ANALYSIS ON A MODIFIED PATH AGGREGATION NETWORK TOWARDS ACCURATE SCENE TEXT DETECTION

RAAKHAL RAPOLU[1]

[1]Dept. of Computer Science, Bennett University, Uttar Pradesh, India

---

ABSTRACT : In a scene text image, there are two types of information: visual texture and semantic meaning. Research on mining semantic information to aid text recognition has made tremendous progress in the previous several years but has garnered less attention; only RNN-like structures have studied to implicitly model semantic information. Multiple object recognition frameworks have been used to text detection in recent years, and they've had great success doing so. Text detection concerns have grown in prominence over time as the spectrum of application scenarios has expanded. When the text is distorted in any way, using a quadrilateral detection box in natural scenes becomes more difficult. Text identification may be difficult due to small targets and imbalanced data, however most networks have improved target sample balancing. The Progressive Scale Expansion Network (PSEN) is used in this study to develop an efficient scene text identification method (PSENet). A Mixed Pooling Module (MPM) employs multiple pooling methods to better extract text shape information at various distances. Using two extensions of ResNet, namely ResNeXt and Res2Net, for the backbone network improve feature extraction. It has been shown that our method has a greater accuracy than the original PSENet.

1. INTRODUCTION

Many visual tasks have been greatly improved by advances in deep learning. Progress in natural scene text detection is one of them. Image features may be extracted and text classifiers can be trained using traditional CNN networks CNNs have inspired a variety of other networks, including segmentation, regression, and end-to-end approaches. Using increasingly advanced algorithms, deep learning yields even more amazing results.

It's tough to recognise text in genuine settings due to a range of elements, such as the variety of text direction rotation and size ratio changes, the lighting of a real street or shopping mall scene, the inclined shooting angle, and the difficulties created by a change in text language. There is still fierce competition. A lack of performance gain from network topologies means that they are rarely used. As a rule of thumb, models with high outcomes tend to have a high number of parameters and huge models, while complicated systems take a long time to execute. There is still a considerable unmet need for algorithms that can be used in batch mode, and many of these algorithms remain in the research stage. A portable device's lightweight form and the demands of the application scenario necessitate that an application-based algorithm produce results that are theoretically accurate.

In the realm of machine vision, text detection is a crucial and tough task. Current text detection methods use predefined anchor frames to identify word candidate regions of high quality. Even though these treatments have been shown effective, they have the following drawbacks:

(1) The size, direction, and quantity of predefined anchor boxes have a significant impact on the outcomes of text detection.

(2) It is impossible to capture all occurrences of text using predefined anchor boxes because of the changing size, shape, and alignment of text in real environments.

(3) Using a high number of anchor boxes to improve text identification requires a tremendous amount of time-consuming calculations.

By employing Zhong's four scaling factors and six aspect ratios (resulting in 24 prior bounding boxes for each sliding point), DeepText improved its accuracy. As compared to Faster R-CNN, it has 2.6 times more than anchors.

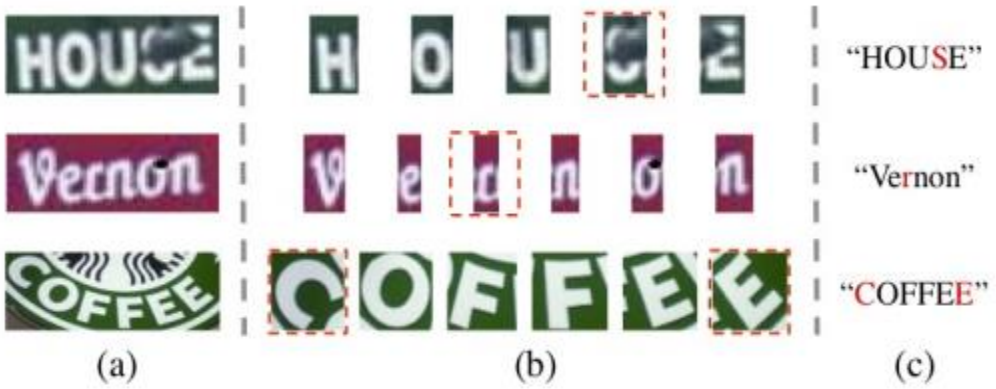

Figure 1. Examples of real-world text occurrences. (a) are some challenging images of text from a scene, (b) individual characters can be retrieved from a larger chunk of text. (a), and (c) are the semantic word contents that correspond to each other.

