2022 13th International Conference on Information and Communication Technology Convergence (ICTC) | 978-1-6654-9939-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICTC55196.2022.9952873

Improving Scene Text Recognition With A Combinative Image Augmentation Approach

Ngan-Linh Nguyen*,†, Gia-Huy Lam*,†, Hoang-Thong Vo*,†,
Trong-Hop Do*,†, Anh-Tien Tran‡, and Sungrae Cho‡

* Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City, Vietnam.
† Vietnam National University, Ho Chi Minh City, Vietnam.
‡ School of Computer Science and Engineering, Chung-Ang University, Seoul, South Korea.

***Abstract*—Scene text recognition plays an important role in various intelligent systems today such as robotic process automation and self-driving cars. These systems require knowledge of the surrounding scenery, where the words in the scene hold a lot of valuable information. For instance, scene text recognitions can serve the development of smart tourism, smart museums, and self-propelled robots. To increase the practical applicability of the solution, the recognition model needs to meet efficiently both accuracies and processing time. However, before constructing the model, most existing scene text recognition frameworks underrate the importance of a reliable and well-served augmentation stage, all the images in the dataset are often applied with common augmentation functions. In this paper, we present a combinative augmentation framework that takes random augmentation functions from different types into combinations for each individual image. In addition, the framework also randomizes the number of functions taken and every specific parameter for a particular function on the image. Thus, all these tweaks help to greatly increase the pattern diversity in images and pose an improvement in the model evaluation. Evaluated on both seen and unseen test datasets, the framework increased the accuracy by an average of 5,02% on the NRTR model and 2.36% on the VietOCR model.**

***Index Terms*—Scene text recognition, Combinative augmentation, Deep learning**

I. Introduction

It has been seen that the number of words in any language vocabulary is inherently a lot, especially the words on the street scenes, it is especially diverse in terms of appearance and semantics, many words on the street do not even appear in the lexicon of the language under consideration and appear or repeated very little in the whole dataset in particular. To increase the diversity of words but still ensure that this augmentation process keeps the sustainability and does not cause any problems for the learning of the model such as overfitting or underfitting, we specifically concern about creating a framework that can perform image augmentation diversely and randomly right from the way it works, making it compatible with many different data sizes from small to large.

Real-life applications of screen-recognition text have attracted a great deal of interest from the computer vision community, mentioned by Rowel Atienza in a paper introducing the STRAug [1] framework, a framework dedicated specifically to revolutionize the augmentation process in screen text recognition, a reliable augmentation framework eases the burden of finding labeled datasets large enough and publicly available or the collections of automatically annotated synthetic text images for training such as MJSynth [2] or Synth90k [3], SynthText [4], Verisimilar [5], and UnrealTex [6]. In addition to the work of improving the performance of text recognition, it is also too heavy and harmonious if we keep focusing on improved model architecture or learning algorithms, that is when we need to touch on an aspect that is not much exploited in screen text recognition - image data augmentation.

Deep learning STR models [7] [8] [9] [10] [11] [12] have supplanted the performance of algorithms with handcrafted features [13] [14]. By constructing this framework, the purpose of our research is to contribute to the existing street scene text image augmentation methods a better and more efficient way for the STR models to better understand the language, especially some complicated languages with all the unexpected appearances. Our research is implemented on the VinText dataset [15], which have its text labels in Vietnamese - a complicated language indeed since the language's syllables are formed by a combination of the initials, finals, and tones, we then work in corporate with recognition models including SRN [16], NRTR [7] and SAR [17], VietOCR and the framework augmentation that we introduce. The STRAug library plays an important role in our framework because it provides augmentation functions that are already classified into groups and each function can be fully customizable based on the parameter input. Further framework details could be found in later sections of this paper.

II. Related work

Learn to augment [18] by Luo et al. was proposed to generate more proper training images for the recognition model by using a set of custom fiducial points, the proposed method is flexible and controllable. Furthermore, the authors bridge the gap between the isolated processes of data augmentation and network optimization by joint learning. The research has shown improvement on ICDAR2015 (IC15) [19], SVT Perspective (SVTP) [20] and CUTE80 (CT) [21]. The disadvantage of Learn to Augment is that it is meant to be a fixing framework that only focused on distorted text, one of

the many causes of data distribution shift in STR, and requires an additional agent and augmentation networks that must be trained with the main STR network. Since this framework works inside the learning stage, this results in a more complex setup, additional 1.5M network parameters, difficult to reuse algorithm, and a very long training time.

