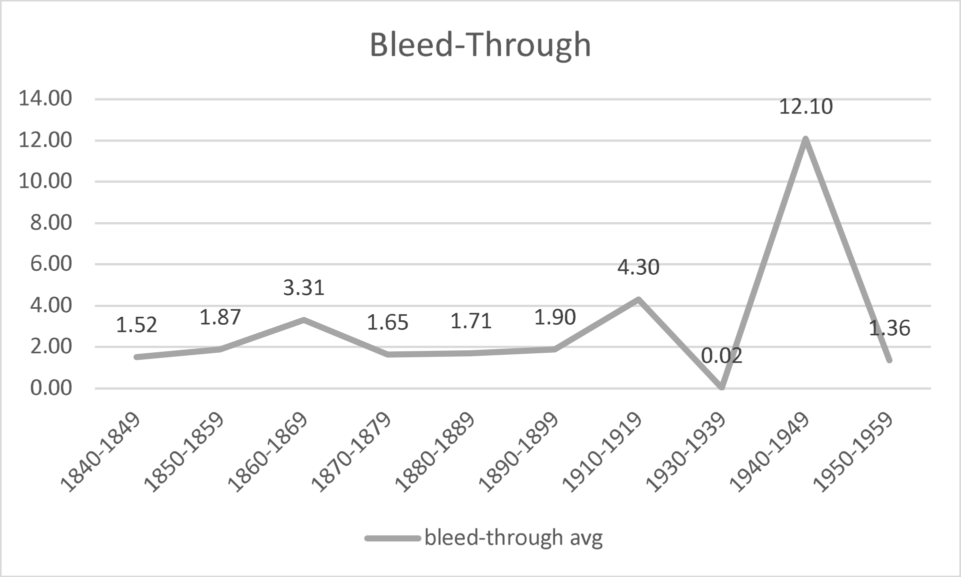


Figure 6 Bleed-through Score Analysis

Differentiation between Microfilm and Scanned images

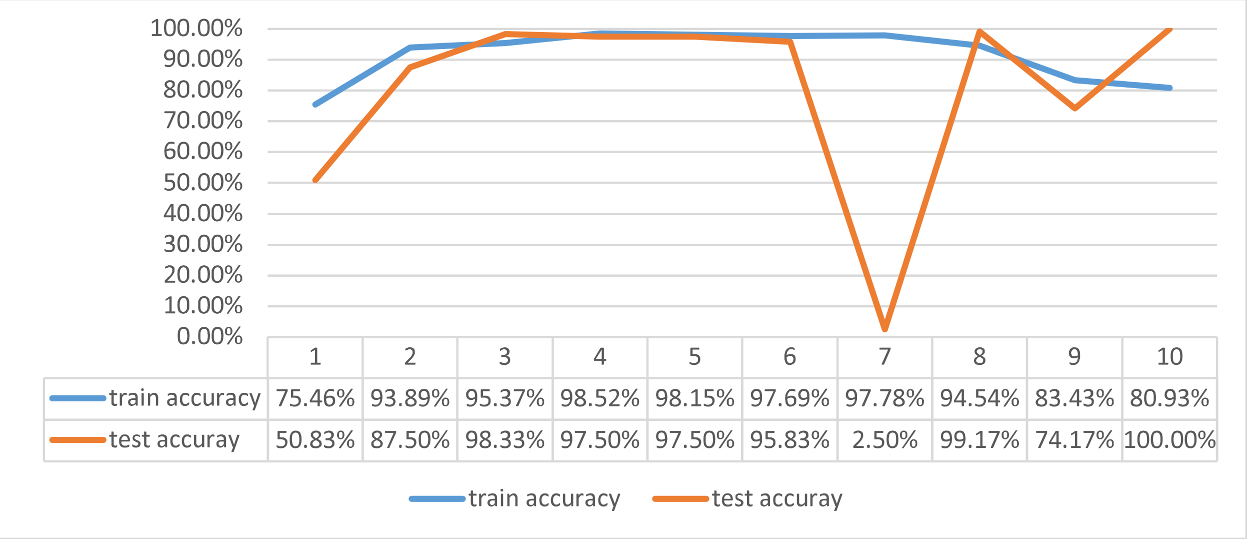
The types of digitization that generated the document images are mixed. There are both microfilms and scanned images in the collection. As a result, techniques developed for one type might not work for the other. In our DIQA suite of image processing tools, for example, we assume that the document images 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.

