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.

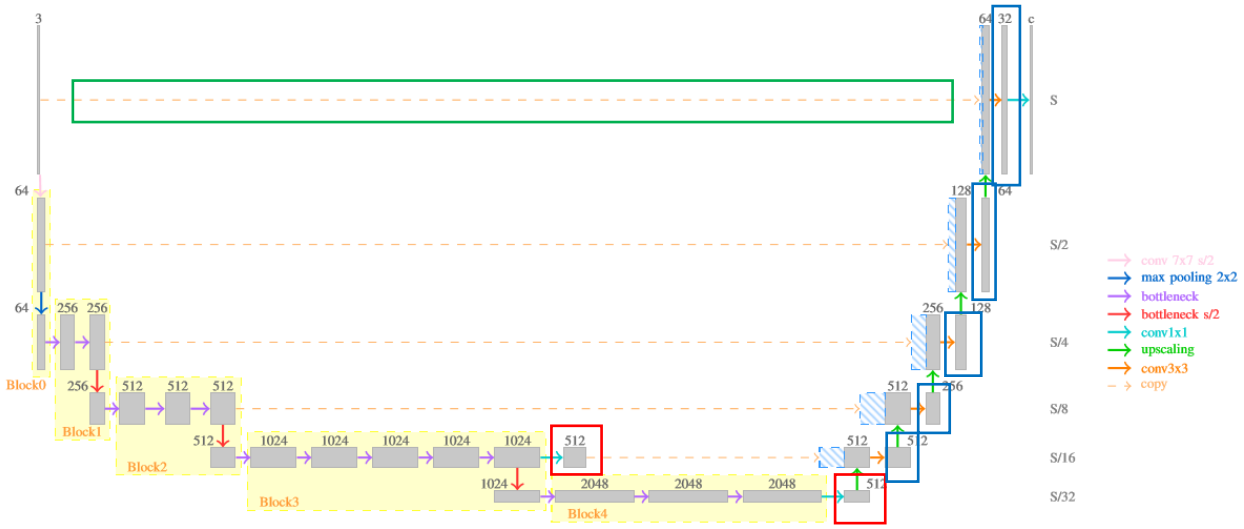


Figure 1 Network architecture of dhSegment

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.

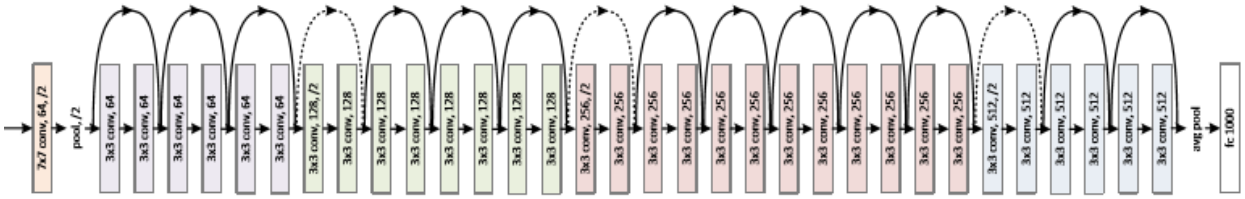


Figure 2 Network architecture of ResNet

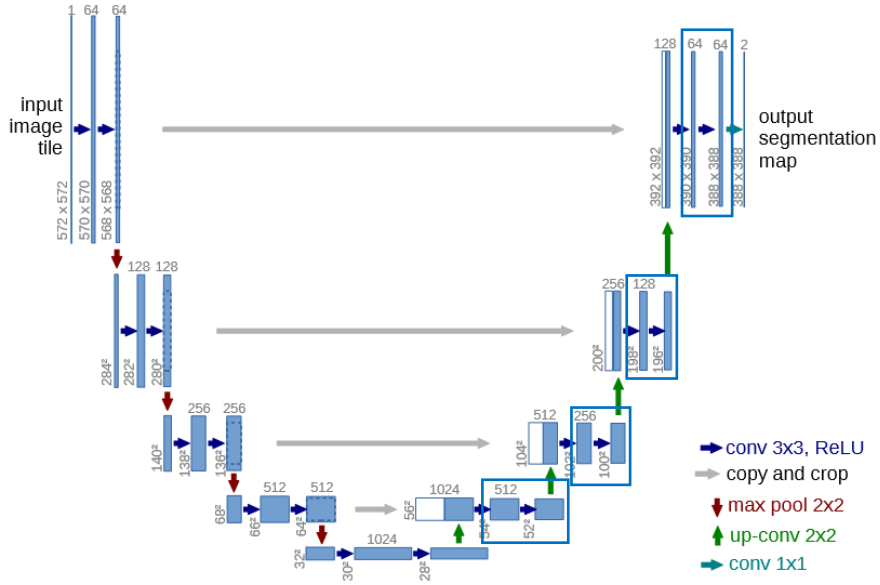


Figure 3 Network architecture of U-Net

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.

