A Brief Introduction to LingDIAR System

Our DIAR system is based on a multi-task deep network model trained with synthetic document images. The model uses a 13-layer Resnet as back-bone network, in which 7 max-pooling layers down sample the feature map into 1/128 of the original image size. Then a FPN (feature pyramid network) is used to merge the 7 scale feature maps as the shared input representation features for the predication network of each task. There are 3 main tasks in our DNN model, including page segmentation and classification, textline detection and textline recognition.

For page segmentation and classification task, we use a FCN based pixel-wise labelling network, as in the work of [1] and [2]. To avoid the merging error for neighboring text regions, we also add a page object contour detection output to the network, as in [1]. During evaluation, the page segmentation results are refined with contour detection results via some post-processing steps.

For textline detection task, firstly we split textline into small segments along the textline orientation as in [3], and then use a FCN based regression network to directly predict the score, bounding box and orientation of each text segment as in [4], finally the text segments are grouped into textlines with some geometrical rules.

For textline recognition task, a spatial transformer network and some vertical max-pooling layers are firstly used to transform the representation features of each textline into canonical sequential features, as in the work [5], then a bi-directional LSTM is used to capture their range dependencies. For each time frame, a fully-connected network is used to output the character classification scores. Finally, CTC is applied to transform frame-wise classification scores to label sequence in the training phase.

To train the model sufficiently, we use a synthetic method to automatically generate a large number of training samples, as in the work of [1] and [2]. Firstly more than 2000 document images with manually labelled page layouts are collected. Then these page layouts are used as templates to generate more document images by replacing each region with a new element. For figure and table region element, we collect large number of samples from the internet. For text region element, we directly render text on the text region with various fonts, colors and sizes. During the document image synthesing, the corresponding layout annotations are also adjusted as necessary.

In the evaluation phase, document images are first normalized and fed into the trained DNN model. Then the region segmentation and contour detection outputs are combined to obtain the region area and boundary polygon of each page element. After that, textlines are obtained by grouping the text segments produced from the textline detection network in each region area. Given all textline bounding boxes in a text region, its contour polygon can be simplified to less number of points. In the end, features of each textline are normalized and fed into the textline recognition network to get the character strings, and the recognition results are also used to eliminate some segmentation and detection errors in previous steps.

References

[1] He D, Cohen S, Price B, et al. Multi-scale multi-task fcn for semantic page segmentation and table detection[C]//2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2017, 1: 254-261.

[2] Yang X, Yumer E, Asente P, et al. Learning to extract semantic structure from documents using

multimodal fully convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5315-5324.
[3] Tian Z, Huang W, He T, et al. Detecting text in natural image with connectionist text proposal network[C]//European conference on computer vision. Springer, Cham, 2016: 56-72.
[4] Zhou X, Yao C, Wen H, et al. EAST: an efficient and accurate scene text detector[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017: 5551-5560.
[5] Liu X, Liang D, Yan S, et al. Fots: Fast oriented text spotting with a unified network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5676-5685.