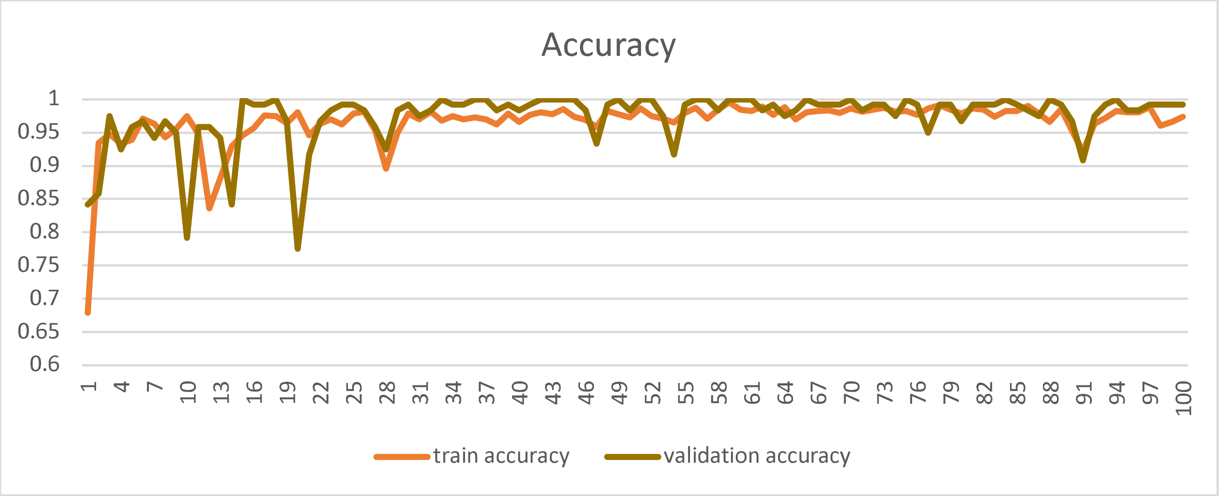


Figure 1 Training and validation accuracies

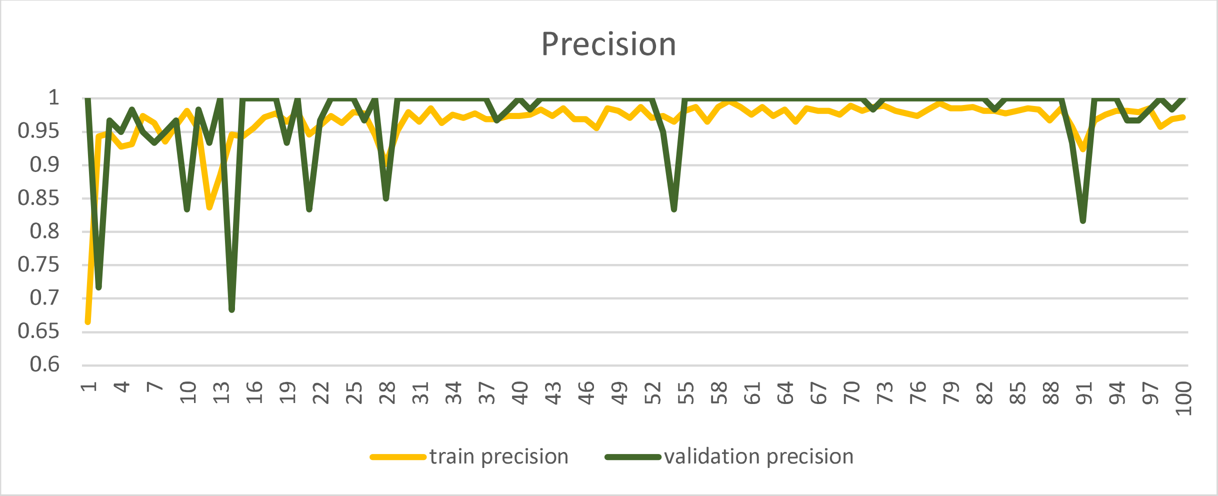


Figure 2 Traing and validation precisions

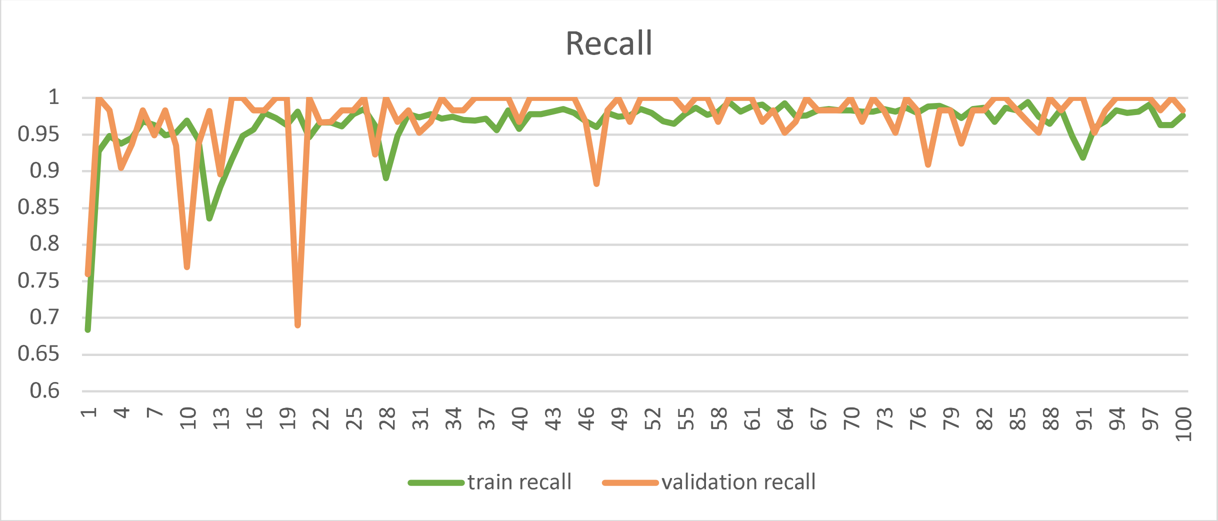


Figure 3 Training and validation recall

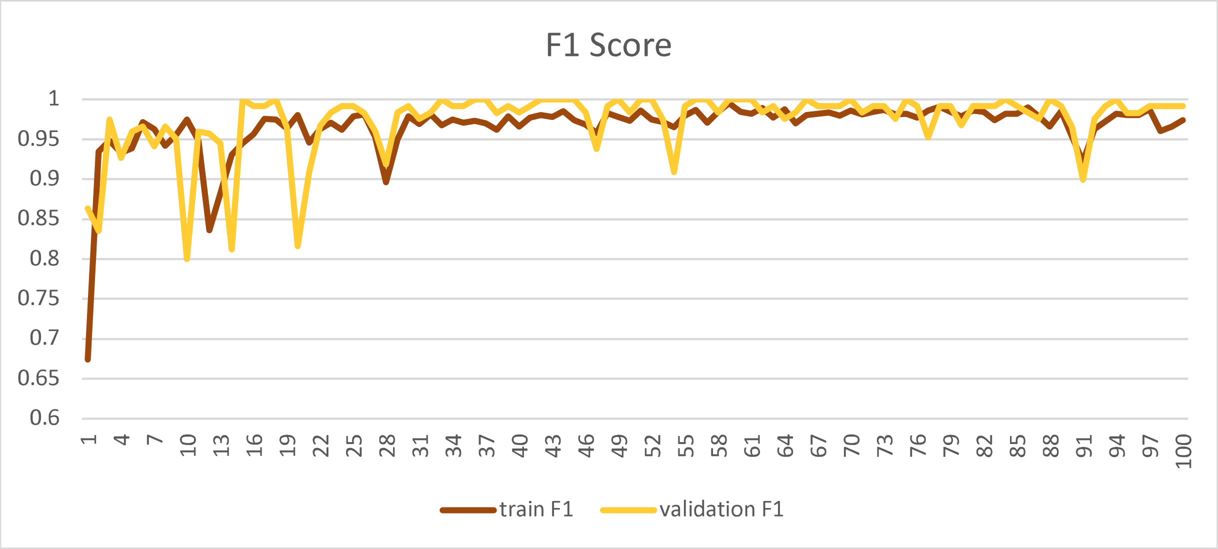


Figure 4 Training and validation F1 score

Table 1 Prediction Results

Total	Predicted Microfilmed Documentation	Predicted Scanned Documentation
36103	2834	33269

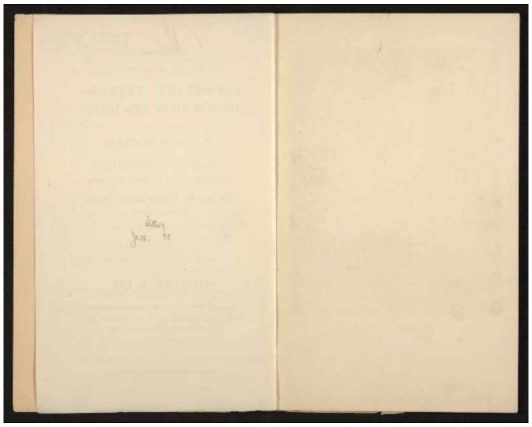
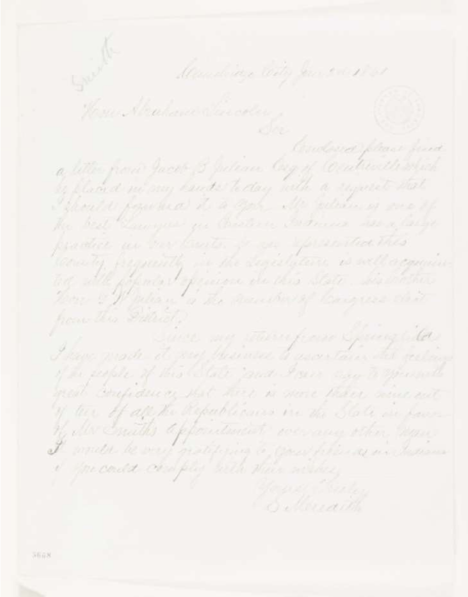


Figure 5 Type (1) mis-classification: "blank" document image



Smith

Cambridge City Jan 2d 1861

Hon. Abraham Lincoln
Sir

Enclosed please find a letter from Jacob B. Julian Esq of Centreville which he placed in my hands today with a request that I should forward it to you. Mr Julian is one of the best Lawyers in Eastern Indiana has a large practice in our Courts & has represented this County frequently in the Legislature is well acquainted with popular opinion in this State. his brother Hon. G. W. Julian is the member of Congress elect from this District.

Since my return from Springfield I have made it my business to ascertain the feelings of the people of this State and I can say to you with great confidence that there is more than nine out of ten of all the Republicans in the State in favor of Mr Smiths appointment over any other man. It would be very gratifying to your friends in Indiana if you could comply with their wishes.