As part of recent scene text recognition research, more robust and effective visual information has been extracted by updating backbone networks and adding corrective modules to the system. Humans, on the other hand, are able to recognise scene text because of their ability to process visual information as well as their ability to interpret high-level text semantic context. If you merely look at the characters outlined in red dot boxes in Fig. 1, for example, it's nearly impossible to tell them apart based on visual cues alone. People are more likely to get the right answer if they take into account semantic context information.

2. LITERATURE REVIEW

Convolutional networks were employed by Wang and et.al., to integrate the text box's corner points, centre points and regression points.

It's possible to resolve document inclination to some extent, but the method's precision was hampered by the difficulties of determining the corners. Using text secondary prediction and link secondary prediction, The PixelLink algorithm was developed by Deng et al. to overcome the problem of neighbouring text portions being segmented.

When Zhang et al. were looking for candidate characters, they employed the maximum stable extremism area approach and split the characters into words or textlines according to the rules of the time.

Mask RCNN's power segmentation framework and context information were combined to recognise any shape text in a supervised context network created by Xie et.al..

In order to prevent erroneous sample detection, the method uses Text Context Module and ReScore Module. In order to overcome the challenge of curving text for the first time, Long et al. came up with the TextSnake semantic segmentation technique.

Multi-scale prediction was made possible by using Feature Pyramid Network (FPN) instead of segmentation-based approaches to construct feature maps of various scales. Adjacent text segmentation is also addressed by using the progressive scale expansion technique. A single maximum pooling operation meant that PSENet couldn't adapt to changing scene text lengths and record long and short text characteristics of the scene object in their original form.

PSENet differs from anchor-free approaches in a few ways. The FPN network divides the fusion properties of different scale outputs. The shrinkage method is used to reduce each text instance to a number of separate text segmentation maps of varying scales. In order to identify the final text, the breadth-first search technique mixes segmentation maps of different scales, aiming to recreate the whole text instance in order to identify the final text. To interpret text without an anchor, the progressive scale expansion algorithm can better distinguish scene text from close or stuck-together text.

3. METHODOLOGY

In order to provide accurate and effective text detection, a scene text detection system based on PSENet is recommended. In the end, enhanced scene text identification models resulted from the development of the MPM and backbone network. Figure 2 depicts the proposed scheme's flowchart.

