Visual and Textual Information Fusion Method for Chart Recognition

Chen Wang[1,2(✉)], Kaixu Cui[1,2], Suya Zhang[3], and Changliang Xu[1,2]

[1] XinHua ZhiYun Inc., Hangzhou, China
1804652205 4@163.com
[2] State Key Laboratory of Media Convergence Production Technology and Systems, Beijing, China
[3] State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, China

Abstract. In this report, we present our method in the ICPR 2020 Competition on Harvesting Raw Tables from Infographics, which is composed of Chart Classification, Text Detection/Recognition, Text Role Classification, Axis Analysis, Legend Analysis, Plot Element Detection/Classification and CSV Extraction. The image classification models of ResNet are adopt in Chart Classification. We adopted a two-stage based pipeline for end-to-end recognition, considering detection and recognition as two modules in Text Detection/Recognition. An ensemble model with LayoutLM and object detection model is adopted in Text Role Classification. A two-stage pipeline with two detection model is adopt in Legend Analysis. The final results are discussed.

Keyword: Chart recognition

1 Introduction

Charts are an effective data visualization tool often used to supplement textual content. They communicate information more efficiently and are common in the media, business documents, and scientific publications [1]. The goal of the ICPR 2020 Competition on Harvesting Raw Tables from Infographics is to provide common benchmarks and tools for the chart recognition community. This competition is composed of a series of sub-tasks (shown in Table 1) for chart data extraction, which when put together as a pipeline go from an input chart image to a CSV file representing the data used to create the chart. Two sets of datasets (Adobe Synth, UB PMC) are provided. The Synthetic Dataset is based on a large number of synthetic chart images (created with mat-plot-lib) with corresponding automatically derived annotations. The UB PMC Dataset is based on a smaller number of chart images extracted from Open-Access publications found in the PubMedCentral (PMC) [2].

In this report we present our solutions on task1, task2, task3 and task5. In the task1, an image classification model is adopted. In the task2, a two-stage based pipeline for

S. Zang—Intern at XinHua ZhiYun Inc.

A. Del Bimbo et al. (Eds.): ICPR 2020 Workshops, LNCS 12668, pp. 381–389, 2021.
https://doi.org/10.1007/978-3-030-68793-9_28

Table 1. Tasks in Competition on harvesting raw tables from infographics

Taks num	Task name
1	Chart classification
2	Text detection/Recognition
3	Text role classification
4	Axis analysis
5	Legend analysis
6	a Plot element detection/Classification
	b CSV extraction
7	End-to-End data extraction

end-to-end recognition is adopted. In the task3, Multi-modal technology is adopted. In the task5, a method based on object detection is adopted.

2 Tasks and Methods

In this section, we present our solutions for chart image classification, text detection and recognition, text role classification and legend analysis. The train datasets of these tasks are divided into train set (90%) and validation (10%). The models with best performance in validations are chosen and the final models fin-tune with both train and validation sets.

2.1 Task1. Chart Image Classification

In this task, model of ResNet [3] with different backbones were valid. According the validation results, the ResNet-50 was used for Synthetic dataset and ResNet152 was used for UB PMC dataset. Additionally, the ensemble learning was adopted for UB PMC dataset to enhance the generalization of the model. In training phase, ten ResNet-152 models were trained with different sub dataset of the whole train dataset. In inference phase, Average voting was used to ensemble the results from the ten models.

The models of ResNet-50 and ResNet-152 were the pre-trained model with ImageNet1000 dataset [4], and were fine-tune by the UB PMC and Synthetic datasets. The parameters of solver were shown as Table 2.

2.2 Task2. Text Detection and Recognition

For this task, we adopted a two-stage based pipeline for end-to-end recognition, considering detection and recognition as two modules. Our team adopted the Mask RCNN [6] based method to detect the line block of text (text with multiple lines considering one text block) in chart images and used the ResNeXt-152-FPN as the backbone network. Deformable convolutions [7] were used in the last three stages to enhance features. Cascade architecture [8] was used in the model to achieve higher detection accuracies. The

Table 2. Parameters of solver.