Yours Truly
S. Meredith

3658

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

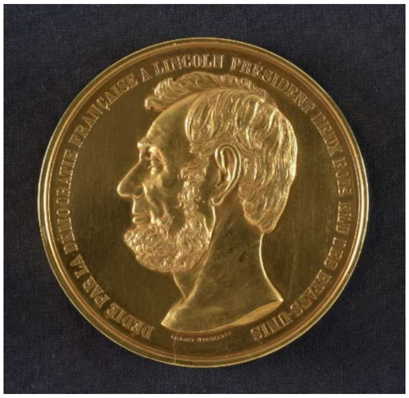

DÉDIÉE PAR LA DÉMOCRATIE FRANÇAISE A LINCOLN PRÉSIDENT DEUX FOIS ÉLU DES ÉTATS-UNIS

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

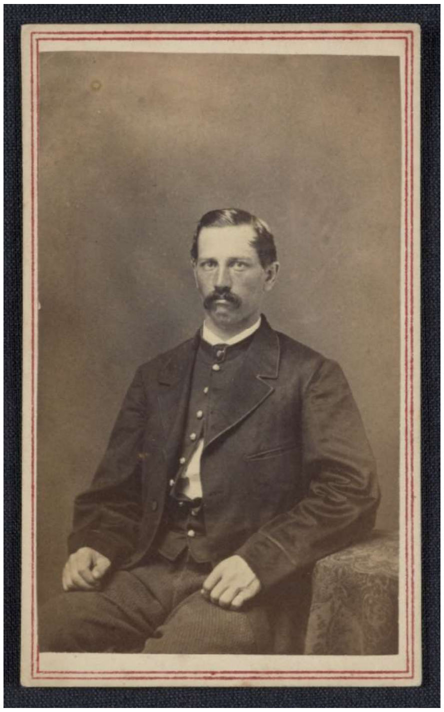


Figure 8 Type (4) mis-classification: graphical images (photo)

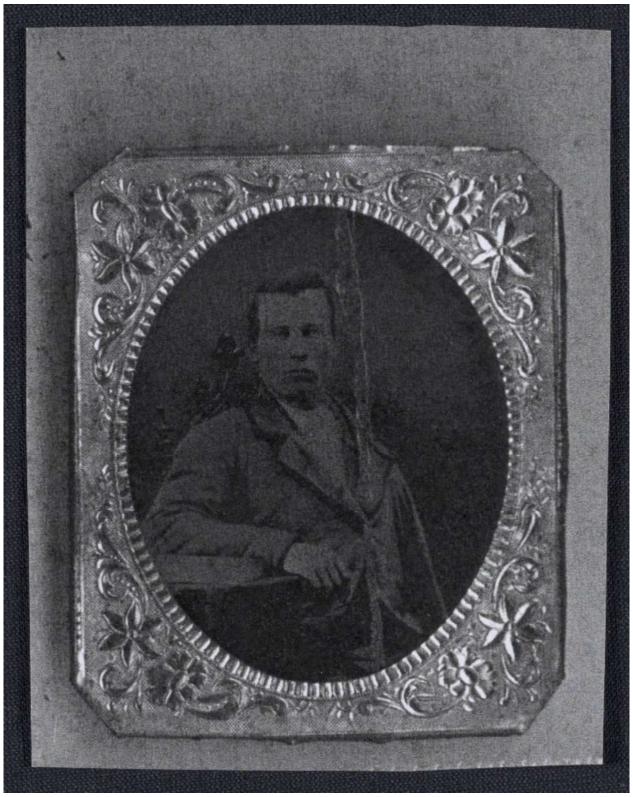


Figure 9 Microfilmed frame-photo being correctly classified

University of Nebraska-Lincoln
Progress report

Proposal# 129426
Sept 20th, 2019

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.

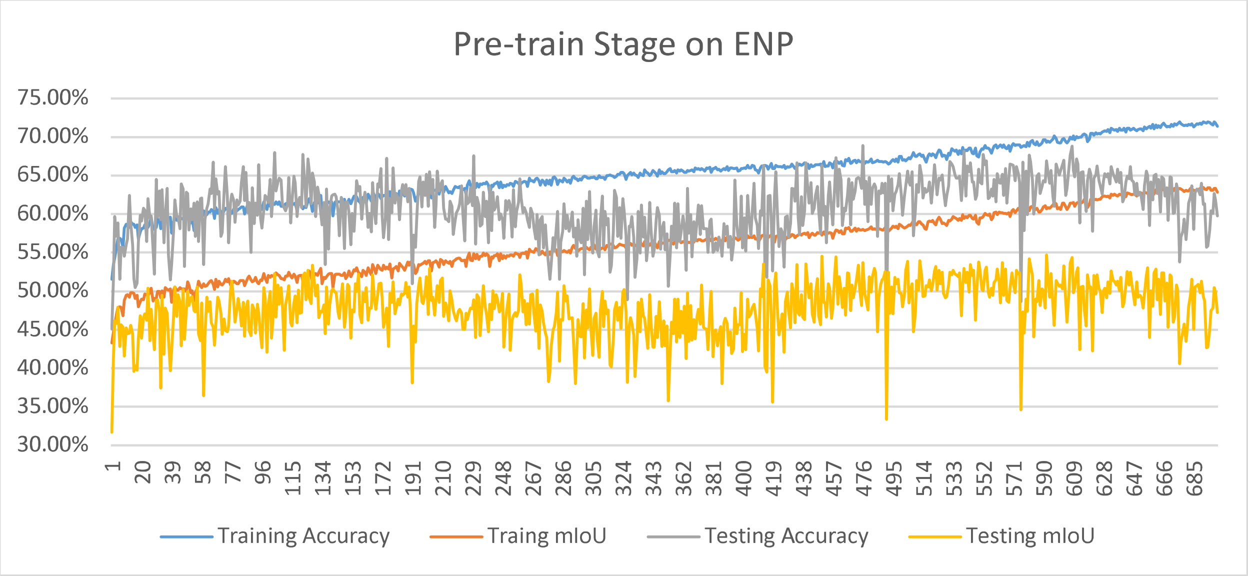
Shown in Figure 1, the training performance of pre-training stage on the ENP reached 72% on training accuracy and 63% on training mIoU. And the testing performance of the pre-training stage reached 68% on testing accuracy and 53% on testing mIoU.

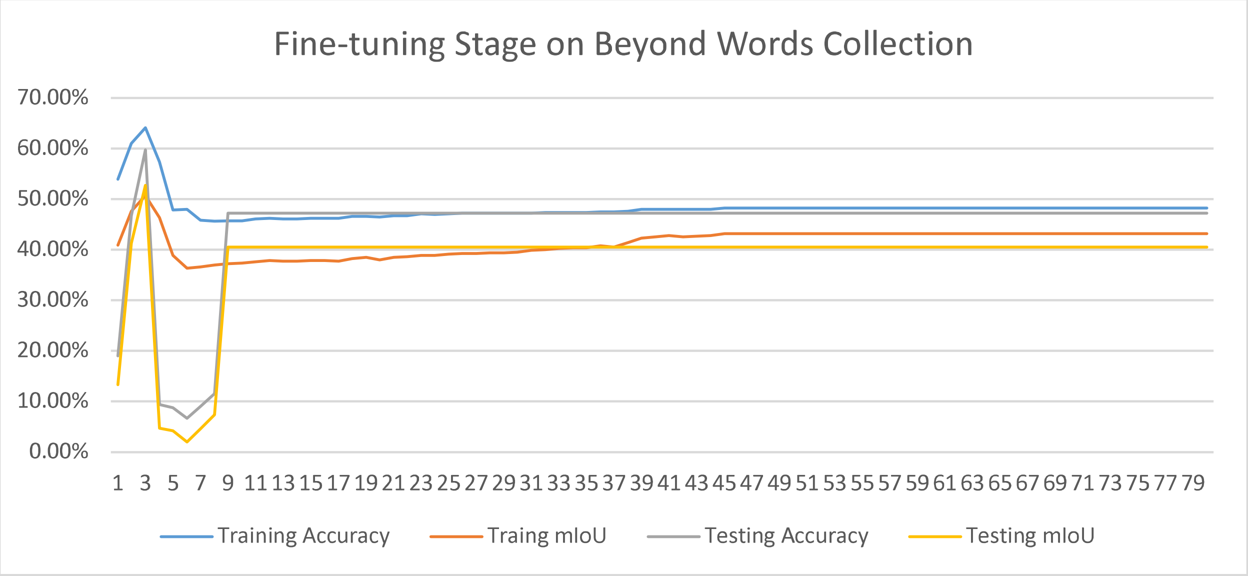
Shown in Figure 2, the fine-tuning stage on the BW was able to reach 64% on training accuracy and 47% on training mIoU, and reach 59% on testing accuracy and 52% on testing mIoU at an early iteration (i.e., at third iteration). However, the fine-tuning stage tried to classify all pixels as background 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