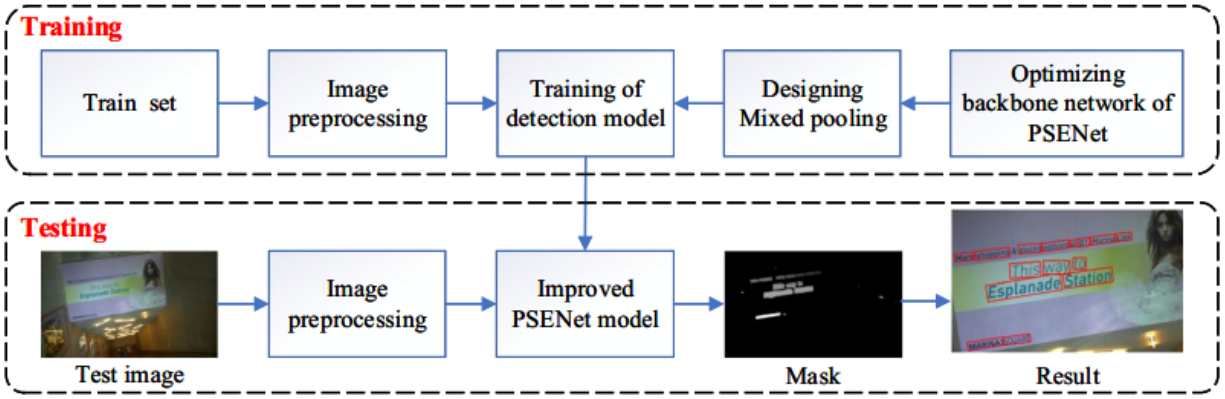

Fig. 2 Schematic diagram of the scheme

Figure 3a shows how ResNet is used to extract fundamental image features from the PSENet backbone network. ResNet's gradient fading disappearance problem was alleviated by introducing short connections, which at the same time helped to create deeper network topologies. Only four feature scales may be produced using various convolution algorithms in ResNet.

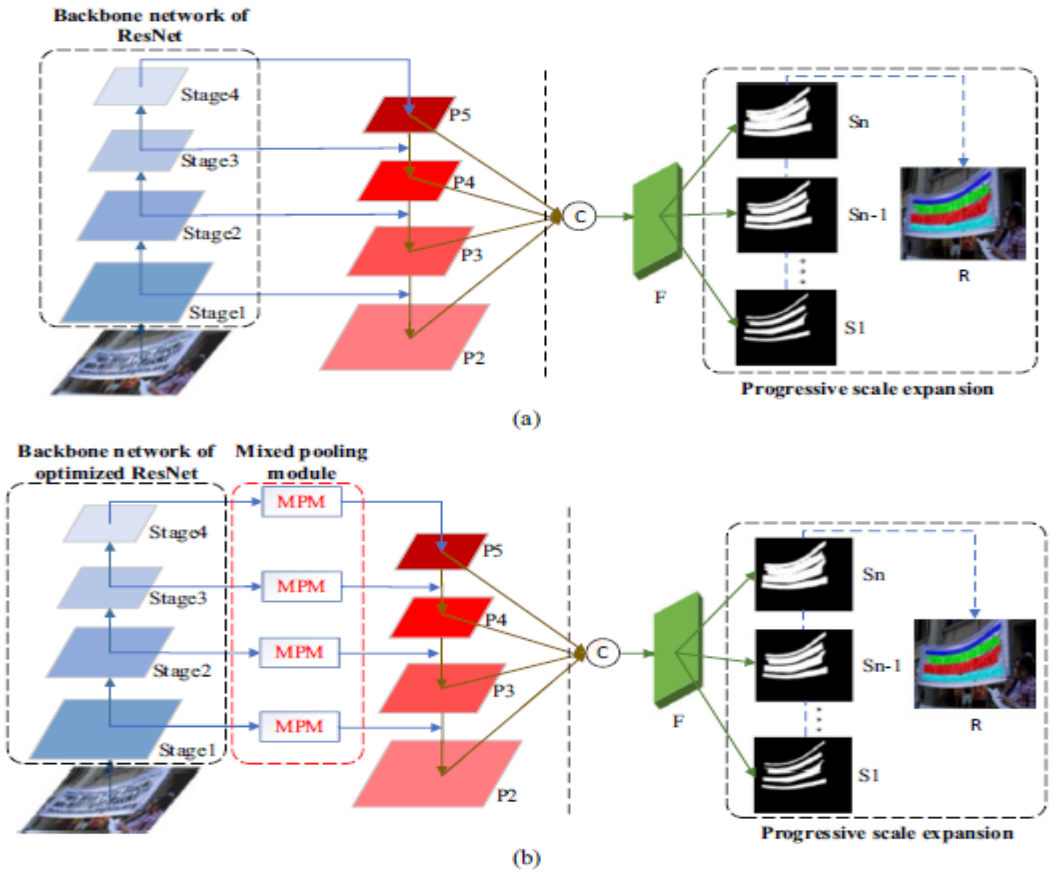

Fig. 3 Original PSENet and suggested scheme: a) the original PSENet; b) the proposed system.

