Vision-Based Layout Detection from Scientific Literature using Recurrent Convolutional Neural Networks

Huichen Yang, William H. Hsu
Department of Computer Science, Kansas State University, USA
Laboratory for Knowledge Discovery in Databases
{huichen, bhsu} @ ksu.edu

Abstract

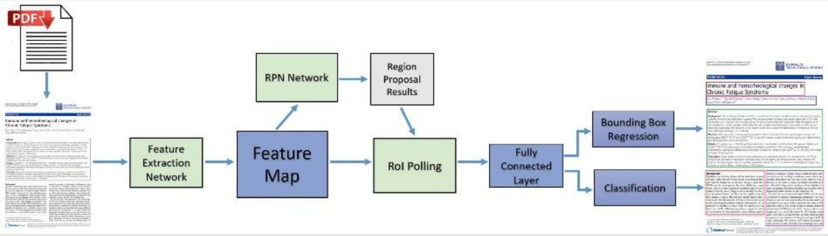
We present a novel approach to developing an end-to-end learning framework to segment and classify major regions of a scientific document. We consider scientific document layout analysis as an object detection task over digital images, without any additional text features that need to be added into the network during the training process. Our technical objective is to implement transfer learning via fine-tuning of pre-trained networks and thereby demonstrate that this deep learning architecture is suitable for tasks that lack very large document corpora for training ab initio. As part of the experimental test bed for empirical evaluation of this approach, we created a merged multi-corpus data set for scientific publication layout detection tasks. Our results show good improvement with fine-tuning of a pre-trained base network using this merged data set, compared to the baseline convolutional neural network architecture.

Introduction

- ➢ Unstructured scientific literature contains huge valuable information
- ➢ Rate of published scientific literature is growing rapidly into huge data set
- ➢ Lack of metadata information extraction for existing tools, e.g., OCR
- ➢ Scientific literature layout detection presents a solution of automatic construction of large corpus for downstream tasks of NLP
- ➢ Consider scientific literature layout detection (SLLD) as object detection task of computer vision
- ➢ Present novel approach to detect the main regions of scientific articles, and output blocks and their corresponding labels
- ➢ Create synthesis dataset for scientific literature documents layout detection (SLLD) task

Methodology

- ➢ Use two-stage object detection framework
- ➢ Use pre-trained VoVnet-v2 [1] on MS COCO dataset as backbone for feature extraction
  - Better performance – aggregate concatenation feature only once in last feature map
  - Residual connection enables to train deeper networks
  - Squeeze-and-Excitation (eSE) attention module improves feature extractor performance
- ➢ Use K-means to select Anchors aspect ratio
  - Analyze distribution bounding box sizes-based ground truth synthesis dataset
  - Use K-means cluster anchor box selection to get aspect ratios for different blocks ranging from 0.1 to 4.0



Scientific literature layout detection framework

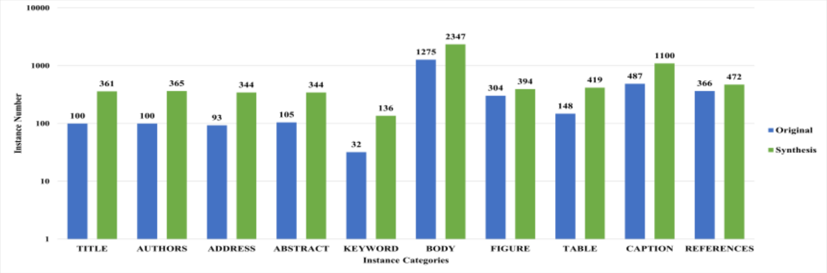
Synthesis dataset

We propose synthesis dataset to relieve imbalance issue with merged three dataset together

- ➢ Region annotations dataset – 822 images from 100 PDF scientific literature [2]
- ➢ ICDAR-2013 – 150 table images from 76 PDF documents [3]
- ➢ GROTOAP- 113 annotated PDF scientific literature [4]

Final synthesis dataset

- ➢ 1550 images from 363 PDF documents
- ➢ Convert these images to fixed size – 612 x 729 at 200 dpi

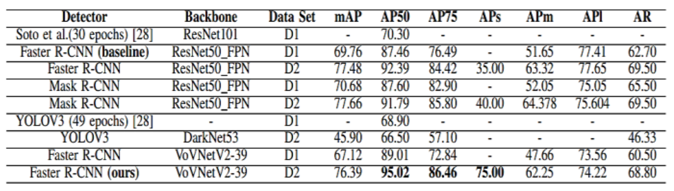


Instances comparison between region annotations dataset and our dataset by labels

Results

We use IoU (intersection of unit) to evaluate our model on two dataset with different methodologies

- ➢ Data set1(D1): original data set [2] - 600 image for training, 222 images for testing
- ➢ Data set2(D2): synthesis data set - 1225 images for training, 325 images for testing



Detector	Backbone	Data Set	mAP	AP50	AP75	APs	APm	APl	AR
Soto et al.(30 epochs) [28]	ResNet101	D1	-	70.30	-	-	-	-	-
Faster R-CNN (**baseline**)	ResNet50_FPN	D1	69.76	87.46	76.49	-	51.65	77.41	62.70
Faster R-CNN	ResNet50_FPN	D2	77.48	92.39	84.42	35.00	63.32	77.65	69.50
Mask R-CNN	ResNet50_FPN	D1	70.68	87.60	82.90	-	52.05	75.05	65.50
Mask R-CNN	ResNet50_FPN	D2	77.66	91.79	85.80	40.00	64.378	75.604	69.50
YOLOV3 (49 epochs) [28]	-	D1	-	68.90	-	-	-	-	-
YOLOV3	DarkNet53	D2	45.90	66.50	57.10	-	-	-	46.33
Faster R-CNN	VoVNetV2-39	D1	67.12	89.01	72.84	-	47.66	73.56	60.50
Faster R-CNN (**ours**)	VoVNetV2-39	D2	76.39	**95.02**	**86.46**	**75.00**	62.25	74.22	68.80

Overall comparison among SLLD results (%) with different methodologies and data set at IoU = 0.5:0.05:0.95

Conclusion

We introduce a novel end-to-end learning and vision-based framework. The major regions in scientific literature documents will be detected through the model which is trained by our framework. This model not only detects text regions but also figures and tables. Our approach is easy to adapt and implement for a broad range of scientific literature formats and domains, since it does not require extraction of additional features. We also created a merged multi-corpus data set for scientific publication layout detection tasks.

References

[1] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).

[2] Soto, C., & Yoo, S. (2019, November). Visual Detection with Context for Document Layout Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP)* (pp. 3455-3461).

[3] Gobel, M., Hassan, T., Oro, E., & Orsi, G. (2013, August). ICDAR 2013 table competition. In *2013 12th International Conference on Document Analysis and Recognition* (pp. 1449-1453). IEEE.

[4] Tkaczyk, D., Czeczko, A., Rusek, K., Bolikowski, L., & Bogacewicz, R. (2012, June). GROTOAP: ground truth for open access publications. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries* (pp. 381-382).