Beyond additive noise: Evaluating Style Geometric Style Transfer in Scene Text Recognition

Syrine Noamen[1] and Antoine Munier[1]

[1] *Department of Computer Science, EPFL Lausanne Switzerland*

***Abstract*—Even with the rapid evolution of computer vision, Scene Text Recognition (STR) stands as a crucial yet challenging task due to the intricate variability of text in natural scenes which makes most traditional Optical Chracracter Recognition methods fail in STR tasks. The primary obstacle for STR models lies in the limited access to extensive, labeled real-world datasets, leading to reliance on synthetic data for training which creates a gap between the conditions of training data and the unpredictable complexities encountered in real-world. Our goal is to study data augmentation approaches to help the robustness of STR models.**

I. Introduction

Scene Text Recognition is the task of detecting textual information from images of natural scenes.The process includes detecting, localizing and recognizing text within a given image and then converting it into a text. In the field of STR, the choice of training datasets is crucial, as it significantly influences the success of model training and subsequent evaluation. A primary challenge in STR is the scarcity of diverse, real-world labeled datasets. This limitation hinders the development of robust deep learning models, as annotation is both time-consuming and costly. Typically, STR models are trained on synthetic data and tested on real-world images, leading to difficulties in adapting to new, real-world distortions during testing.

To address this, our research focuses on data augmentation as a strategy to enhance the diversity of training images, thereby developing more robust STR models. Our study begins by picking a state-of-the-art model and selecting real-world datasets that effectively measure model robustness. Our methodology involves a thorough assessment of various data augmentations and their impact on the model, followed by training on an augmented dataset.

This approach is particularly relevant in the evolving landscape of STR, where nowadays the trend is towards multimodal methods that leverage both visual and textual features. Despite these advancements, foundational techniques like data augmentation remain indispensable for enhancing model performance. Originally, we considered employing Generative Adversarial Networks (GANs) for applying distortions and style transfer to generate a dataset for STR, but first, we need to conduct a comprehensive evaluation of direct data augmentation methods to establish a baseline understanding of their efficacy.

II. Methodology

A. Evolution of STR models

Scene Text Recognition (STR) encapsulate a modular approach across various stages of development. For feature extraction, models transitioned from the foundational VGG, used to process the whole image to create a feature map and RCNN (Region-based CNN) to ResNet architectures, enabling deeper learning. BiLSTM networks have been pivotal in sequence modeling, adeptly managing the order and context of characters in text sequences. For the prediction there was a shift from Connectional Temporal loss (CTC), which independently decodes character probabilities, to attention mechanisms that focuses on relevant image segments. The latest leap in STR has been the integration of language-aware models. that add linguistic information from language models, blending visual cues with the predictive power of language context. As an example, the the Multi-Granularity Prediction for Scene Text Recognition from Hugging Face that achieves an average accuracy of 93.35% on standard benchmarks (1).

B. State Of The Art model

Initially, we started with Recurrent Convolutional Neural Network (RCNN) as the backbone with VGG style, using Connectionist Temporal Classification (CTC) for decoding. CTC independently decodes character probabilities. However the implementation is extremely slow, which is why we moved to another model. While it is larger than CRNN, it achieves higher accuracy due to its use of ResNet and the attention mechanism.

1) Description: The TPS-ResNet-BiLSTM-Attn (TRBA) model consists of four components: **TPS**, Thin Plate Spline is for spatially transforming distorted or curved text, **ResNet** is for deep feature extraction, **BiLSTM** is for sequence modeling in both

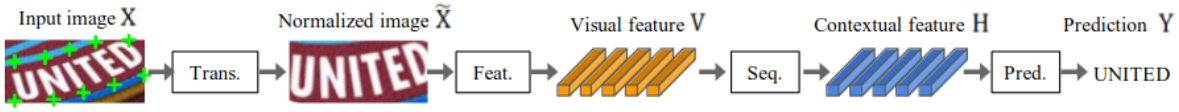

Figure 1: Visualization of the stages (2)

directions, and an **Attention** Mechanism for focused prediction.