Parameters	Value
Max epoch	20
Batch size	128
Optimization	Momentum SGD momentum = 0.9 [5]
Learning rate	0.01
Normalization	*L2 regularization Weight = 3e5*

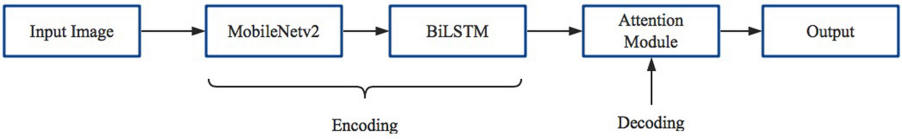

Fig. 1. Structure of recognition network

training dataset were UB PMC Dataset and Synthetic Dataset. For the recognition part, CNN and RNN with attention was adopted shown as follows (Fig. 1).

According to results of the first stage, the cropped images were achieved as input of the recognition model. The encoder first extracted a feature map from the input image with a stack of convolutional layers to enlarge the feature context, we employed a Bidirectional LSTM (BLSTM) network [9] over the feature sequence. The BLSTM network analyzed the feature sequence bidirectionally, capturing long-range dependencies in both directions. The attentional sequence-to-sequence module [10] was built to translate the feature sequence into a character sequence. In the training phase, firstly, the recognition model was train by training data including the publicly available datasets, i.e., ICDAR 2019-LSVT, ICDAR 2013 [11], ICDAR 2015, IIIT5K, ICDAR 2017, ICDAR 2017-MLT [12] English + Chinese, ICDAR 2017 RCTW-17 [13], ICPR 2018-MTWI, and synthetic images [14]. Then the final model fine-tuned with the UB PMC and Synthetic Datasets.

2.3 Task3. Text Role Classification

The text semantic roles of charts were related with both the visual and textual information. Therefore, only applying with object detection or text classification technology could not achieve good results. BERT-like models became the state-of-the-art techniques on several challenging NLP tasks, they usually leveraged text information only for any kind of inputs. The LayoutLM [15] was a new BERT-like pre-trained model, which both text and layout information were jointly learned in a single framework. However, it was not enough for some roles, i.e., Tick grouping, Legend Title, Legend Label, Data Marker Label, and Others.

The results were shown as Table 3. The LayoutLM produced perfect scores for the Synthetic Dataset but very poor scores for some roles in UB PMC Dataset. Therefore,

Table 3. Results of UB PMC and synthetic validation datasets.

Roles	F1 score in UB PMC		F1 score in synthetic
	LayoutLM	LayoutLM + Detection	
Chart title	0.746	0.815	1.0
Axis title	0.976	0.983	1.0
Tick label	0.984	0.992	1.0
Tick grouping	0.558	0.961	–
Legend title	0.214	0.889	1.0
Legend label	0.842	0.976	–
Value label	0.961	0.946	–
Data marker label	0.299	0.918	–
Other	0.594	0.866	–

we adopted object detection to enhance the results of UB PMC Dataset. The detection model was as same as the model for task2. According the results in Table 3 we chose 5 classes (Tick grouping, Legend Title, Legend Label, Data Marker Label, and Others) which reply more on visual information. The final flow chart was shown as follows (Fig. 2):

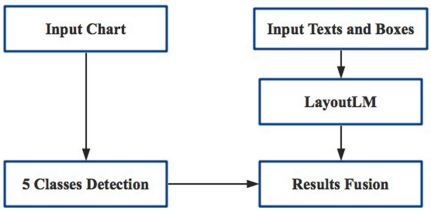

Fig. 2. Flow chart of task3

In the inference phase, we fused the results from the two models. The predicted role class from LayoutLM was changed to the predicted role class from detection model if the Intersection Over Union (IOU) between the predicted bounding box of detection model and the GT bounding box was over 0.5. The merging results were shown in Table 2. It was shown that the visual information made a great improvement on the task3. The confusion matrix of UB PMC validation dataset is shown as Fig. 3. In particular, legend title and mark label texts were more easily misclassified as text roles named other. The reason may be that the mark label and legend title text only appear on a few charts and the presented model was over-fit on these classes. The text roles named other were misclassified as tick label and value label texts. This may be due to the serious imbalance of category. In train dataset, the number of value label and tick label were much more than the other class, which makes model predict more value label and tick label. Some