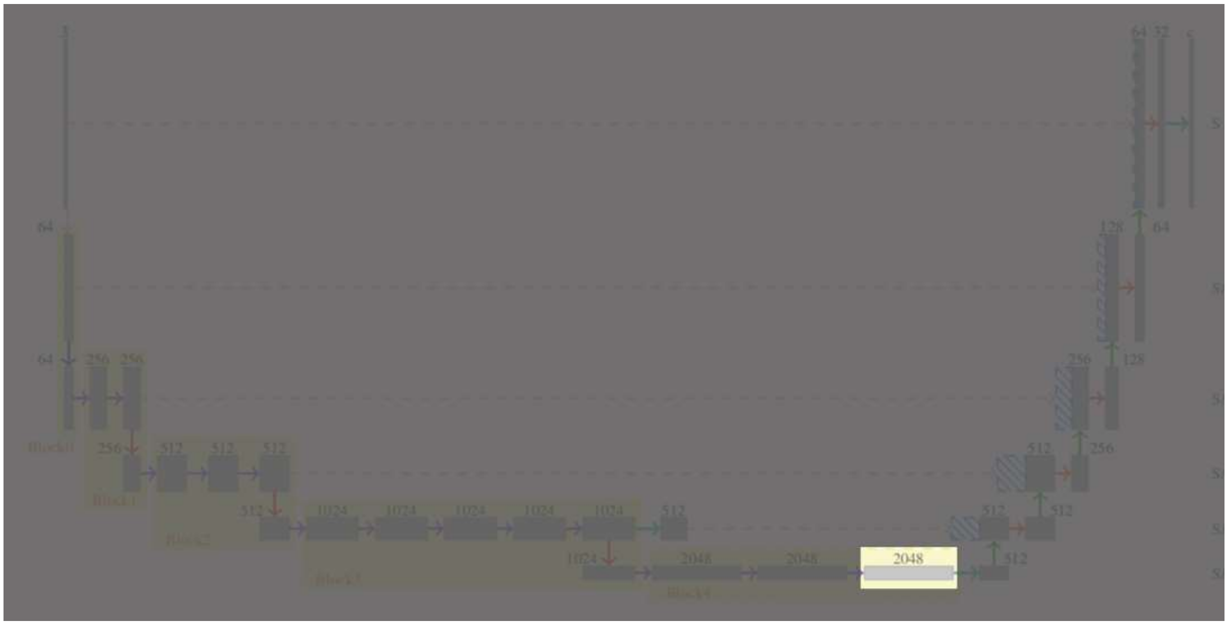


Figure 1. Latent space of *dhSegment*. This model is a combination of ResNet-50 and U-net. From this model, we extract only the highlighted block, which is known as a latent space.

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.

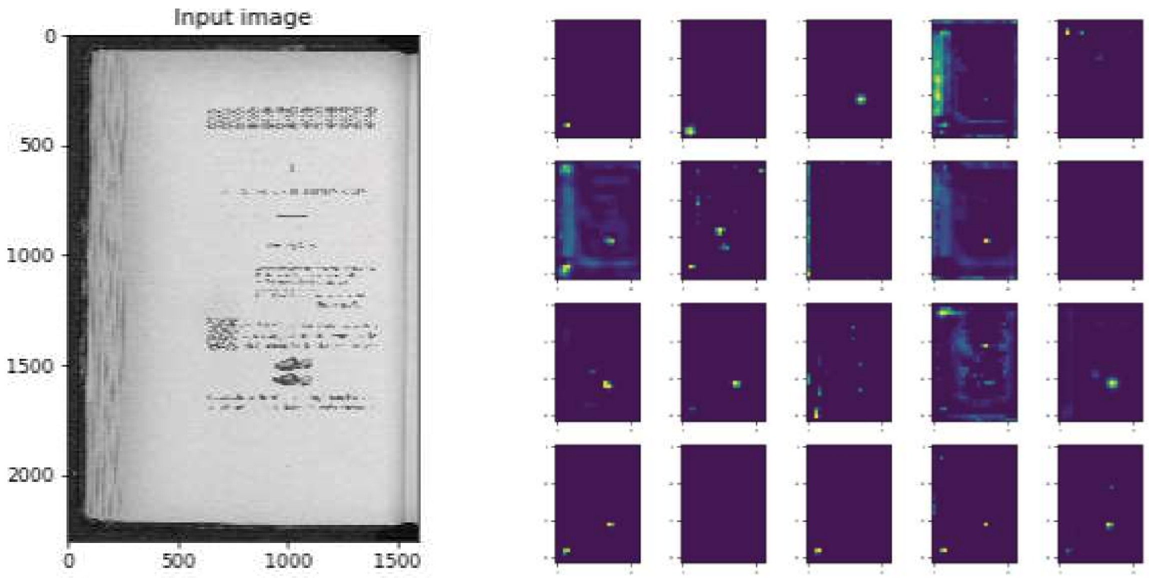


Figure 2. **Left**: Input image; **Right**: Visualized latent space down-sampled to *W*/32 x *H*/32 x 25

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.

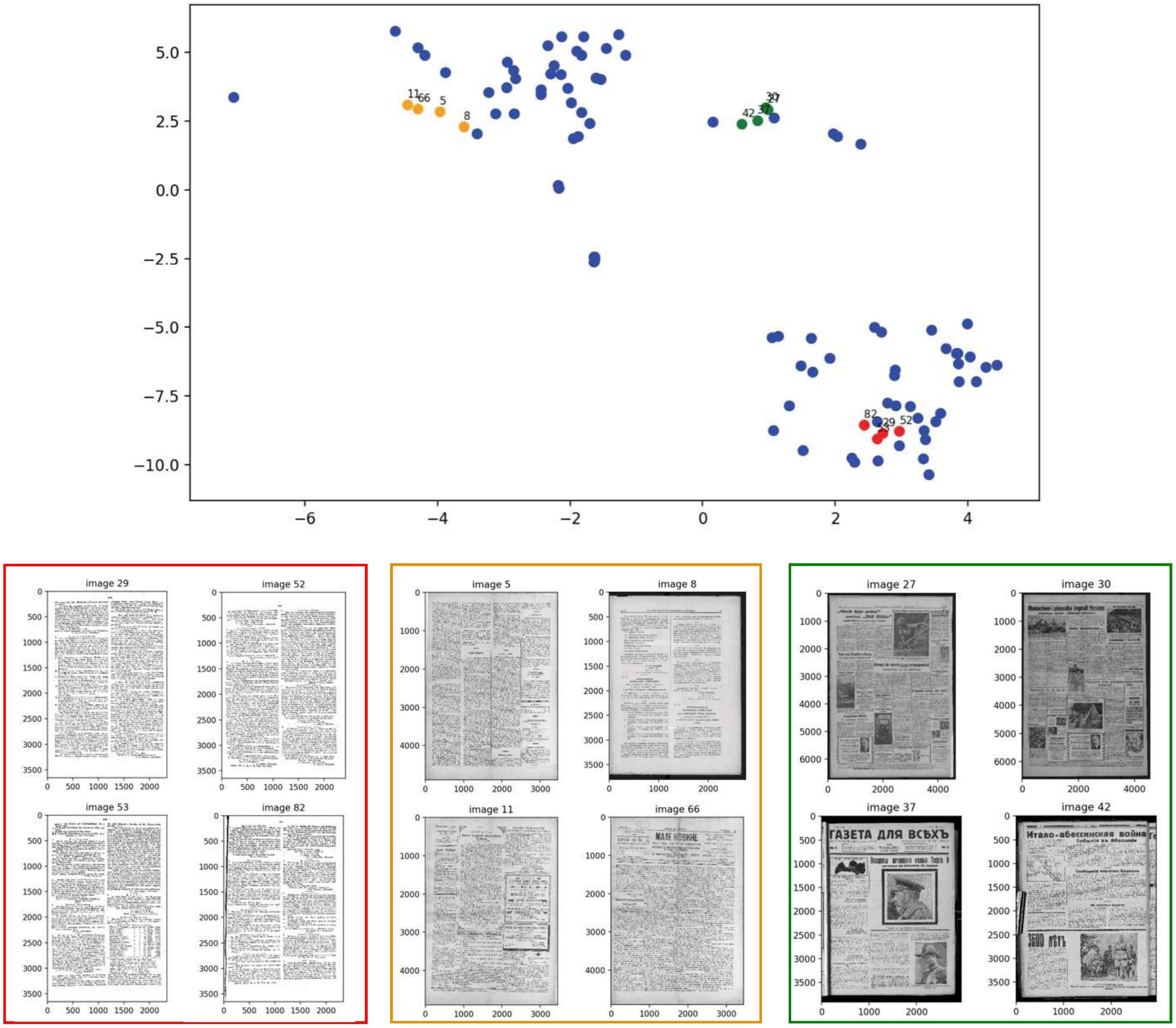


Figure 3. **Top row:** Visualization of latent space of ResNet-50 mapped into a low-dimensional space using t-SNE. For visual similarity inspection purpose, we selected four datapoints from different clusters labeled them in different colors; orange, green, and red. On the top-right of each exemplary 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.

