A Deep Learning Method for Mathematical Formulas Detection in PDF Documents

Nghia Vo Trong
Faculty of Information Technology
University of Science, VNU−HCM
Ho Chi Minh City, Vietnam
20120536@student.hcmus.edu.vn

Van-Loc Nguyen
Faculty of Information Technology
University of Science, VNU−HCM
Ho Chi Minh City, Vietnam
20120131@student.hcmus.edu.vn
ORCID: 0000-0001-9351-3750

Minh-Tam Nguyen Kieu
Faculty of Information Technology
University of Science, VNU−HCM
Ho Chi Minh City, Vietnam
20120572@student.hcmus.edu.vn

Dang Nguyen Hai
Faculty of Information Technology
University of Science, VNU−HCM
Ho Chi Minh City, Vietnam
nhdang@selab.hcmus.edu.vn

***Abstract*—Write the abstract here.**

***Index Terms*—**

I. Introduction

II. Related Works

III. Method

IV. Experiments

In this section, we will describe the implementation of our mathematical formula detection system and dataset in detail.

A. Dataset

Our data is from the IBEM dataset. This originally comprises 600 documents, with 8273 pages in total. Those documents are parsed from mathematical papers, then each page is annotated with a bounding box of 2 types: isolated and embedded. The dataset is then split into various sets for ICDAR 2021 Competition on Mathematical Formula Detection, including Training, Test, and Validation sets.

Training

- Tr00: 4082 pages.
- Tr01: 760 pages.
- Tr10: 329 pages.

Test

- Ts00: 736 pages.
- Ts01: 380 pages.
- Ts10: 699 pages.
- Ts11: 329 pages.

Validation

- Va00: 577 pages.
- Va01: 380 pages.

Our experiment uses Tr01, Tr10, Ts01 for training, Va01 for validation, and Ts11 for testing with 2178 pages in total ($\sim 26.33\%$ of the original dataset), and an approximate ratio of $4.47 : 1.16 : 1$. The reason for this small subset is for the purpose of evaluating the ability of the model on small subsets, and the performance it gives (F1-score) through time (minutes).

B. Implementation Details

Our baseline model is Faster R-CNN with ResNet50 as the backbone. We have trained on Kaggle with a 4-core CPU, 12GB RAM, and a NVIDIA Tesla P100 GPU [1]. The images are resized to 1447×2048 with the same ratio. The size of the region crops from the image is 1200×1120 to fit the limitation of the machine. They are also flipped and padded for data augmentation. For the feature aggregation, we use FPN (2-6). The loss function for the classifier is Cross-Entropy Loss and for the bouding box is L1 Loss. Test images are resized to 1583×2048 due to the distribution of the test dataset, flip augmentation is also applied. For post-processing, Non-Maximum Suppression (NMS) with 0.5 IoU threshold to remove redundant boxes. All models are trained based on the MMDetection toolbox and config given by Yuxiang Zhong. The optimizer for this baseline is Stochastic Gradient Descent (SGD) with a learning rate of 0.02.

C. Remarks

We have tested on 3 configs: Faster R-CNN with schedule 1x (12 epochs), Dynamic R-CNN with schedule 1x (12 epochs) to check if it is better than the faster one and Faster R-CNN with schedule 2x (24 epochs) to check if the model is underfitting with low epochs.

[1]https://www.kaggle.com/docs/notebooks

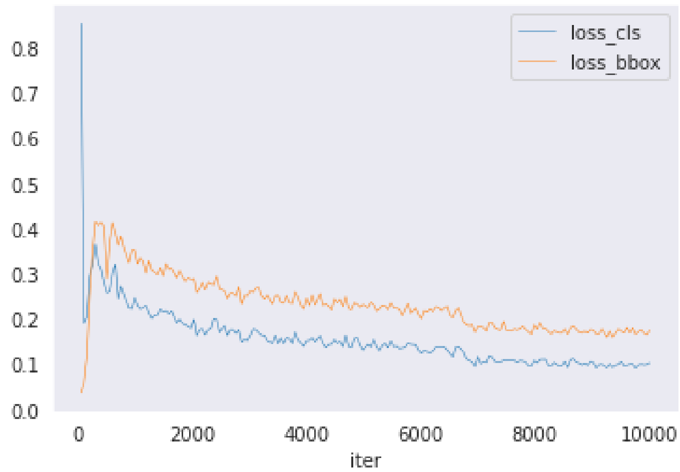

Fig. 1. Faster R-CNN with schedule 1x

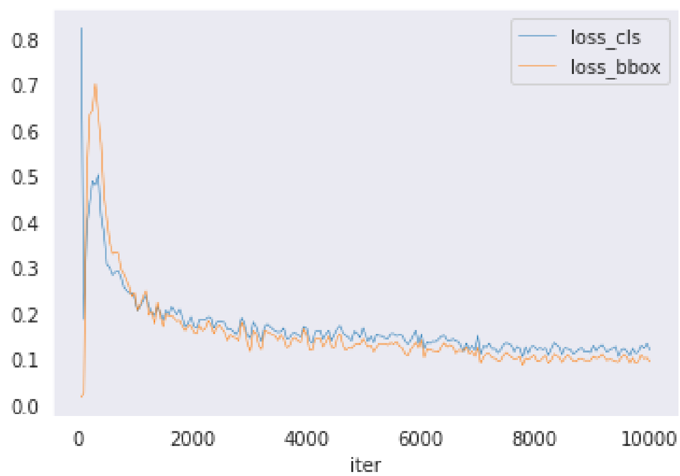

Fig. 2. Dynamic R-CNN with schedule 1x

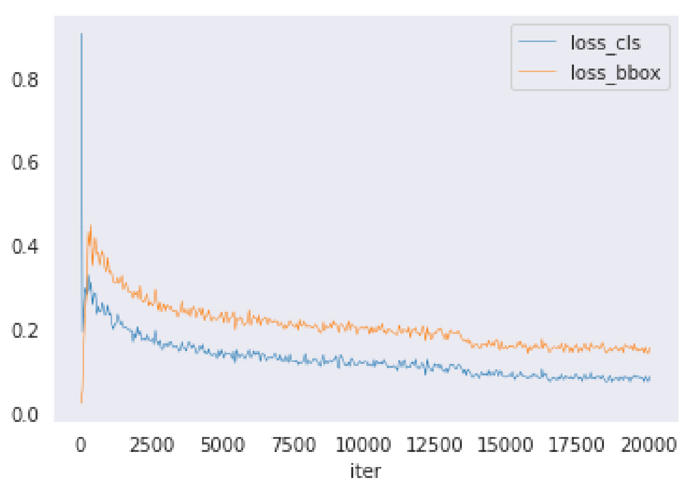

Fig. 3. Faster R-CNN with schedule 2x

V. Future Works

VI. Conclusion

Acknowledgment

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.