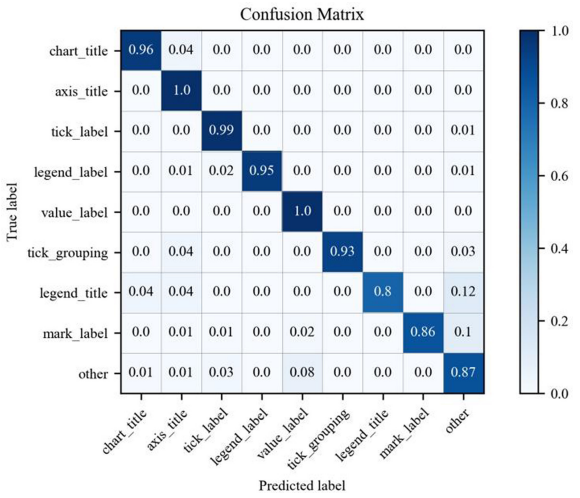

Fig. 3. Confusion matrix of UB PMC validation dataset

tricks for category imbalance (e.g. focal loss, positive sample) were adopted and only a small improvement achieved. It seems the key to obtain higher performance was to add more chart contains legend tile, mark label and other.

2.4 Task5. Legend Analysis

The sizes of Legend markers were mainly from 4 * 4 to 32 * 32, which were not suited for common object detection method. Increasing input image size was one way to address the small object detection, but in this task the sizes of legend markers were so small and increasing image size had little effect. Therefore, we adopted a two steps method to solve the issue, which was shown as follows (Fig. 4):

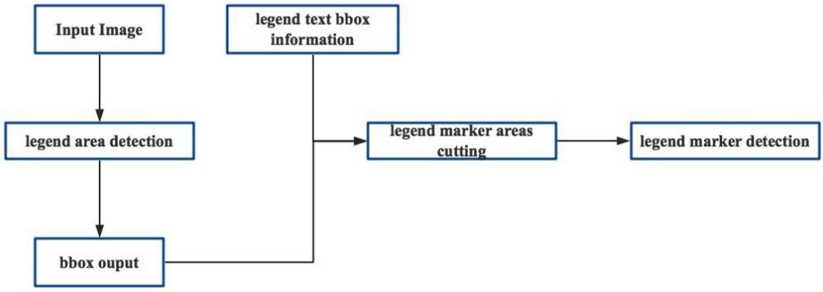

Fig. 4. Flow chart of task5

Firstly, the legend areas including all legend texts and markers were detected by a detection model whose config file was as same as detection model in task2. According

with the results of task3, the cutting image including text and corresponding marker can be easily figure out. Secondly, another detection model was apply to detect the bounding box of marker with the cutting image as input. The model of second detection is Cascade RCNN with ResNet-50 as backbone. In the training phase, the GT of legend areas were generated by annotations of Task5, specially, the maximum and minimum of the legend texts' and markers' coordinates were used to be the bounding boxes. For the second detection task of task5, the cutting images were generated shown as Fig. 5.

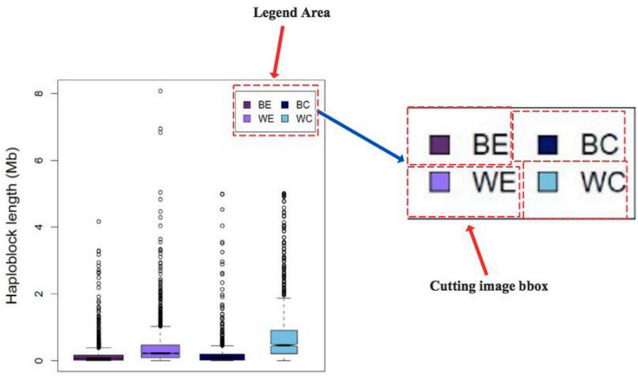

Fig. 5. Legend area and cutting images

We combining the UB PMC and Synthetic datasets for pre-training. Secondly, fine-tune the model on the corresponding dataset.

2.5 Final Results and Discussion

The Final results of the above tasks were shown as follow table (Table 4).

Table 4. Final results of the four tasks.

Task name	UB PMC	Synthetic
Chart classification	90.48%	99.44%
Text detection/Recognition	72.85%	92.56%
Text role classification	81.71%	99.93%
Legend analysis	92.00%	99.30%

The results show that the performance of the above method is very high on the Synthetic Dataset, but the scores of UB PMC left room for considerable improvement. Considering the much complex layout of real charts, our methods on the UB PMC dataset are also good solutions.

The confusion matrix of Chart Classification in UB PMC test dataset is shown as Fig. 6. In particular, area, scatter-line and vertical interval charts were more easily misclassified as line charts. One possible reason is that above misclassified charts have similar linear structure with line charts and the errors were magnified by category imbalance. A special classification model and data augmentation for the above classes may make sense.

