RDCL2019: ICDAR Competition on Recognition of Documents with Complex Layouts

Yassine Ouali, Céline Hudelot.
MICS, CentraleSupélec, 91190 Gif-sur-Yvette, France.

In this short document, we briefly explain the basis of our segmentation approach we used in our RDCL2019 submission, by describing the model, the training procedure, the post processing steps and some inference details.

The model

The model we used is very similar to PSPNet [1] with some additionnal changes, the backbone is a Renset 50 [3] with an *output_stride* of 8 (i.e., the spatial dimensions of the output is 1/8 of the input) instead of 32, this is done by replacing the *conv* 3×3 with stride 2 in the last two blocks with a *conv* 3×3 with dilation rates of 2 and 4 respectivelly, to maintain a similar receptive without any further reduction of the spatial dimensions, the outputs are then fed into a PSP (Pyramid Scene Parsing) module to add a global information into low level segmentation and help the network detect inconspicuous classes, and given that in documnent segmentation the low level information contrains a significant learning signal, we combine the outputs of the intermediate resnet layers using a succesion of symetric filters and bilinear upsampling similar to [2], we combine these features and add them to the PSP module, and finally we upsample the feature maps to obtain an output with the same size of the input image (see Figure 1).

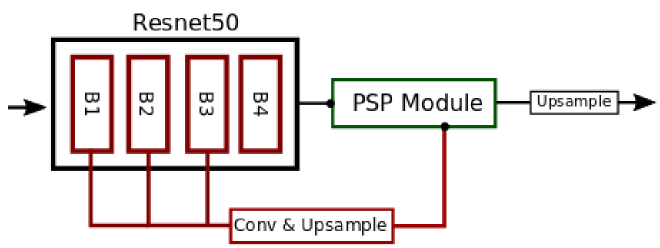

Figure 1: High level disciption of the semantic segmentation network used.

Training

For training the network, we use the PRImA Layout Analysis Dataset [4], and to resist overfitting we adopt a comprehensive data augmentation, we use random horizental flipping, random resize of the input between 0.5 and 1.5, random Gaussian blur and finally a random crop of the image.
Given the limited GPU memory, we use a batch size of 8 with a base learning rate of 5×10^{-4}, we adjust the learning rate following the one cycle procedure [6], starting from $\frac{lr_{base}}{25}$ to lr_{base} for 30% of the epochs which are set to 100 epochs, and then decreasing back to $\frac{lr_{base}}{25}$ for the rest of the training time.
The training is done in an iterative way, first by only leaning the added modules to the backbone; then fine tunning the whole model, and then traning for some additionnal epochs using larger image crops and smaller batches.

Inference and Post Processing

During inference, we predict the segmentation mask for various input scales (e.g., [0.4, 0.6, 0.8, 0.9]), upsample them to the original image size and take their average, the result is a single mask of size $H \times W$, each element $\in [0, C-1]$ where C is the number of classes (i.e., C = 12), we then apply a post processing step, first we apply a fully connected CRF [7] to the output probabilities, this is done to refine the segmentation mask based on both the model's output and the image colors, we then transfrom the output mask into C binary masks to apply basic morphological operators (closing followed by an openning) to further refine the masks and filter out small connected components, and finally in order to transform the detected regions into a set of polygons, we extract the blobs in each binary mask as a set of coordinates and write them in the PAGE XML format [5].

References

[1] Hengshuang Zhao1, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, Jiaya Jia, *Pyramid Scene Parsing Network*, CVPR 2017.

[2] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, Jian Sun, *Large Kernel Matters – Improve Semantic Segmentation by Global Convolutional Network.*

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, *Deep Residual Learning for Image Recognition, CVPR 2016.*

[4] A. Antonacopoulos, D. Bridson, C. Papadopoulos, and S. Pletschacher, *A Realistic Dataset for Performance Evaluation of Document Layout Analysis*, ICDAR2009.

[5] Stefan Pletschacher ; Apostolos Antonacopoulos, *The PAGE (Page Analysis and Ground-Truth Elements) Format Framework*, 20th International Conference on Pattern Recognition, 2010.

[6] Leslie N. Smith, *A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay*

[7] Philipp Krähenbühl, Vladlen Koltun, *Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials*