

Attacking Sequential Learning Models with Style Transfer Based Adversarial Examples

Zhilu Zhang¹, Xi Yang¹, Kaizhu Huang^{1*}

¹School of Advanced Technology, Xi'an Jiaotong-Liverpool University

E-mail: *kaizhu.huang@xjtlu.edu.cn

Abstract. In the field of deep neural network security, it has been recently found that non-sequential networks are vulnerable to adversarial examples. There are however few studies to investigate the adversarial attack on sequential tasks. To this end, in this paper, we propose a novel method to generate adversarial examples for sequential tasks. Specifically, an image style transfer method is used to generate for a Scene Text Recognition (STR) network adversarial examples, which are only different from the original image on the style. While they will not interfere with the recognition of image information by human vision, the adversarial examples would significantly mislead the recognition results of sequential networks. Moreover, based on a black-box attack, both in digital and physical environments, we show that the proposed method can use cross text shape information and attack successfully the TPS-ResNet-BiLSTM-Attention (TRBA) and Convolutional Recurrent Neural Network (CRNN) models. Finally, we demonstrate further that physical adversarial examples can easily mislead commercial recognition algorithms, e.g. iFLYTEK and Youdao, suggesting that STR models are also highly vulnerable to attacks from adversarial examples.

1. Introduction







The security issues in machine learning have a long history. As one of the most successful machine learning paradigms [1], deep learning has achieved great success in speech recognition [2], image classification [3], and auto driving [4]. Recently, many investigations have revealed the vulnerability of deep learning models by manipulating the inputs with imperceptible adversarial perturbations called adversarial examples [5, 6, 7].

Current research on adversarial examples mainly focuses on non-sequential tasks, for example image classification and semantic segmentation [8]. There are little research efforts on digital attacks of STR model [9]. The sequential STR model is a variable length sequence, while previous adversarial attacks mainly target a single tag; this makes it more difficult than non-sequential tasks. As shown in Table 1, when a carefully crafted disturbance is added to a stop sign, similar to rust, it does not feel strange in human vision. However, the modified stop sign succeeds in enabling the self-driving car to recognize the stop sign as another object with a high confidence. This kind of attack will bring great security risks. Similarly, cameras are ubiquitous in today's environment, which means that the information around us is at risk of being leaked at any time. Especially for some text messages that need to be kept secret, it is important to protect effectively the security of this information. While considering attacks digitally based on gradient optimization, the traditional methods and the existing STR model attack method [9] generate very small perturbations and may not lead to effective physical attacks. Therefore, it

is interesting to investigate a physical method that does not interfere with human vision but could also attack successfully physical sequential targets.

In this paper, we design an effective adversarial attack method particularly on sequential learning models i.e. STR models, TRBA and CRNN. Different from the traditional adversarial attacks, which are not easily noticeable with a lower attack intensity, we engage the style transfer mechanism [10] that can greatly modify the style of the original image while retaining the original content information.

Table 1: Adversarial examples crafted by style transfer.

Original image	Style	Adversarial example	Prediction
			stop → sore
			open → ooen

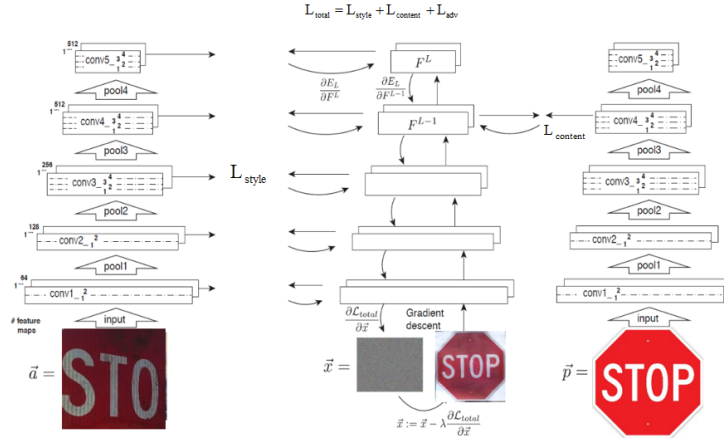


Figure 1: Adversarial example generate process.