The model works through stages: TPS normalizes text shapes and geometry making it easier for next stages, ResNet processes images into visual feature maps that represent a distinguishable receptive filed, BiLSTM preserves contextual information in both directions of sequences, and the attention mechanism enhances character-level prediction accuracy and enables learning the dependencies between characters. It helps the model focus on the most relevant features at each step of the sequence 1.

To evaluate the model under evaluations, we have selected 5 evaluation datasets or benchmarks.

Type	Dataset	# Images	Characteristics
Regular	IIIT5K	3,000	Sourced from Google image searches: billboards, signboards, etc.
	Street View Text (SVT)	647	Outdoor images from Google Street View, including noisy, blurry, or low-res images.
Irregular	SVT Perspective (SVTP)	645	Images with perspective distortions from Google Street View.
	CUTE80 (CT)	288	Cropped images from natural scenes, focusing on curved text.
	ICDAR2015 (IC15)	2,077	Captured via Google Glasses, featuring noisy, blurry, and rotated images.

Table I: Overview of Real-world Datasets Used for Evaluation

2) Regular vs. Irregular Datasets: Regular datasets typically consist of text that appears in well-aligned, horizontally placed lines with clear spacing and standard fonts. Irregular datasets, on the other hand, contain text that may be curved, distorted, occluded, or presented in unusual layouts or artistic fonts.

C. Augmentation techniques

Type	Applications	Description
Blur	GaussianBlur	Soft blur for out-of-focus effect.
	MotionBlur	Simulates blur from movement or shake.
Noise	GaussNoise	Adds Gaussian noise for graininess.
	MultiplicativeNoise	Random pixel value alteration.
Camera-related	RandomBrightnessContrast	Adjusts brightness and contrast.
	ImageCompression	Applies JPEG-like compression.
Dropout	PixelDropout	Randomly removes pixels.
	RandomToneCurve	Alters tone curve for color balance.
Artistic Effects	Posterize	Reduces color levels for a poster effect.
	Solarize	Inverts colors based on threshold.
	Equalize	Balances histogram for enhanced contrast.
	Emboss	Creates a raised or 3D effect.
Geometric	Affine (Zoom)	Scales image for zoom-in effect.
	ElasticTransform	Distortion for wavy effect.
	Perspective	Changes viewpoint for 3D look.

Table II: Image Augmentation Techniques

These distortions are kept "minor" because of the fragility of any STR model. We carefully designed transformations that are not too aggressive and that don't create visual artifacts, but simulate natural distortions in images taken from the phone camera or street surveillance cameras. The text should remain legible and not challenging for the inference.

D. Impact of the augmentations on the performance

Dataset	Drop in Accuracy Individually (%)							Comb. Drop (%)
	Blur	Noise	Camera	Dropout	Process	Affine	Geometric	
IIIT5k	5.167	0.167	0.234	1.700	3.100	-0.300	7.500	2.734
SVT	3.474	1.001	0.538	2.238	2.392	2.856	7.802	2.574
CT	0.697	0.348	-0.348	0.348	1.742	6.272	3.136	4.53
SVTP	16.744	0.775	-0.155	4.031	4.806	3.876	11.628	5.891
IC15	10.562	1.405	0.364	4.058	3.330	1.561	9.001	4.682

Table III: Summary of Drop in Accuracy Across Transformations, Comb. is the combination of the transformations applied randomly each time

a) Interpretation: Different real-world datasets used in STR, each has unique characteristics which can lead to varied responses to the transformations.

- IIT5k and CT are dominated by clean text and simple background. Even with the curved text that appear in the natural scenes, the TPS normalizes the curved character and aligns it horizontally as shown in the example image from CT in figure 1
- SVT and SVTP are both outdoor images of business logos, but SVTP has variety of viewpoints and orientations, blur and geometric transformations can only make it more difficult.
- IC15 has images that are noisy, blurry, rotated, and with low resolution.
- Noise has a weaker effect because during the generation of the training dataset MJ the synthetic engine adds some noises like Gaussian noise and applies JPEG compression. So the model is more able to adapt to noise.
- The minor accuracy gains resulting from camera and zoom effects, can show how the model can benefit from light processing of the testing inputs.