Backbone networks must include multi-scale feature representations in order to locate objects of varying scales in the picture. The feature extraction efficiency shown in Fig. 3b is improved by combining ResNeXt and Res2Net in

the backbone network. MPM is a part of the backbone network, so it can measure the correlation of the distances between different points. It is also possible to better extract the shape of the scene text from backbone networks, hence increasing the accuracy of the model's text detection capabilities.

3.1 ResNet and FPN

The detector has two primary parts: the backbone network and the neck portion, which can be employed in either stage. Various target detection algorithms have also been developed and improved to better target certain locations. The convolution layer consists of nonlinear and downsampling layers. It is possible to describe images by tracing their characteristics across the entire receptive field. It is possible to improve performance by constantly expanding the network. It's more crucial to be able to understand features than it is to have a huge number of parameters. This difficulty is alleviated by the use of the BN layer in batch normalisation, which magnifies small changes in parameters. It has become the typical design for convolution networks because of its higher performance. As a result of ResNet, input and output have a direct link. The robust parameterized layer primarily focuses on learning the residual between the input and output in order to improve gradient explosion and gradient disappearance.

VGG, ResNet, and other detection systems form the basis of target detection. While the VGG16 backbone is first utilised in CTPN, SSD network also uses VGG16 as primary network. Feature extraction in Yang et altechnique .'s first used the ResNet-50 module, and most subsequent networks have adopted the ResNet series. Many good networks, such as DenseNet, have benefited from the backbone component. In order to achieve a strong training impact, DenseNet leverages feature reuse and bypass settings rather than increasing the number of network layers in ResNet or expanding the topology of the network in Inception. Overfitting is efficiently suppressed by using the bottleneck and translation layers, which minimises the network's parameters and makes it smaller. DenseNet is used as a backbone by numerous detectors for feature extraction.

3.2 Principle of the Method

PSENet, a text detection framework, is used to analyse ResNet and FPN in new directions without a single anchor in the process. The SENet and MPANet modules make up the bulk of the proposed structure. Remaining PANet uses the remaining structure of ResNet to process layers in the vicinity. To increase the effect of the network, MPAnet incorporates the characteristics of neighbouring layers into the network's graphs. Fig. 4 depicts the scene text identification algorithm's suggested architecture in great detail.