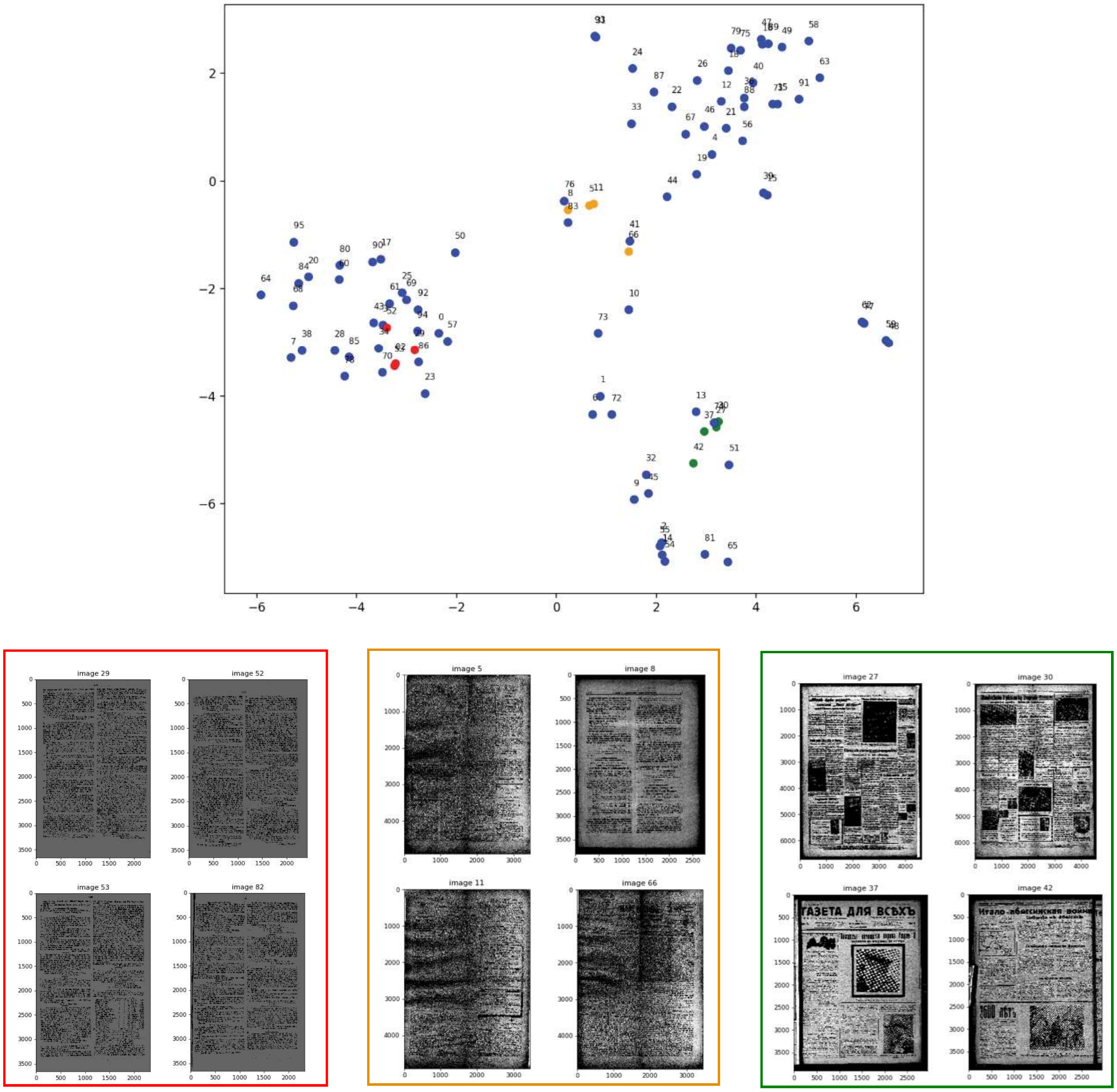


Figure 4. **Top row:** Visualization of latent space of ResNet-50 with normalization mapped into a low-dimensional space using t-SNE. For visual similarity inspection purpose, we selected four datapoints from different clusters labeled them in different colors; orange, green, and red. On the top-right 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.

University of Nebraska-Lincoln
Progress report

Proposal# 129426
Oct 12th, 2019

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.

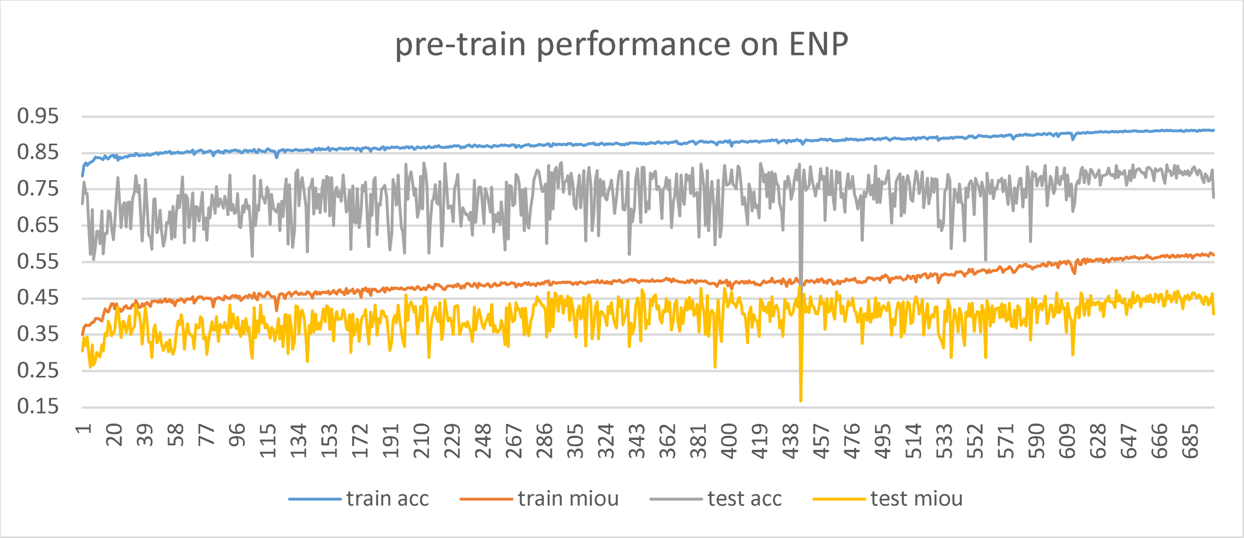


Figure 1 Pre-train ENP classifier performance.