Figure. 1 introduces the process to generate adversarial examples using the proposed method. More specifically, the original image and style image is fed into a VGG16 network to extract content features and style features respectively. Combining target network adversarial loss to generate adversarial examples, we can increase significantly the attack loss to bring great attack performance. Importantly, the generated adversarial examples can achieve both digital attacks and physical attacks since the adversarial intensity is large enough. We also conduct a migration test and compare TRBA and CRNN under different scenarios. The results show that the adversarial samples generated based on the high-accuracy model can easily attack the CRNN model with low accuracy, and the confrontation generated based on CRNN examples, it is difficult to successfully attack the TRBA model. Furthermore, we try to cross the shapes of the two characters to destroy the STR model’s ability to perceive the characters. The results show that this method is more effective in physical environment attacks. It also successfully attacks commercial STR models (e.g. iFLYTEK, and Youdao). Our key contributions in this paper are summarized in the following:

- We develop a new approach to generate adversarial examples by crossing the shape features of multiple characters that would affect largely the STR model. The attack results show that sequential networks are also highly vulnerable to adversarial attacks.
- With an excellent transferability, the proposed method can be successfully applied in attacking AI models under both digital and physical environments. Particularly, the generated scene text adversarial examples will easily fool the systems while accepted perceptibly by human vision.
- We successfully applied our method on the STR models including those famous commercial STR models.

2. Method

2.1. Model Structure

The framework of the proposed method is composed of a feature extraction network (VGG16) and a target network (sequential learning models). Many studies have proved that the shallow layers of VGG16 network can extract spatial features, and the deep layers can extract semantic features. Inspired by this conclusion, the original image and style image are respectively fed into VGG16 network to extract style information and content information from different layers from the network. As shown in Figure. 1, the network on the left is used to extract style features, and that on the right is used to extract content features. Then, the adversarial examples are generated by attacking the target network. Concretely, the first adversarial example is generated by randomly crafted adversarial noise. The adversarial example is input to the target network and VGG16 network to get adversarial loss, content loss, and style loss. Finally, adversarial examples will be continuously optimized in the direction of reducing all loss functions.

2.2. Loss Functions

Given test images $x \in R^m$ with labels y , y could be the class label or a sequence of labels. A DNN model is used to map images to a discrete label set. Adversarial examples x' are craft to let $F(x') \neq y, x' = x + \delta$, where δ represents adversarial perturbations generated by x' . Then, adversarial example x' for x craft process could change to optimize Eq. 1 as follows:

$$\min_{x'} L(y_x, y_{x'}) + \lambda D(x, x'), \quad s.t. x' = x + \delta. \quad (1)$$

L represents the attack loss function which changes with the target network change. λ is a hyperparameter which is used to balances style transfer and adversarial intensity. If λ increased, the generated adversarial example tends to reduce attack intensity to reduce the distance D between the adversarial image and the original image. We combine the image style transfer mechanisms to meet the purpose of disguise, while the adversarial attack method satisfies the attack effect. In addition, the use of a content smoothing loss makes the adversarial example smooth locally. The entire loss function consists of four parts: style loss L_s , smoothness loss L_m , adversarial loss L_{adv} , and content loss L_c .

Style loss and content loss both propose by [10]. The style loss can effectively reflect the difference between the style of the adversarial example and the original image. The content loss can show the difference in text information between the original image and the adversarial example. Reducing style loss and content loss can ensure that the adversarial example retains the text information while incorporating a specific style. We leverage the Eq.2 as the Style loss, Eq.3 as the content loss:

$$G_{ij}^l = \sum_k F_{ik}^l F_{jk}^l; \quad E_l = \frac{1}{4N_l^2 M_l^2} \sum_{i,j} (G_{ij}^l - A_{ij}^l)^2; \quad L_s(x, x') = \sum_{l=0}^L W_l E_l. \quad (2)$$

$$L_c(x, x') = \frac{1}{2} \sum_{i,j} (F_{ij}^l - F_{ij}'^l)^2. \quad (3)$$

G_{ij}^l denotes the pixel feature relationship of generated noise of the l -th layer, $A_{ij}^{l,2}$ represents the style image's style feature value, and E_l represents single layer style loss, where N represents the filter number of each layer and M represents each feature size. F_{ij}^l represents the characteristics of the layer where the image is located. The high-level features will capture the content information of the image, such as the object and position, but not the pixel level of the image. Thus, in this paper, we will capture the highest-level information to construct the content loss.

The image smoothness can be converted to reducing the degree of the change between near pixels. For the adversarial sample, the smoothness loss could be defined as:

$$L_m(x, x') = \sum (x'_{i,j} - x_{i+1,j})^2 + (x'_{i,j} - x_{i,j+1})^2)^{\frac{1}{2}}, \quad (4)$$

where $x'_{i,j}$ is the pixel at coordinate (i, j) of image x' .