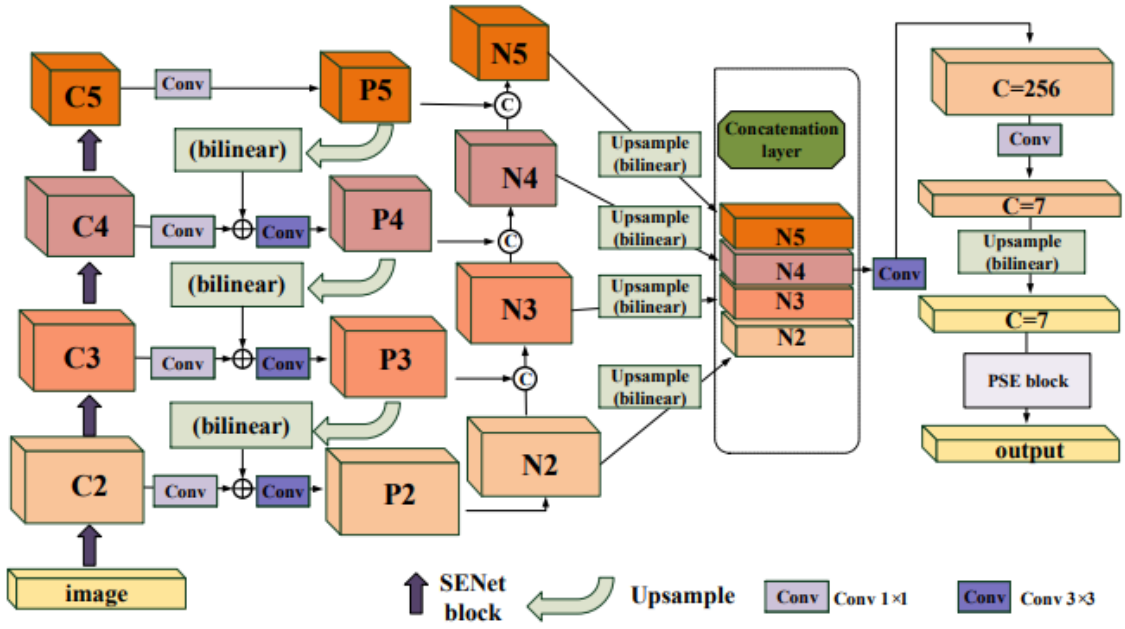

Figure 4. An illustration of our framework.

3.3 Backbone Network

As a backbone network, we combine hierarchical feature maps from ResNet50 stages 3, 4, and 5. To put it another way, ResNet50+FPN employs a 1/8 of the image feature map and has 512 channels. Non-local phenomena have prompted us to use our transformer unit to capture the global spatial interconnectedness more effectively. A feed-forward module, a location encoder, and a multi-head attention network make up this unit. Using an 8-head multi-head attention stack transformer unit, 2D feature maps are fed into two stack transformer units with 512 heads of feed-forward output.

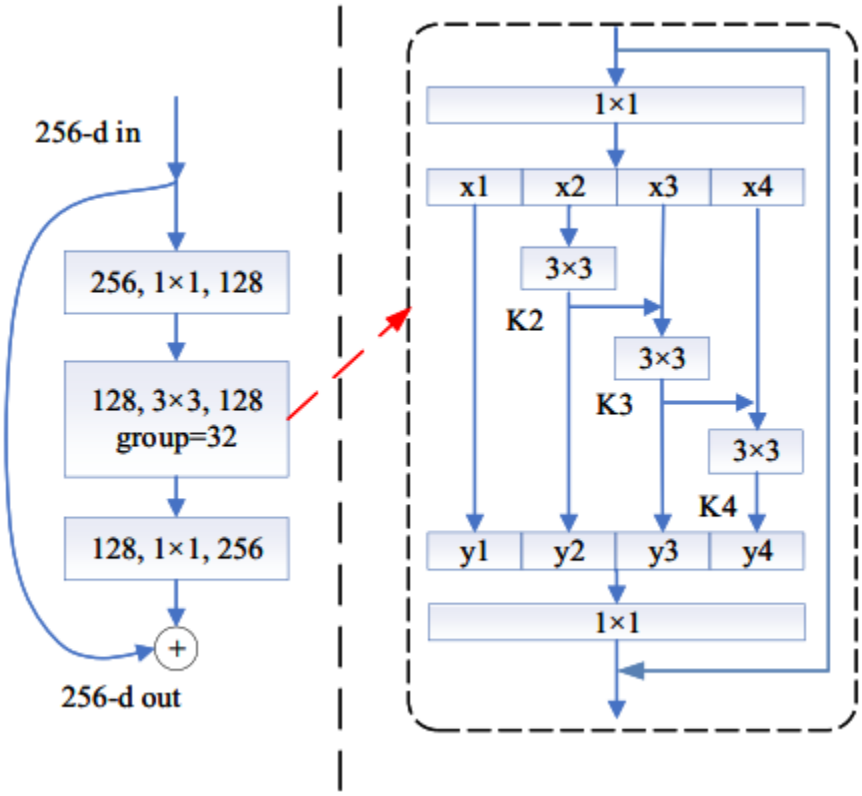

Fig. 5 The ResNet is optimized by combined ResNeXt and Res2Net (Variable cardinality C=32, scale dimension s=4)