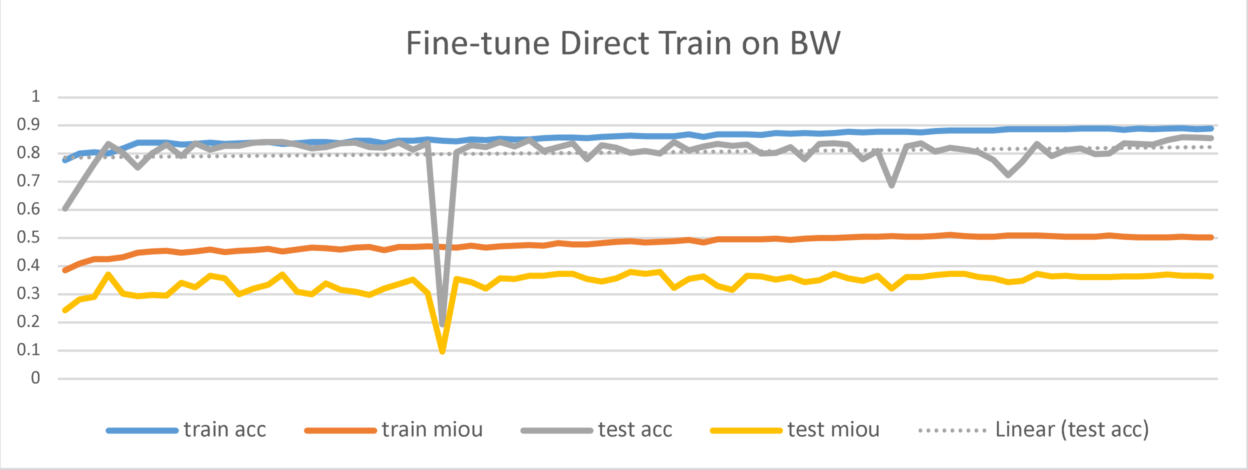


Figure 2 Fine-tuning experiment 1 - the baseline.

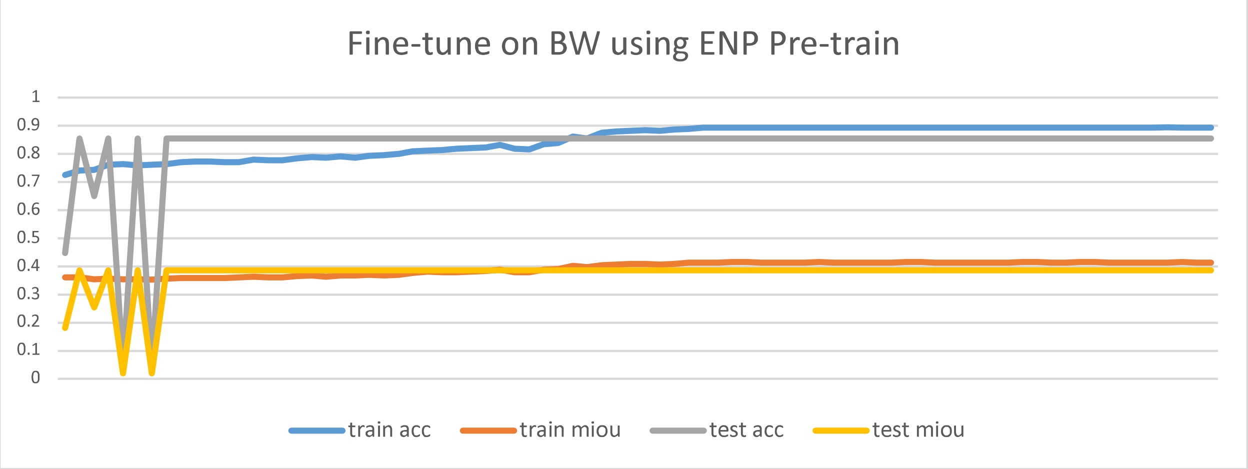


Figure 3 Fine-tuning experiment 2 - using pre-trained ENP classifier.

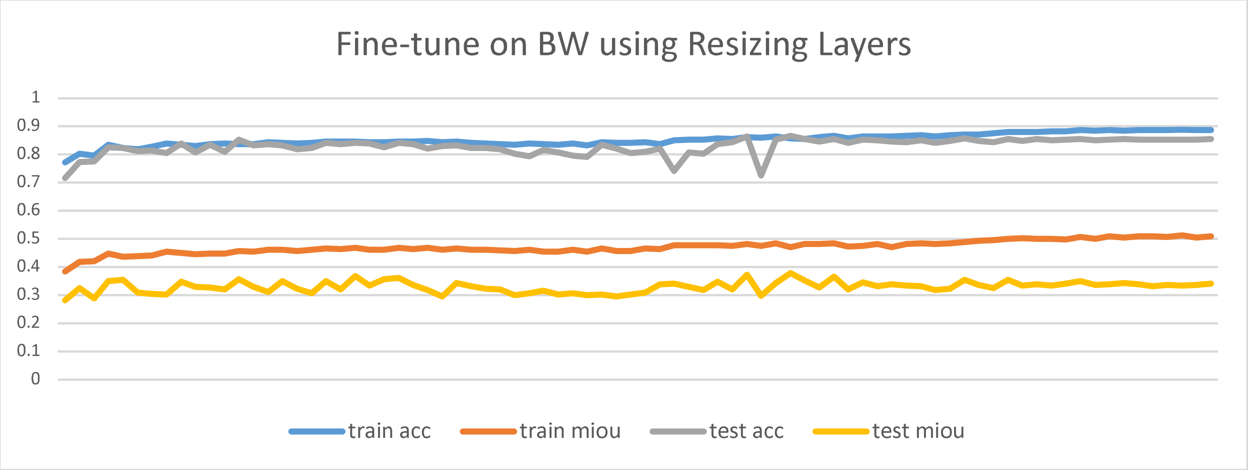


Figure 4 Fine-tuning experiment 3 - using resizing layers.