III. Training on Augmented Dataset

The model is trained on two commonly used synthetic scene text datasets, MJSynth(MJ) and SynthText (ST), and for the validation the union of the training sets IC13, IC15, IIIT, and SVT as the validation data are used. After applying the augmentations as

shown in Table II randomly on each image sample and training the TRBA from scratch. As the baseline for our study, the model was trained on two commonly used synthetic scene text datasets: MJSynth (MJ) and SynthText (ST). For validation purposes, we used a unified set comprising the training subsets of IC13, IC15, IIIT, and SVT. We have applied a series of data augmentations, as detailed in Table II, to each image sample. These augmentations were implemented randomly, with each application choosing a function from that group to simulate a variety of real-world conditions. The TRBA model was then trained from scratch using these augmented datasets to evaluate the impact of these modifications on its performance. Please refer to the appendix, Detailed information about the datasets used can be found in Table VI and the training conditions in Table VII.

Notes about the Training: Due to time constraints, our training was limited to 41,000 iterations, requiring approximately 70 hours with the computing resources at our disposal: 2 GPU Tesla V100s 32GB. The model, encompasses 49×10^6 trainable parameters. Meanwhile, the baseline model in Baek et al.'s study (2) underwent 300,000 iterations. Our training achieved accuracy of **78.2%**, compared to the **84.1%** accuracy reported for the baseline. This limitation in training time limits our ability to fully assess the effectiveness of the data augmentation techniques employed. The reduced number of iterations may have constrained the model's learning capability, and impacting the overall evaluation of the augmentation's efficiency in enhancing robustness. Our first trial of training was not successful because we used a subsample of the training data in the first training trial and the performance got affected. It shows how large data is a requirement for STR. Therefore, we had to train again using the whole sample.

We validated every 500 iterations and we saved the model that achieves the best accuracy which was obtained in iteration number 39,000, See figure 2.

A. Results

The normalized Edit Distance is a standard metric that measures the minimum number of insertions, deletions, and substitutions required to change the predicted text into the correct text, normalized by the length of the ground truth.

Dataset	Accuracy (%)		Acc. Change (%)	Edit Distance		Norm. Ed. Dist. Change
	Clean	Aug.		Clean	Aug.	
IIIT5k	84.467	82.633	1.834	0.947	0.936	0.011
SVT	84.389	81.917	2.472	0.945	0.936	0.009
CT	66.551	64.808	1.743	0.835	0.826	0.009
SVTP	73.953	65.9	8.053	0.902	0.86	0.042
IC15	67.417	65.365	2.052	0.861	0.849	0.012

Table IV: Comparison of Accuracy and Edit Distance on Clean and Augmented Data using our trained model

The accuracy of our trained model exhibits a minimal discrepancy between the clean and augmented data sets, unlike the baseline. Our trained model has achieved *total accuracy* **75.356%** on the original datasets and **71.288%** on the augmented datasets compared to the baseline that had **81%** and **77.146%** respectively.

B. Conclusion

Training on augmented data significantly enhances the model's robustness against real-life variations. This improvement is evident when comparing the changes in accuracy between the baseline model and our trained model on both clean and augmented datasets taking into account the difference in training accuracy. It is also revealed that the model has high variance when using different seeds. This discrepancy suggests that while the model has not fully captured all patterns present in the training data yet, it exhibits an ability to generalize well to the unseen data in the testing set. Extending the training periods would have provided more insights into the model's learning dynamics and potential. Despite not capturing all the patterns in the training data, it is still adequate enough for the patterns present in the testing/evaluating data and is capable to generalize.

C. Summary

Dataset	Original Accuracy (%)	Obtained Accuracy (%)
IIIT5k	87.367	84.467
SVT	87.400	84.398
CT	74.216	66.551
SVTP	80.155	73.953
IC15	75.390	67.417

Table V: Comparison of baseline and our trained model

The table shows that our model has learned patterns in some of the real-world datasets.