ResNeXt and Res2Net, two residual network expansions, are combined to optimise ResNet (ResNet). There are two times fewer parameters in ResNeXt than in ResNet or Inception, thanks to a technique known as group convolution. As a result of the Res2Net, the capacity to extract multiscale attributes from a single residual block has been improved. There are a limited number of groups that can be formed in group convolution, and the variable cardinality is 32. One of the advantages of the optimised ResNet is that it can greatly reduce computing time by restricting the cardinality of the variables, resulting in higher operating speeds. In further experiments, the hypothesis is confirmed. The Res2Net concept adds a short residual block to each residual module in order to optimise the ResNet. Replacement for Res2Net's 3x3 attention mechanism module is the adoption of tiny groups of filters that connect distinct filter groups hierarchically. As a result, the model's performance is significantly improved by connecting filter groups in a similar residual stacking manner.

4. RESULTS AND DISCUSSION

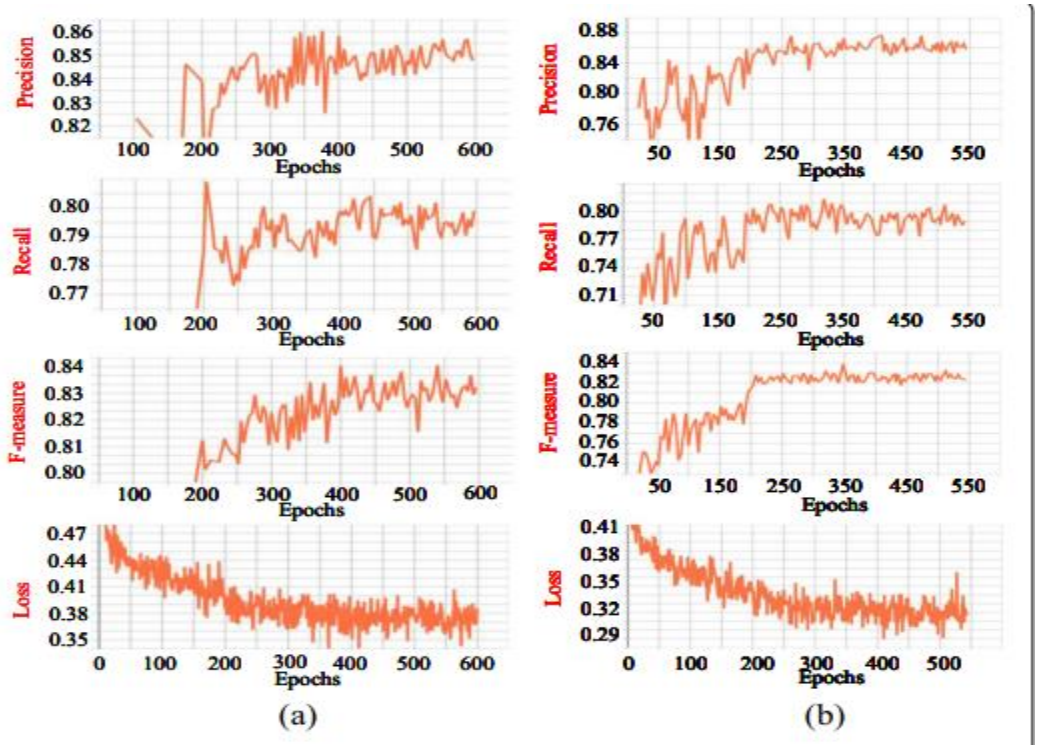

Fig. 6 Comparison of ResNet-152 with ResNet-MPM-50 performance: There are two different backbone networks in use: a for the ResNet-152 and b for the ResNet-MPM-50.

Fig. 6 depicts the results of testing ResNet-MPM-50 and ResNet-152 after optimizations. With 540 epochs, the ResNet-152 based model achieves a precision of 85.51 percent, an 80.69 percent recall rate, and an F-measure of 83.03 percent. Figure 6b shows the improved ResNet MPM-50-based model's recall of 80.79 percent, precision of 87.26 percent, F-measure of 84.0 percent, and loss of 0.3275 after 350 iterations. Based on our testing, we can conclude that our proposed scheme training model beats ResNet-152. Thus, the proposed scheme training model is smaller than ResNet-152's training model in terms of its scale.