Same as traditional attack methods, style transfer adversarial loss is also based on the gradient. The adversarial loss can invert the output directly, Eq. 5 represents the output of the attention-based STR model. $f(\cdot)$ represent RNN, LSTM or Bi-LSTM cell, the probability $P(l_t = i | x, l_1, \dots, l_{t-1})$ of the t -th character l_t base on hidden state h_{t-1} and previous character L_{T-1} . Finally, attention based STR model loss could be defined as follows:

$$P_t = \text{softmax}(f(h_{t-1}, L_{T-1})); \quad P(l_t = i) = \frac{\exp(P_t^i)}{\sum \exp(P_t^i)}; \quad (5)$$

$$L_{adv}(x, l) = -\log P(l|x) = -\sum_{t=2}^T \log P(l_t = i | x, l_1, \dots, l_{t-1}).$$

3. Experiment

3.1. Experimental Setup

CRNN model (base on CTC) and TRBA model (base on attention) [11] are selected as the target attack networks and two models are pre-trained base on MJSynth [12] and SynthText [13] separately. We compared the invisibility with amazon's adversarial examples method. The PyTorch toolbox is used to implement our attack algorithm and conduct experiment on the DIDI Cloud platform with P4 GPU.

3.2. Style Transfer Adversarial Examples

3.2.1. Attack Setting In order to verify the effectiveness of the adversarial attack, we designed an attack experiment to compare the attack success rate under different adversarial intensity parameters. Considering the iteration situation, the adversarial intensity parameter is set to 0.1. By trial and error, we found that if the value of the adversarial intensity parameter is large, the generated image will be distorted before the style generated, as shown in Figure. 2. Correspondingly, both the content loss parameter and style loss parameter are set according to the optimal situation of [10], which are 0.025 and 5, respectively. Besides this, we set epoch=100 and save a total of 20 adversarial examples for every 5 images generated. As shown in Figure. 3, which is a complete adversarial sample generation result of available images. Starting from the first generated adversarial example, the attack is successful, but the style transfer effect has not been completed at this time. Generally, after the fourth image is generated, the adversarial examples will have the specific style characteristics from the style image and can successfully attack the network.

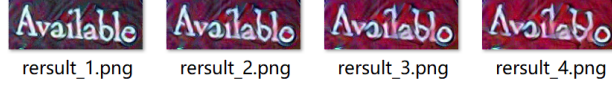


Figure 2: Too high adversarial intensity leads to failure of style transfer.

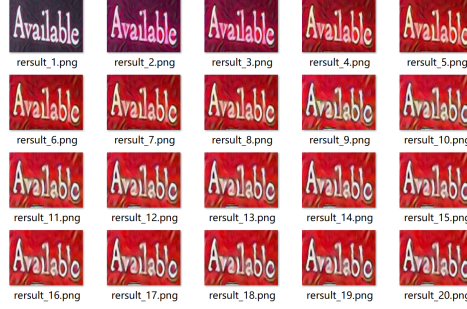
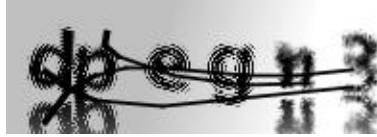


Figure 3: All adversarial examples of one sample.



(a) Adversarial example with style tranfer.



(b) Amazon adversarial example.

Figure 4: Change words texture result.

3.2.2. Attack Result In the attack experiment, the original images are come from the cute80 database, while the first picture in the database is used as the style image. The attack results are shown in Table 2. The adversarial examples generated by the proposed method have a great style transfer performance, while they remain the content information from original images. It is clear that both TRBA and human vision can correctly recognize all original test images, but only 5.26% adversarial examples can be correctly recognized by TRBA. As for human vision, even choose iter=10 adversarial examples, it still have better performance than TRBA on Amazon adversarial examples. Different from the method that makes the noise prevent undetected by looking for the smallest perturbations, usually, the STR model can only misrecognize one or two words. However, generating the adversarial examples by the style transfer could have great attack strength, such as the result shown in Figure. 4, the text information in the original picture is ‘samsung’, and the generated adversarial samples have been identified as ‘samsunie’, which completely mislead the sequential learning models even for the amount of text information has been changed. Table 2 also compares STR model and human vision accuracy of recognition of adversarial examples generated by the style transfer method in iter5 and iter10 and Amazon adversarial examples. Comparing the results, it can be clearly found that the style transfer based method can effectively attack the STR model without affecting human visual recognition. While Amazon’s adversarial sample misleads the STR model, it also has a huge interference with human visual recognition.

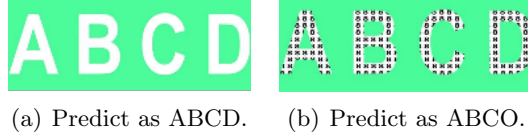


Figure 5: Change words texture result.

