Geometric layout analysis with CNNs

Klára Janoušková, Michal Bušta, Jiří Matas
Center for Machine Perception, Department of Cybernetics
Czech Technical University, Prague, Czech Republic

1 Method description

The method consists of the following steps:

1. Obtaining segmentation maps from a CNN
2. Post-processing steps
3. OCR

1.1 Segmentation maps

We train a CNN on the PRImA Layout Analysis Dataset extended by about 1000 automatically annotated images from arXiv using the pdfMiner python library (axis aligned bounding boxes for text regions). The 4-channel output consists of the probability of a pixel forming part of a text region, the probability of a pixel forming part of an edge of a text region, the probability of a pixel forming part of a separator region and the probability of a pixel forming part of any of the remaining regions (tables, graphs, images, ..). The final segmentation maps are obtained by applying a threshold of 0.9 on the predicted probabilities, each channel forming one segmentation map.

1.2 Post-processing

We perform multiple post-processing steps - first, we obtain the polygon representation of the regions by extracting connected components from he segmentation maps for text, separator and other regions. We split text regions merged in the text segmentation map by detecting horizontal lines in the text region edge map. We also dilate the text polygons to make up for some small mistakes in segmentation. Then we discard very small regions that are likely to be false positives.

1.3 OCR

As a last step, we run a CNN based OCR on the cropped axis-aligned bounding box of each text region to obtain the text transcription.