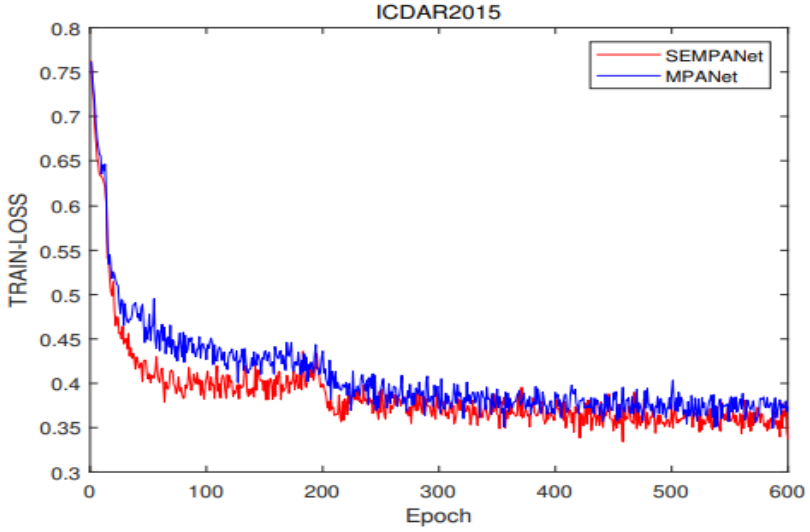

Figure 7. Ablation study of an SE block on ICDAR2015.

MPANet-trained ResNet 50 and SE blocks are used in these results.

SEMPANet (SEM) has a significantly lower train loss than MPAnet (MP) without a SE block (MPANet). SEMPANet's loss function declines quicker on ICDAR2015, as shown by this graph.

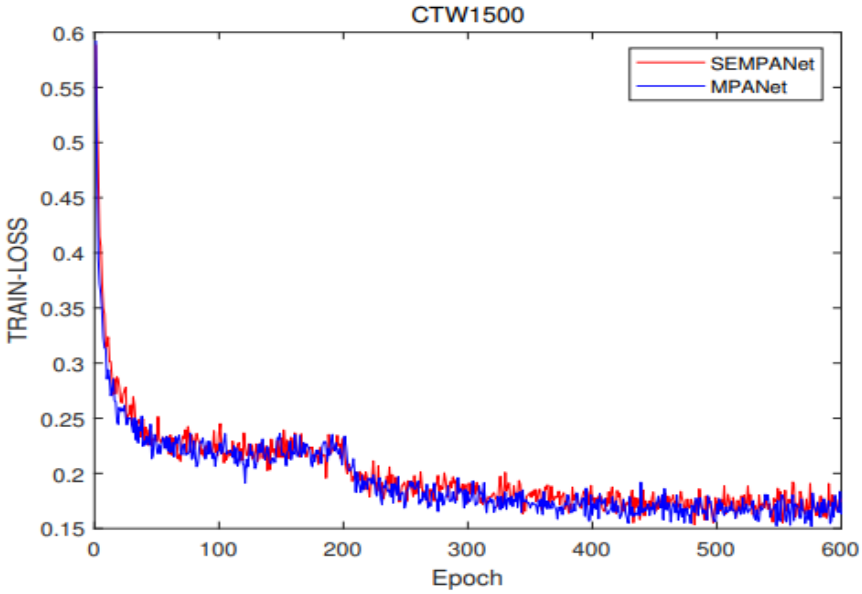

Figure 8. Ablation study of an SE block on CTW1500.

MPANet was used to train both (ResNet 50 and SE block) and (ResNet 50 block).

In general, the MPANet model's loss function has a faster impact on CTW1500's convergence than the CTW1500 model without a SE block.