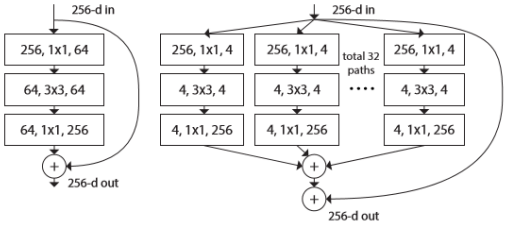


Figure 4 ***(Left)*** *Residual blocks of ResNet.* ***(Right)*** *Residual blocks of ResNext.*

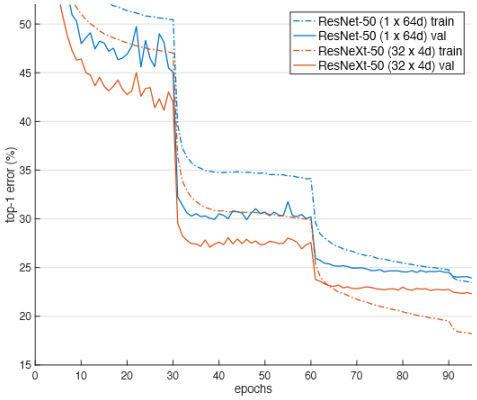


Figure 5 Comparative results between ResNet and ResNeXt on ImageNet-1K dataset.

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).

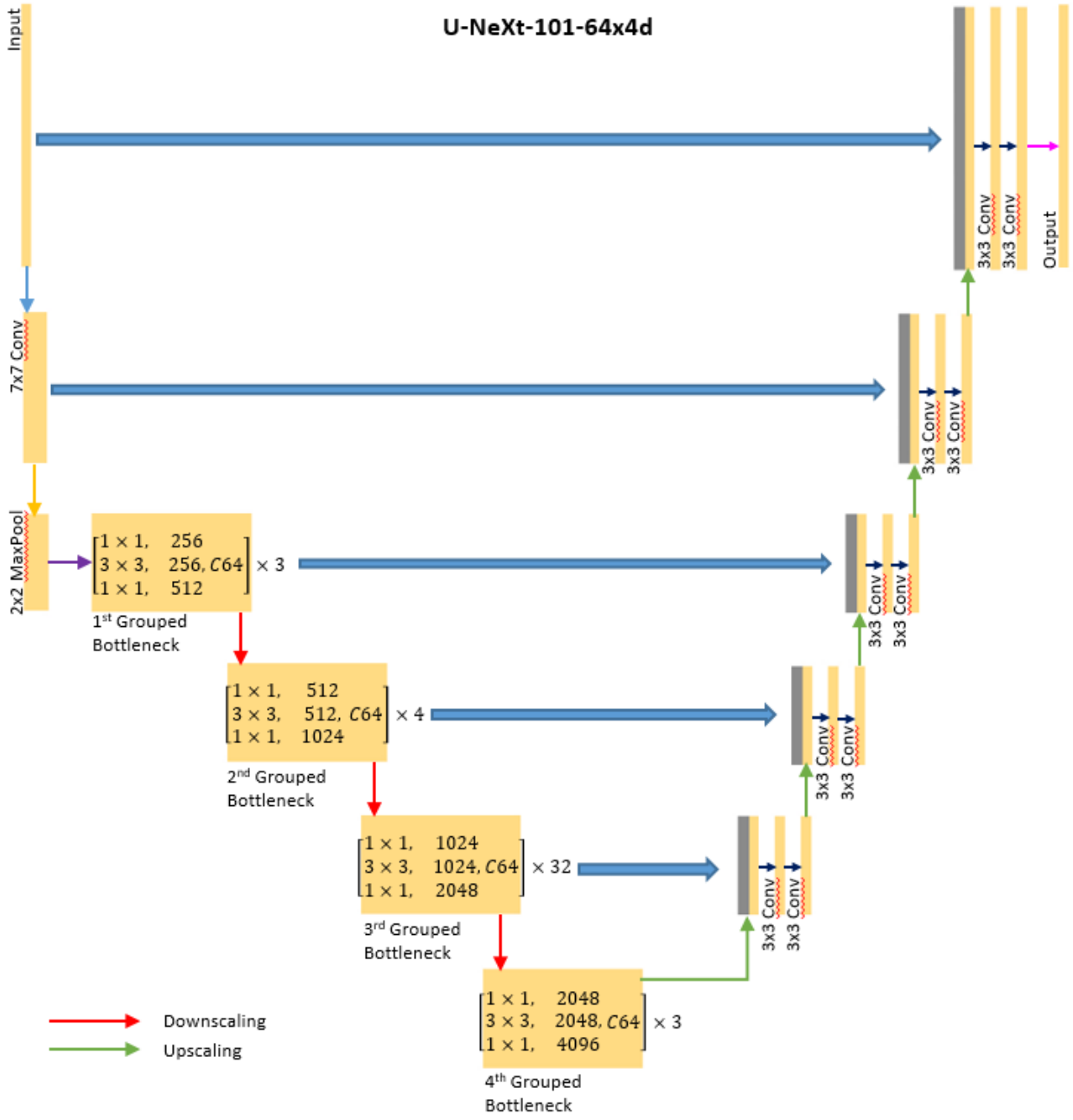
[4] Oliveira, S. A., Seguin, B., & Kaplan, F. (2018, August). dhSegment: A generic deep-learning approach for document segmentation. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)* (pp. 7-12). IEEE.

[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
Output
Downscaling
Upscaling