Rowel Atienza proposed STRAug which mainly focused on comprehensive study and empirical evaluation on different data augmentation functions, which is also developed from their previous reseacrh [22]. STRAug provides 36 data augmentation functions, creating 8 logical groups, analyzing the effect of each group, and systematically combining all groups to maximize their overall positive impact. However both "individual group performance" and "combined group performance" describes how the author evaluated each group or each multi-group to decide and rank the best groups performance for a specific model thus the chosen functions that were already randomized will stay the same for the whole training process, for every image. Our approach is to create a more randomized framework to the work on the dataset generation level, which applies different combined functions on each image and makes the whole dataset more diverse in terms of language appearance.

III. Proposed system

The overall pipeline of the proposed framework is depicted in Fig. 1. The original data set first goes through the augmentation stage that we would discuss further in Experiment procedure and settings section. Both the augmented and original are trained on the models and get evaluated and compared in results of accuracy [23].

To build our framework for street scene text augmentation, there must be a core library dedicated to providing augmentation functions. There are countless libraries available that can be implemented in our framework, however, our main purpose in building this framework is to be able to experiment with a large number of functions/ function groups to make the dataset as diverse as possible, that is why STRAug is chosen among all existing libraries, not to mention its reliability and the capability to generate decent quality augmentations.

STRAug is designed exclusively to work with STR, offers up to 36 augmentation functions in 8 groups, each group represents a different purpose of transforming an image and each function supports 3 levels or magnitudes of severity or intensity.

IV. Experiments

A. Dataset

VinText is a street scene text image data set. This is a challenging data, containing busy and chaotic scenes with scene text instances of various types, appearance, sizes, and orientations. Each text instance is annotated with a quadrilateral bounding box and and word-level transcription. This data set can be a good benchmark for measuring the applicability and robustness of scene text spotting algorithms or even augmentation solutions like ours. Figure 2 shows some examples of images from VinText dataset.

B. Experiment procedure and settings

Figure 3 shows some images in the detection stage. This stage does not directly concerns the research of this paper which are augmentation and recognition models so we won't go through in details.

In the augmentation stage, by default, our framework generates nine augmented images from an original image. For each augmented image, one to three out of eight function groups are randomly selected. One function will be randomly selected from each selected group. Any parameters mentioned early are configurable. For example, we took an image from the VinText dataset, which has the label "theo" on it. Figure 4 shows the original image. We then randomly select one to three function groups to be applied on this image. The selected function groups are "blur", "warp", and "weather". One function in each group will be randomly selected to aply on the image. The results are that GaussianBlur from "blur", Distort from "warp", and Frost "weather" are the three exact functions to be apllied on the original image. Figure 5 shows the image that has been applied to 3 functions from different groups.

We learned that using functions from different groups is more efficient than letting the combination have any functions that come from the same group. Functions from the same group usually have similar deformative effects which when combined, are somehow useless or may make the image overly deformed due to their conflicts in effect. Furthermore, we also let the numbers of functions applied to be randomized every time, for example, 1 to 3, since we want the difficulty of the augmented image to be diverse from easy to difficult, not fixed on a specific number of functions for all image in the dataset.

Figure 6 shows augmented images generated from one original image, any function combinations within one image range are made sure not to be repeated. Notice how the sixth image and ninth image share the same Curve function but have different angles of curving, which explains the "magnitude" parameter from the function is naturally randomized if not specified, this detail from STRAug makes our framework even more diverse in augmenting capability.

In the recognition stage, we implemented our augmented dataset with the experiment settings for the three models on the https://github.com/PaddlePaddle/PaddleOCRPaddleOCR framework, including SRN, NRTR, SAR as following: lowering label, using Cosine learning rate schedule with learning rate is 0.001, L2 regularizer with factor is 0.00004, number of epochs is 30, warm up epoch is 2, space char is True. In addition, https://github.com/pbcquoc/ vietocrVietOCR model is also used in the paper with default settings (only lowering label)

C. Result and discussion

Figure 7 shows some augmented images that are generated by our framework from the original images. These images are