Effects of MPANet

MPANet will be tested in a series of ablation studies on ICDAR2015 and CTW1500 datasets (see Table 1). The models are all trained using certified training images. Table 1 shows that MPANet's F-measure on ICDAR2015 and CTW1500 improved by 1.01 percent and 1.21 percent, respectively.

Table 1. MPANet's performance improvement, as measured by ICDAR2015 and CTW1500. The FPN network model was applied to ICDAR2015 and CTW1500 in PSE, yielding FPN * and FPN as the final results.

Method	Recall	Precision	F-Measure
FPN *	79.68	81.49	80.57
MPANet *	79.97	83.26	81.58
Gain *	0.29	1.77	1.01
FPN †	75.55	80.57	78.00
MPANet †	75.52	83.29	79.21
Gain †	−0.03	2.72	1.21

CONCLUSION

Scene text recognition methods have a PSENet-based solution in place to handle the problem of missing or incorrect detection. An image's text and object borders can be accurately detected using the Mixed Pooling Module, which captures the dependency between various text spots and obtains context information. Incorporating ResNeXt and Res2Net together considerably improves the network's ability to extract multi-scale feature extraction. In comparison to current scene text identification algorithms, the experimental findings reveal that the proposed technique has a lower missing detection rate and a larger detection precision. Specifically, the suggested technique improves the accuracy of the original PSENet by more than 5%. Mathematical tools for research and discussion are also a goal of mine. For multispectral images, a geometric algebra-based technique retrieves features that may be studied. Researchers could benefit from further study in areas like L1-norm reduction and hashing networks.

REFERENCES

1. Aljuaid, H.; Iftikhar, R.; Ahmad, S.; Asif, M.; Afzal, M.T. Important citation identification using sentiment analysis of in-text citations. Telemat. Inform. 2021, 56, 101492. [CrossRef]

2. Dashtipour, K.; Gogate, M.; Li, J.; Jiang, F.; Kong, B.; Hussain, A. A hybrid Persian sentiment analysis framework: Integrating dependency grammar based rules and deep neural networks. Neurocomputing 2020, 380, 1–10. [CrossRef]
3. Zhong, Z.; Jin, L.; Zhang, S.; Feng, Z. DeepText: A Unified Framework for Text Proposal Generation and Text Detection in Natural Images. arXiv 2016, arXiv:1605.07314.
4. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 39, 1137–1149.
5. X. Wang, K. Chen, Z. Huang, C. Yao, W. Liu, "Point linking network for object detection," arXiv preprint arXiv: 1706. 03646 (2017)
6. D. Deng, H. Liu, X. Li, D. Cai, "PixelLink: detecting scene text via instance segmentation," In The National Conference on Artifcial Intelligence (AAAI), 6773–6780 (2018)
7. Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, X. Bai, "Multi-oriented text detection with fully convolutional networks," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4159–4167 (2016). https:// doi.org/10.1109/CVPR.2016.451
8. X. Enze et al., "Scene text detection with supervised pyramid context network," In The National Conference on Artifcial Intelligence (AAAI), 9038–9045 (2019)
9. L. Shangbang et al., "TextSnake: a fexible representation for detecting text of arbitrary shapes," In European Conference on Computer Vision (ECCV), 20–36 (2018). https://doi.org/10.1007/978-3-030-01216-8_2
10. W. Wenhai et al., "Shape robust text detection with progressive scale expansion network," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 9336–9345 (2019). https://doi.org/10.1109/CVPR.2019.00956
11. Wang, W.; Xie, E.; Li, X.; Hou, W.; Lu, T.; Yu, G.; Shao, S. Shape robust text detection with progressive scale expansion network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 9336–9345.
12. Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep structured output learning for unconstrained text recognition. arXiv preprint arXiv:1412.5903, 2014.
13. Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. In NeurIPS, 2014.
14. Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. IJCV, 116(1):1–20, 2016.
15. Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In NeurIPS, pages 2017–2025, 2015.