IV. Limitations

Data augmentation technique is more complicated for STR than other tasks in Computer Vision. The application of such augmentations must be judicious to avoid introducing misinterpretations of characters. For instance, blurring, if not applied carefully, can lead to character ambiguity, causing a model to mistake a 'q' for an 'o', or a 'y' for a 'v'. (3) reports that in STR the peak of accuracy gain is between N = 2 and N = 4, where N is the number of augmentations. For TRBA specifically, the relative accuracy gain from the baseline is about 1% for N around 3. (4) states that TRBA model has benefited from 0.9% gain accuracy relative to the baseline, which can showproof the limited capacity of this type of model.

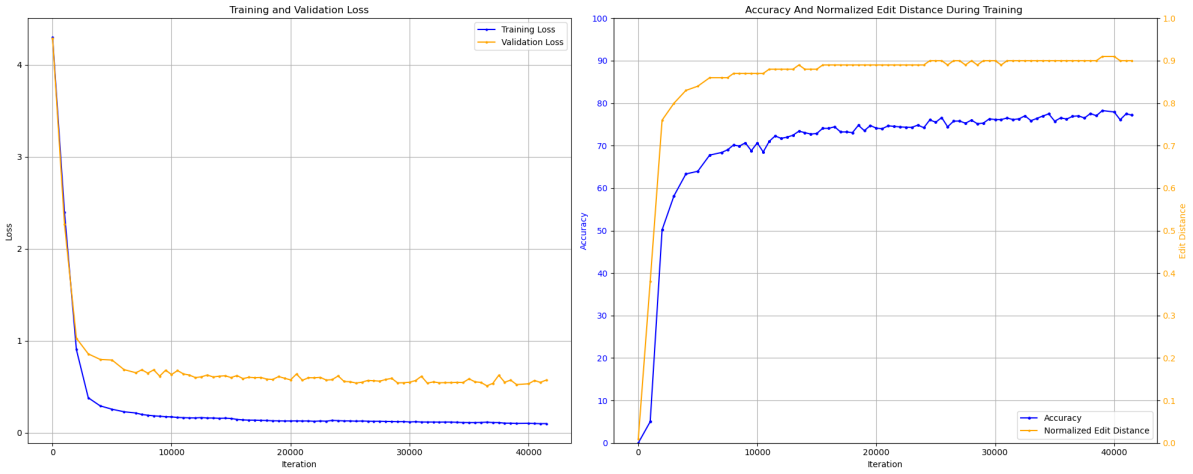

Figure 2: Visualization of Training and Validation Loss (left) alongside Accuracy and Normalized Edit Distance (right) over training iterations.

Methodical approaches of data augmentation are essential for fortifying the STR models. For instance, Mu et al. propose a nuanced Random Blur Region technique (5) that segments a small image patch into squared units, applying localized blurring to each. This method retains the clarity of character shapes while effectively simulating environmental variations (5). Moreover, although relatively few studies have explored data augmentation in depth for STR, the work by Atienza (3), finds that the optimal accuracy gains are achieved with a moderate number of augmentations, specifically between 2 and 4. In the case of the TRBA model, a relative accuracy gain of 0.9% from the baseline is observed when around three augmentations are applied (3). Additionally, Baek et al. report a 0.9% accuracy improvement for the TRBA model compared to the baseline, suggesting a potential limit to the model's capacity to learn the real-world irregularities with their various characteristics. (4).The beneficial augmentations were of type Blur and Crop that resembles our affine transformation.

V. Further Experiments

STR models encounter a variety of challenges. There remains a range of alternative strategies and methodologies that could potentially enhance the robustness.

A. Fine-tuning on real-world data

we could have studied transfer learning with simple fine tuning. Training the model on augmented synthetic data then fine tuning on real-datasets to enhance the model's generalization ability. It first gains the basic learning from the synthetic then get introduced to complexities and variety of natural scenes. But this should be done carefully as the labeled real-world data is scarce and the distribution of the trained subsets during fine tuning should match the distribution of the testing subset, to not worsen the performance of the model.