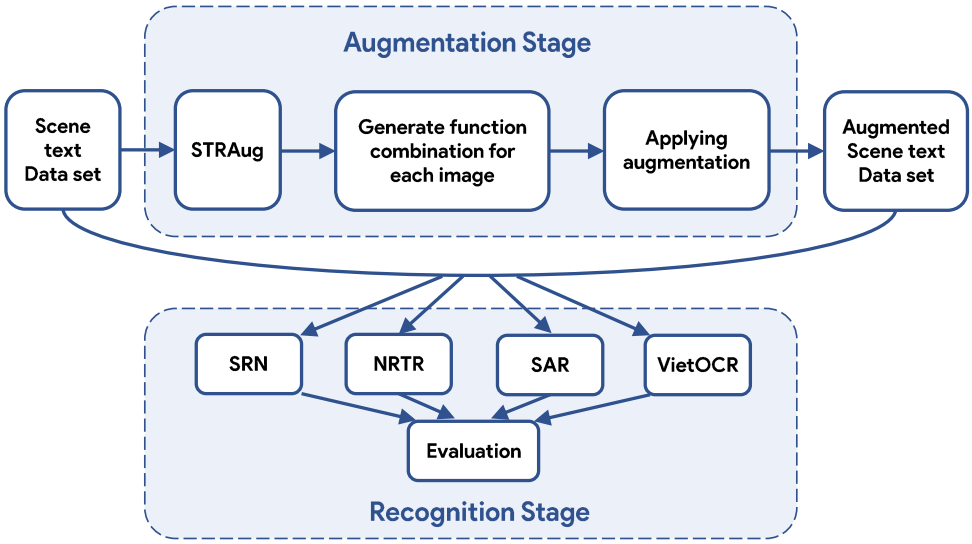

Fig. 1: Overall pipeline of proposed framework.

Fig. 2: Examples of images from VinText dataset

ready to be trained on the four mentioned models in order to compare the improvement in efficiency.

Fig. 7: Some generated augmented samples ready for training.

In this section, we show the efficiency of the SOTA models on the two experiment scenarios, including using the augmentation framework and not using the framework. To compare

Fig. 3: Example images with bounding boxes from detection stage.

the efficiency, accuracy is used as metric of the models.

TABLE I: Experiment results on unseen test image dataset.

Model	Original	Augmentation
SRN	0.8112	0.8095
NRTR	0.7234	**0.7736**
SAR	0.7922	0.7935
VietOCR	0.7938	**0.8174**

Following Table 1, using our framework for training doesn't help much to improve the efficiency of models except the NRTR model (the accuracy increases from 72.34% to 77.36%) and VietOCR model (the accuracy increases from 79.38% to 81.74%). In the most of experienced models, VietOCR whose performance is the best with 81.74% accuracy. Thus,

Fig. 4: Original image

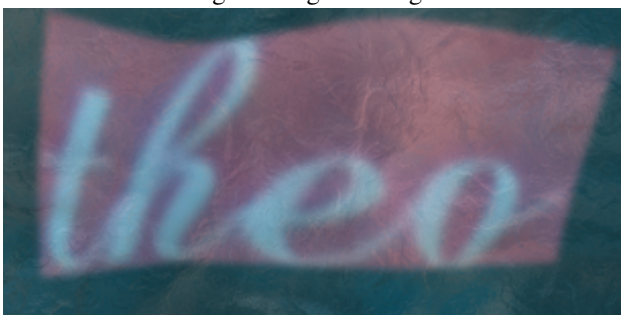

Fig. 5: Augmented image with Distort, GaussianBlur, Frost applied.

Fig. 6: Example of augmented images from the original images

depending on the model architecture that enhances partially supported data in increasing model performance .

V. Conclusion

Optical character recognition problems still face many challenges, especially in daily life. Tough challenges such as busy and disorganized scenes with scene text instances of diverse types, appearance, sizes, and directions. In this study, we propose a new system to enhance the performance by flexibly augmenting scene text images. In addition, we also contribute experiments on state-of-the-art word recognition models on the benchmark data set. With our proposed system with strong recognition ability, they will extract the information to understand the surrounding scenery. Applying in the fields of smart tourism, smart museums, self-driving cars, self-propelled robots and bringing practical effects to the social community, production and business.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2022-00156353, Development of End-to-End 8 Ultra-Communication and Networking Technologies).

References

[1] R. Atienza, "Data augmentation for scene text recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1561–1570.

[2] J. Weinman, Z. Chen, B. Gafford, N. Gifford, A. Lamsal, and L. Niehus-Staab, "Deep neural networks for text detection and recognition in historical maps," in *Proc. IAPR International Conference on Document Analysis and Recognition*, Sep. 2019.

[3] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *International Journal of Computer Vision*, vol. 116, no. 1, pp. 1–20, jan 2016.