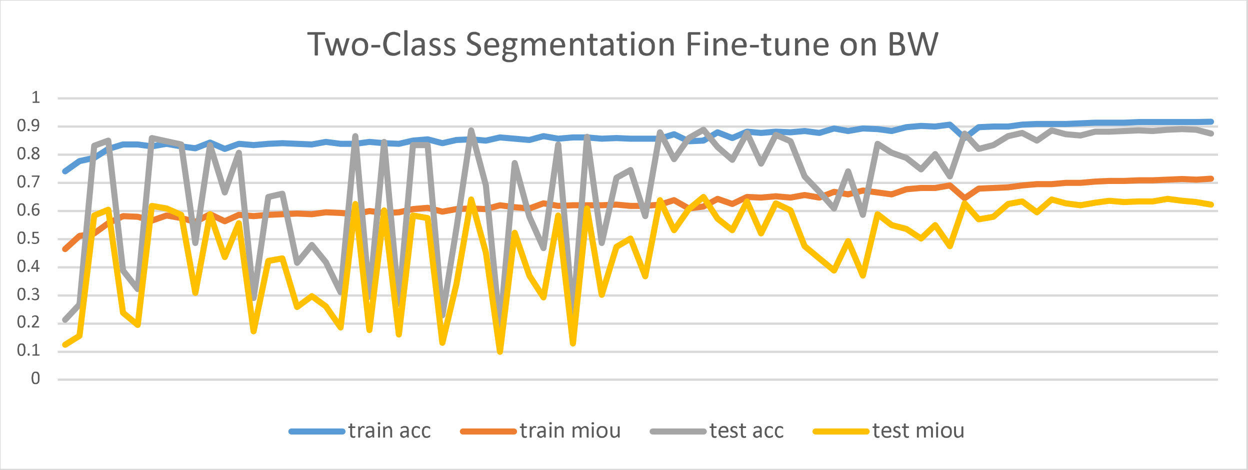
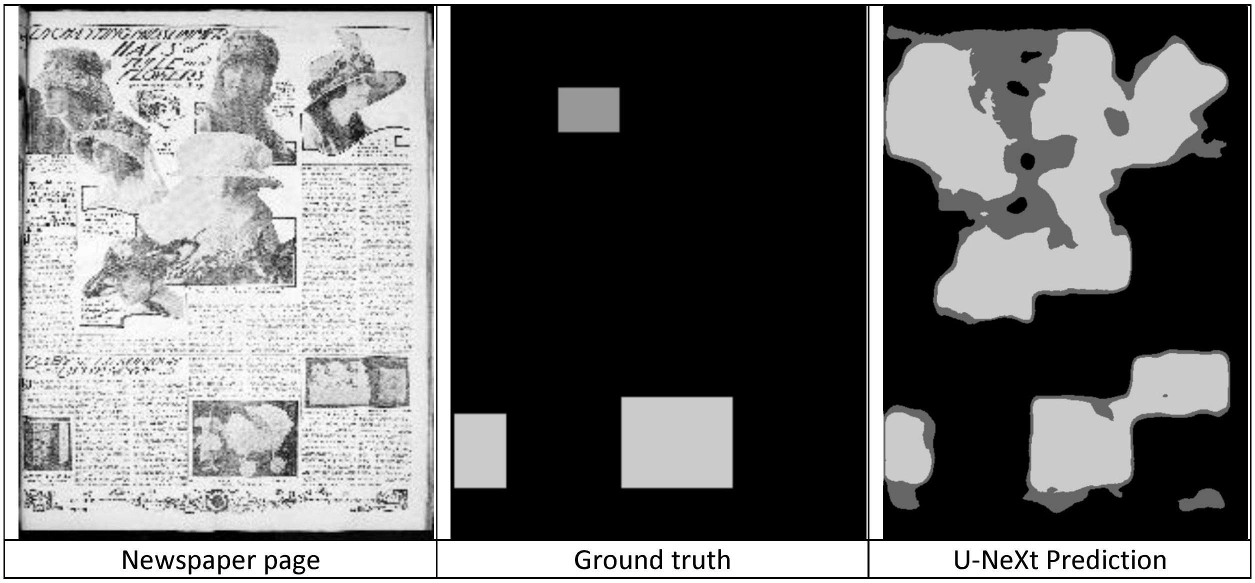
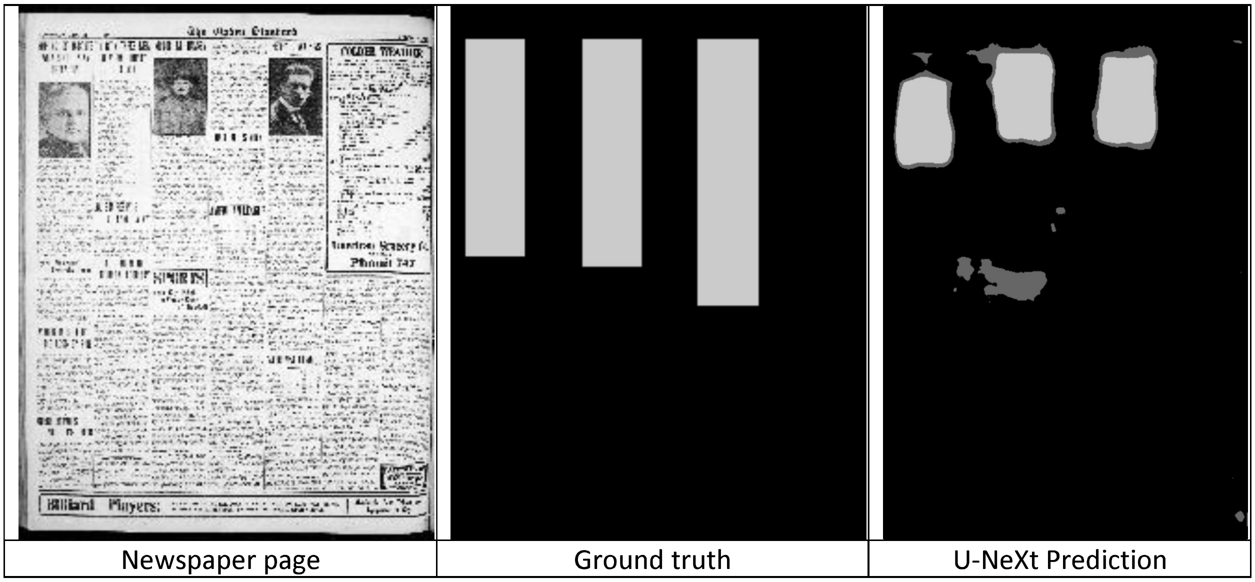


Figure 5 Fine-tuning experiment 4 - combined two-class segmentation.



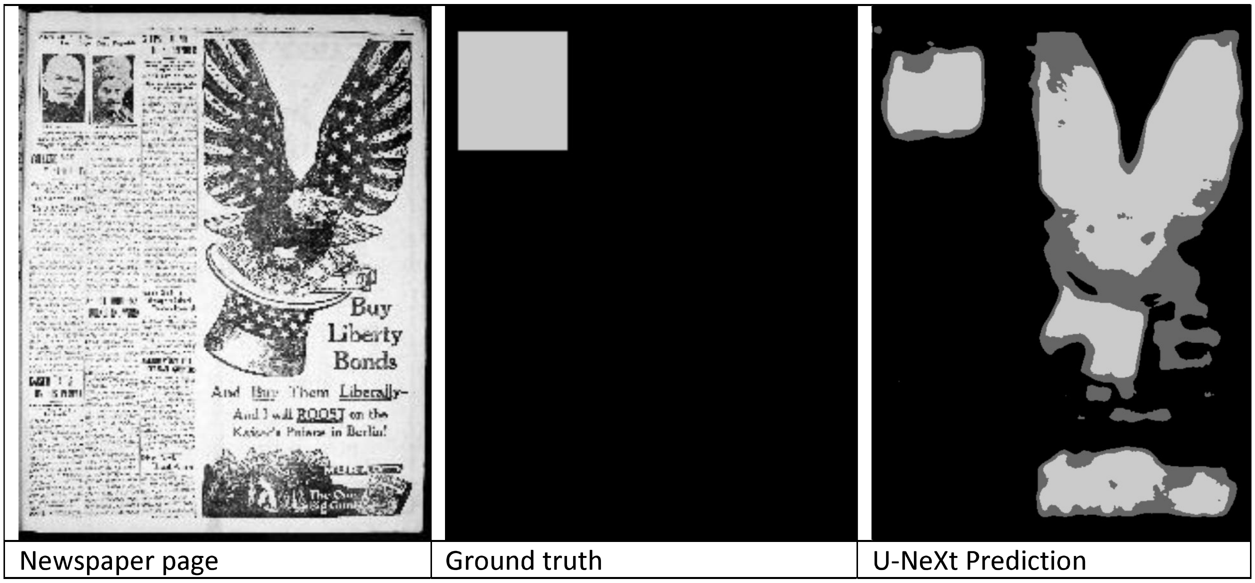
Newspaper page	Ground truth	U-NeXt Prediction

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:

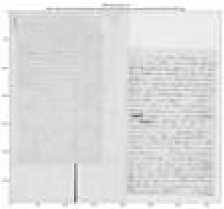
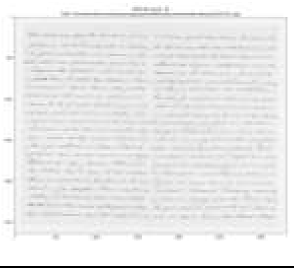
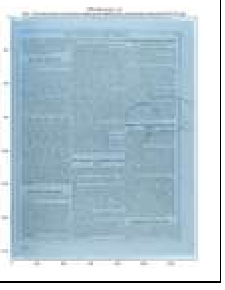
Observation 1. The same *visual feature* deemed to related to the *difficulty score* in *typed* documents is not deemed to related to the difficulty score in *handwritten* documents.

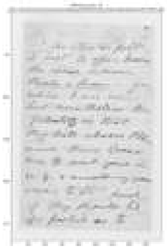
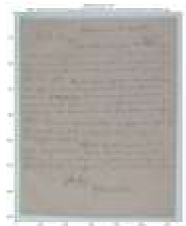
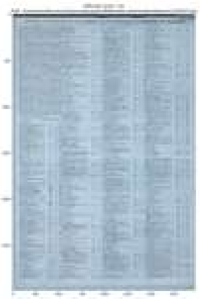
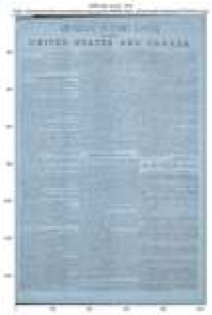
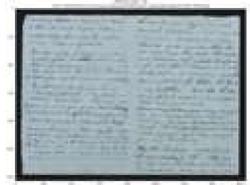
Observation 2. It is hard to expect to find correlation between a simple standalone *visual feature* (e.g., number of characters or low contrast) and a *difficulty score.*

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.