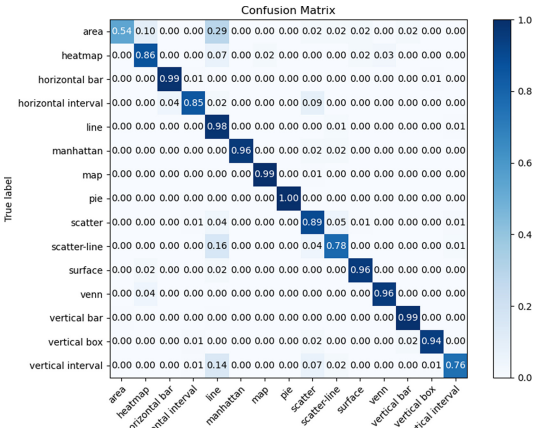

Fig. 6. Confusion matrix of chart classification in UB PMC test dataset

The confusion matrix of Text Role Classification in UB PMC test dataset is shown as Fig. 7. There is a few difference from the validation dataset. The scores of test dataset is obviously lower than the validation, especially for chart title, mark label and other. Chart title, mark label and value label texts were more easily misclassified as the class named other, the legend title and mark label were often misclassified as legend label. The reason may be that the presented models were over-fit on train dataset. A effective data augment and a larger train dataset may make sense.

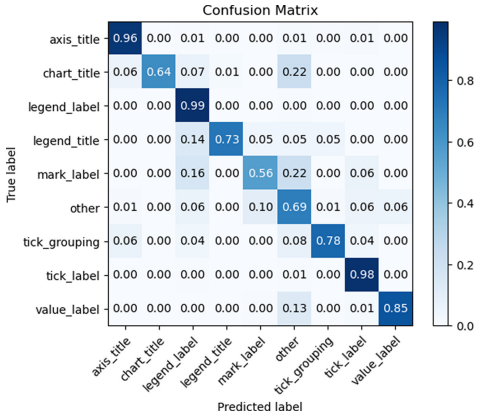

Fig. 7. Confusion matrix of text role classification in UB PMC test dataset

3 Conclusion

In this report we present the solutions on the four tasks (Chart Classification, Text Detection/Recognition, Text Role Classification, Legend Analysis). Our works show that the presented methods achieve good performance at both Synthetic and UB PMC datasets. Additionally, we discuss the drawbacks of the methods in UB PMC test datasets.

References

1. Davila, K., Kota, B.U., Setlur, S., et al.: ICDAR 2019 competition on harvesting raw tables from infographics (CHART-Infographics). In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1594–1599. IEEE (2019)
2. PMC Homepage. https://www.ncbi.nlm.nih.gov/pmc/
3. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
4. Deng, J., Dong, W., Socher, R., et al.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
5. Ruder, S.: An overview of gradient descent optimization algorithms. arXiv preprint arXiv: 1609.04747 (2016)
6. He, T., Tian, Z., Huang, W., Shen, C., Qiao, Y., Sun, C.: An end-to-end textspotter with explicit alignment and attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5020–5029 (2018)
7. Dai, J., et al.: Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 764–773 (2017)
8. Cai, Z., Vasconcelos, N.: Cascade r-cnn: delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6154–6162 (2018)
9. Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., Schmidhuber, J.: A novel connectionist system for unconstrained handwriting recognition. IEEE Trans. Pattern Anal. Mach. Intell. **31**(5), 855–868 (2009)

10. Mnih, V., Heess, N., Graves, A.: Recurrent models of visual attention. In: Advances in Neural Information Processing Systems, pp. 2204–2212 (2014)
11. Karatzas, D., et al.: ICDAR 2013 robust reading competition. In: Proceedings of ICDAR, pp. 1484–1493. IEEE (2013)
12. ICDAR 2017 competition on multilingual scene text detection and script identifification. https://rrc.cvc.uab.es/?ch=8&com=introduction, Accessed 16 Nov 2018
13. Shi, B., et al.: ICDAR2017 competition on reading chinese text in the wild (RCTW-17). In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 1429–1434.11. IEEE (2017)
14. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
15. Xu, Y., Li, M., Cui, L., et al.: Layoutlm: pre-training of text and layout for document image understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1192–1200 (2020)