[4] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[5] F. Zhan, S. Lu, and C. Xue, "Verisimilar image synthesis for accurate detection and recognition of texts in scenes," *CoRR*, vol. abs/1807.03021, 2018. [Online]. Available: http://arxiv.org/abs/1807.03021

[6] S. Long and C. Yao, "Unrealtext: Synthesizing realistic scene text images from the unreal world," *arXiv preprint arXiv:2003.10608*, 2020.

[7] D. Yu, X. Li, C. Zhang, J. Han, J. Liu, and E. Ding, "Towards accurate scene text recognition with semantic reasoning networks," *CoRR*, vol. abs/2003.12294, 2020. [Online]. Available: https://arxiv.org/abs/2003.12294

[8] T. Zheng, Z. Chen, S. Fang, H. Xie, and Y. Jiang, "Cdistnet: Perceiving multi-domain character distance for robust text recognition," *CoRR*, vol. abs/2111.11011, 2021. [Online]. Available: https://arxiv.org/abs/2111.11011

[9] V. Loginov, "Why you should try the real data for the scene text recognition," *CoRR*, vol. abs/2107.13938, 2021. [Online]. Available: https://arxiv.org/abs/2107.13938

[10] M. Cui, W. Wang, J. Zhang, and L. Wang, "Representation and correlation enhanced encoder-decoder framework for scene text recognition," *CoRR*, vol. abs/2106.06960, 2021. [Online]. Available: https://arxiv.org/abs/2106.06960

[11] J. Lee, S. Park, J. Baek, S. J. Oh, S. Kim, and H. Lee, "On recognizing texts of arbitrary shapes with 2d self-attention," *CoRR*, vol. abs/1910.04396, 2019. [Online]. Available: http://arxiv.org/abs/1910.04396

[12] T. Wang, Y. Zhu, L. Jin, C. Luo, X. Chen, Y. Wu, Q. Wang, and M. Cai, "Decoupled attention network for text recognition," *CoRR*, vol. abs/1912.10205, 2019. [Online]. Available: http://arxiv.org/abs/1912.10205

[13] L. Neumann and J. Matas, "Real-time scene text localization and recognition," vol. 38, 06 2012, pp. 3538–3545.

[14] C. Yao, X. Bai, and W. Liu, "A unified framework for multioriented text detection and recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4737–4749, 2014.

[15] Y. He, C. Chen, J. Zhang, J. Liu, F. He, C. Wang, and B. Du, "Visual semantics allow for textual reasoning better in scene text recognition," *CoRR*, vol. abs/2112.12916, 2021. [Online]. Available: https://arxiv.org/abs/2112.12916

[16] F. Sheng, Z. Chen, and B. Xu, "NRTR: A no-recurrence sequence-to-sequence model for scene text recognition," *CoRR*, vol. abs/1806.00926, 2018. [Online]. Available: http://arxiv.org/abs/1806.00926

[17] H. Li, P. Wang, C. Shen, and G. Zhang, "Show, attend and read: A simple and strong baseline for irregular text recognition," *CoRR*, vol. abs/1811.00751, 2018. [Online]. Available: http://arxiv.org/abs/1811.00751

[18] C. Luo, Y. Zhu, L. Jin, and Y. Wang, "Learn to augment: Joint data augmentation and network optimization for text recognition," *CoRR*, vol. abs/2003.06606, 2020. [Online]. Available: https://arxiv.org/abs/2003.06606

[19] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny, "Icdar 2015 competition on robust reading," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, pp. 1156–1160.

[20] T. Q. Phan, P. Shivakumara, S. Tian, and C. L. Tan, "Recognizing text with perspective distortion in natural scenes," in *2013 IEEE International Conference on Computer Vision*, 2013, pp. 569–576.

[21] A. Risnumawan, P. Shivakumara, C. S. Chan, and C. L. Tan, "A robust arbitrary text detection system for natural scene images," *Expert Systems with Applications*, vol. 41, p. 8027–8048, 12 2014.

[22] R. Atienza, "Vision transformer for fast and efficient scene text recognition," *CoRR*, vol. abs/2105.08582, 2021. [Online]. Available: https://arxiv.org/abs/2105.08582

[23] J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S. J. Oh, and H. Lee, "What is wrong with scene text recognition model comparisons? dataset and model analysis," *CoRR*, vol. abs/1904.01906, 2019. [Online]. Available: http://arxiv.org/abs/1904.01906