3.2.3. Transfer Attack In this part, we further verified the performance of style transfer based attacks by using cross text shape information to attack different networks. Specifically, we use the adversarial examples generated on the CRNN model to input to the TRBA model for transfer attack. The results are shown in Table 3, the iter=1 adversarial example can successfully attack the CRNN model with 11.76% correct rate, while the TRBA model still has 82.3% correct rate. It can be seen that TRBA is more robust than the CRNN model. Otherwise, when we using the adversarial examples generated by the CRNN model to attack the TRBA model, most of the adversarial examples fail to attack successfully. However, the adversarial examples generated on the TRBA model can attack the CRNN model with a high correct rate. It can also be explained that for the STR model, the adversarial examples generated by the model with a higher recognition rate are also have high transferability.

Table 2: Adversarial example attack result

correct rate	Style transfer(iter=5)	Style transfer(iter=10)	Amazon
TRBA	5.26%	5.26%	42.10%
Human vision	100%	84.21%	52.63%

Table 3: Transfer attack result


Adversarial Examples	TRBA	CRNN
	university	universis
correct rate	82.3%	11.76%

Table 4: Physical adversarial example attack result.

Image	TRBA	CRNN	iFLYTEK	Youdao
(a)	ABCD	ABCD	ABCD	ABCD
(b)	ABCO	ABCO	ERROR	ERROR

3.3. Physical-world Attacks

In the last setting, we have carried out physical attack experiments on TRBA and CRNN networks. The generated adversarial examples are printed and then converted into digital images which are input into the STR network. However, compared with the non-sequential network, the physical attack effect of the sequential network is not particularly great, especially for the text images without curvature. This phenomenon may be due to the fact that the STR network has better attention to the shape. The above Experimental generated adversarial examples are mainly attack the area outside the target. Therefore, we test attack the texture only inside the text according to the characteristics of the physical attack, that is, the main attack area is transferred from the outside of the target to the inside of the target. As shown in Figure. 5 (a) is the original image, which can be correctly identified as ABCD by the TRBA network. We have made some changes to the texture of the words, fill the number ‘8’ as the texture

inside the word, print it out, and use the iPhone 8 to take photos and input it into the model (as shown in Figure. 5 (b)). This image can successfully attack the model. We deployed the adversarial examples into the physical environment and converted them into digital images, which successfully attacked the STR models of iFLYTEK and Youdao(as shown in Table 4).

4. Conclusion

In this paper, we combine the style transfer method to attack the non-sequential network model and the sequential network model. This method can be used to generate natural antagonistic sample images with large perturbations and can be used to generate antagonistic samples in the physical world, as well as to evaluate the robustness of DNN models. Switch the view to protect the information, this method can be an efficient technique to protect objects, text, or human beings avoid artificial intelligence vision. The text or picture generated by this method can resist the acquisition of information by artificial intelligence, so as to target some places with high confidentiality requirements, it can well protect the information security.

Acknowledgement

The work was partially supported by the following: National Natural Science Foundation of China under no.61876155; Jiangsu Science and Technology Programme (Natural Science Foundation of Jiangsu Province) under no.BK20181189, BK20181190, BE2020006-4; Key Program Special Fund in XJTLU under no. KSF-T-06, KSF-E-26, and KSF-A-10;

References

- [1] Kaizhu Huang, Amir Hussain, Qiufeng Wang, and Rui Zhang. Deep learning: Fundamentals, theory, and applications, springer. In *ISBN 978-3-030-06072-5, 2019.*, 2019.
- [2] Yisen Wang, Xuejiao Deng, Songbai Pu, and Zhiheng Huang. Residual convolutional CTC networks for automatic speech recognition. *CoRR*, abs/1702.07793, 2017.
- [3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.
- [4] Chenyi Chen, Ari Seff, Alain L. Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2722–2730. IEEE Computer Society, 2015.
- [5] Guanyu Yang, Kaizhu Huang, Rui Zhang, John Goulermas, and Amir Hussain. Inductive generalized zero-shot learning with adversarial relation network. In *European Conference on Machine Learning (ECML)*, 2020.
- [6] Shufei Zhang, Kaizhu Huang, Rui Zhang, and Amir Hussain. Generalized adversarial training in riemannian space. In *In IEEE Fifteen Conference on Data Mining (ICDM'2019)*, 2019.
- [7] Chunchun Lyu, Kaizhu Huang, and Hai-Ning Liang. A unified gradient regularization family for adversarial examples. In *In IEEE Fifteen Conference on Data Mining (ICDM'2015)*, 2015.
- [8] Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2774–2783. IEEE Computer Society, 2017.
- [9] Xing Xu, Jiefu Chen, Jinhui Xiao, Lianli Gao, Fumin Shen, and Heng Tao Shen. What machines see is not what they get: Fooling scene text recognition models with adversarial text images. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition,*, pages 12301–12311. IEEE, 2020.
- [10] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423. IEEE Computer Society, 2016.
- [11] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *2019 ICCV*, pages 4714–4722, 2019.
- [12] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *CoRR*, abs/1406.2227, 2014.
- [13] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2315–2324, 2016.