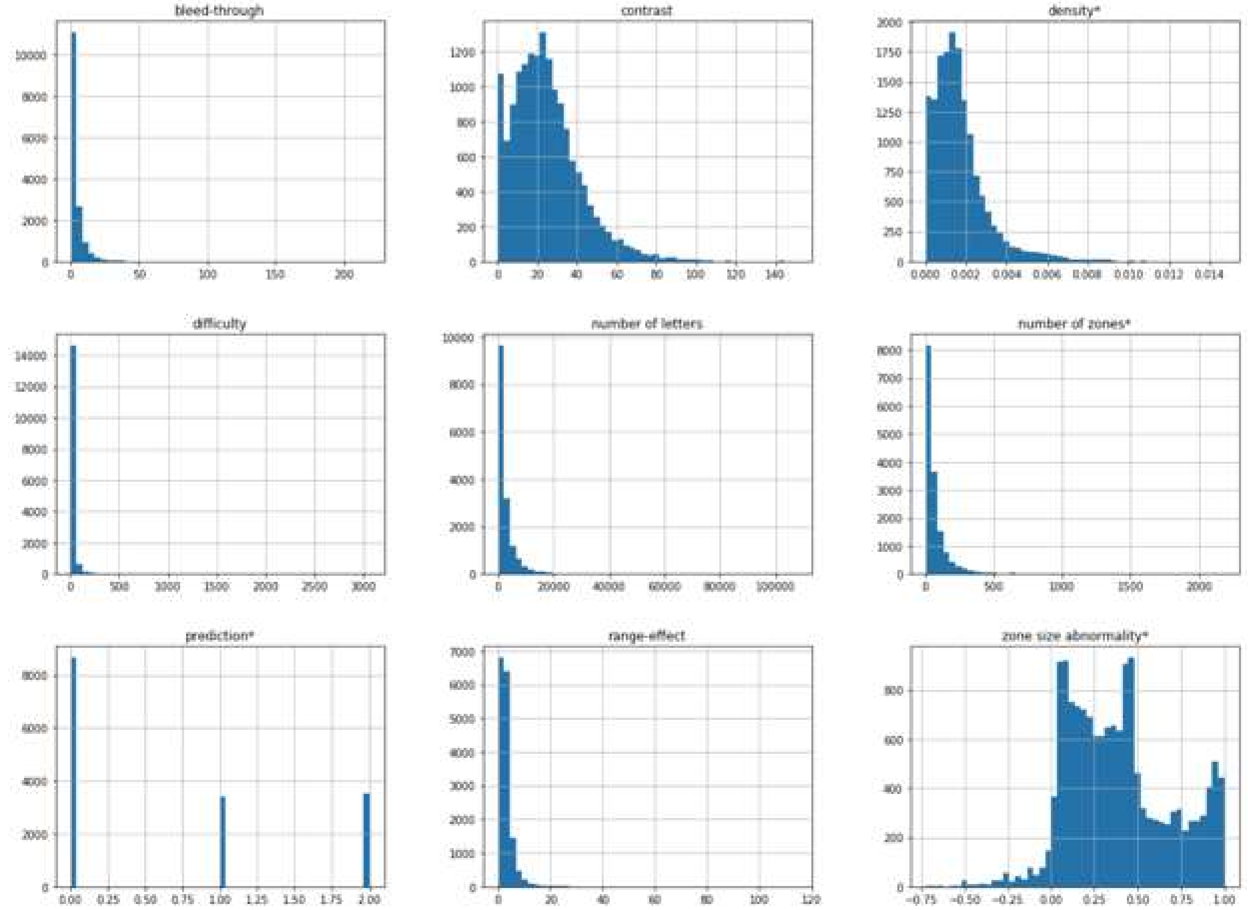


Figure 1. Distribution of visual features over 15,592 images. Note that one of the assumptions behind the Pearson's correlation coefficient is that variables (i.e., visual features) should be normally distributed.

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.

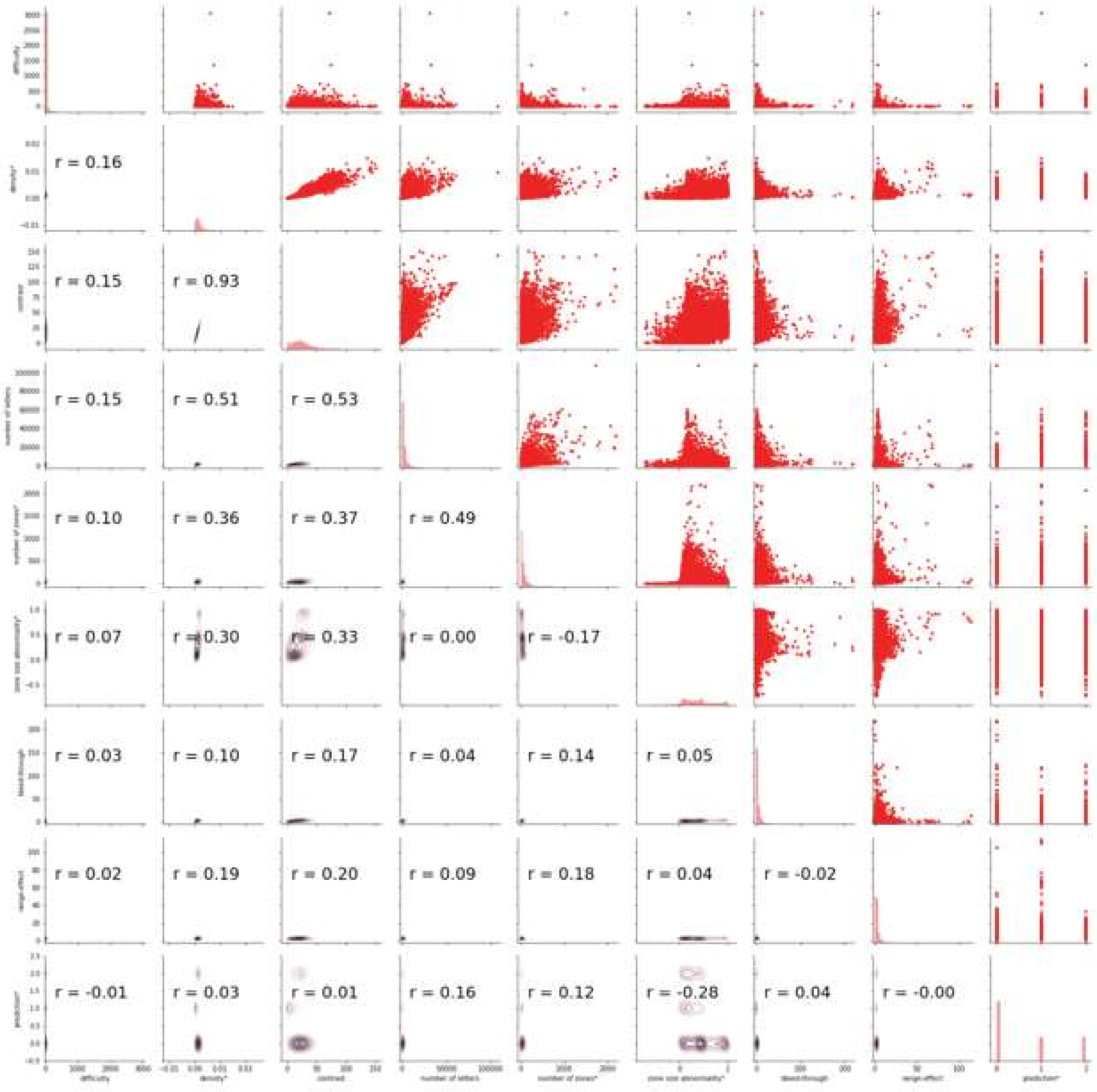


Figure 2. Pairs plot to visualize scattered distribution and relationship between two visual features. The *r* value represents the size of correlation (range from -1 to 1, best value at -1 or 1, worst value at 0.) It is worth noting that the relationship between most of the *visual features* with the *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*.