B. Generative models

GAN can help in generating large datasets that can effectively simulate the complexity and variability of real-world scenes. The generated dataset would have the advantage of containing automatically labeled and realistic images that cover a broad spectrum of scenarios from various real-life applications with a wide range of text styles, fonts, and distortions. solving the scarcity of labeled real-world datasets for STR. The labels are inherent in the generation process, eliminating the need for manual annotation or the use of of seperate models to isolate text, unlike the semi- and self-supervised learning approaches applied to unlabeled real-world datasets that require a text detector component to identify and crop word regions from scene images (4).

C. Augmentative positive learning pair

Like in self-supervised learning, we can show the model different versions of the same text image (positive pairs), where each version might have variations in font, size, orientation, viewpoints, background.. This way it learns the similarity between the pair regardless of the visual alterations.

VI. Technical challenges

The cluster was equipped with CUDA version 10, which needd the installation of older binaries of PyTorch due to the drop of support. This version mismatch extended to other libraries as well, as some required versions were not compatible with the older PyTorch installation. So we seeked out for alternative libraries that were independent of both Torch and Torchvision...

ACKNOWLEDGMENTS

We would like to express our gratitude to Igor Krawczuk, whose expertise, and quick responsiveness guided us invaluably and the Laboratory for Information and Inference Systems – LIONS for providing us with the computing resources.

VII. ETHICAL RISK ASSESSMENT:

Scene text recognition models introduce ethical risks, and one significant concern lies in their potential misuse for resolving captchas, leading to a denial of service on certain websites. In this scenario, the stakeholders impacted include website owners, administrators, and users relying on the security of captchas for various online activities. The negative impact involves a compromised security mechanism, rendering captchas ineffective against automated attacks. The severity of this risk is substantial, as it opens avenues for automated bots to engage in malicious activities, such as spamming, account takeover, or other forms of cyberattacks. The likelihood of occurrence is notable, given the prevalence of automated tools seeking to bypass captchas. Given the intertwined nature of our efforts to enhance the Scene Text Recognition (STR) model with captcha resolution, attempting a complete disassociation from captcha-solving seems impossible. Consequently, modern captchas use complex puzzles, image recognition challenges, and behavioral analysis to differentiate between humans and bots, to be less vulnerable to the visual aspects of text recognition models. This strategic shift acknowledges the technology's limitations while striving to strike a balance between user-friendliness and robust security in online environments. The model could be used for surveillance purposes, impacting vulnerable populations and infringing on individual privacy rights. There is also bias in the Data, most real-world datasets use English language and a few are in Chinese.

APPENDIX

A. Training

Dataset	Validation Images
IIIT5K-Words (IIIT)	2000
Street View Text (SVT)	257
ICDAR2003 (IC03)	1156
ICDAR2013 (IC13)	848
ICDAR2015 (IC15)	4468

Table VI: Overview of Validation Datasets

B. Use cases

In figures 3, the last image was created by DALL.E generative model. (1) the multimodal model for the inference.

Train dataset:	50%MJ+50%ST (8.9M + 5M)
Total batch size:	384
Iterations:	40,000
Parameter init:	He
Optimizer:	Adadelta
Learning rate:	1.0
Adadelta ρ:	0.95
Adadelta ϵ:	$1e^{-8}$
Gradient clipping:	5.0
Image size:	100×32
Channels:	1 (grayscale)

Table VII: Training Conditions

Figure 3: Diverse scene text images with "epfl" written as text, predicted by the baseline model

Figure 4: Example to show how STR models can benefit from learning distortions applied on the AI generated image and infer using the 2022 MGP-STR model

REFERENCES

[1] P. Wang, C. Da, and C. Yao, "Multi-granularity prediction for scene text recognition," 2022.

[2] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "What is wrong with scene text recognition model comparisons? dataset and model analysis," *arXiv preprint arXiv:1904.01906*, 2019.

[3] R. Atienza, "Data augmentation for scene text recognition," pp. 1561–1570, October 2021.

[4] J. Baek, Y. Matsui, and K. Aizawa, "What if we only use real datasets for scene text recognition? toward scene text recognition with fewer labels," 2021.

[5] D. Mu, W. Sun, G. Xu, and W. Li, "Random blur data augmentation for scene text recognition," *IEEE Access*, vol. 9, pp. 136 636–136 646, 2021.