

Adaptive Adversarial Attack on Scene Text Recognition

Xiaoyong Yuan*, Pan He*, Xiaolin Li[†], Dapeng Wu*

*NSF Center for Big Learning, University of Florida

[†]AI Institute, Tongdun Technology

chbrian@ufl.edu, pan.he@ufl.edu, xiaolin.li@tongdun.net, dpwu@ieee.org

Abstract—Recent studies have shown that state-of-the-art deep learning models are vulnerable to the inputs with small perturbations (adversarial examples). We observe two critical obstacles in adversarial examples: (i) Recent adversarial attacks require manually tuning hyper-parameters and take a long time to construct an adversarial example, making it impractical to attack real-time systems; (ii) Most of the studies focus on non-sequential tasks, such as image classification, yet only a few consider sequential tasks. In this work, we propose an adaptive approach to speed up adversarial attacks, especially on sequential learning tasks. By leveraging the uncertainty of each task, we directly learn the adaptive multi-task weightings, without manually searching hyper-parameters. A unified architecture is developed and evaluated for both non-sequential tasks and sequential ones. To evaluate the effectiveness, we take the scene text recognition task as a case study. To our best knowledge, our proposed method is the first attempt to adversarial attack for scene text recognition. *Adaptive Attack* achieves over 99.9% success rate with $3 \sim 6\times$ speedup compared to state-of-the-art adversarial attacks.

Index Terms—adversarial example, deep learning, scene text recognition, multi-task learning

I. INTRODUCTION

Recent studies [3], [6], [21], [23] have shown that deep neural networks are vulnerable to *adversarial examples*, by adding imperceptible perturbations on original images to fool a deep learning model. Adversarial examples have raised significant concerns since deep learning models are prevalent in many security-critical systems.

Accelerating Adversarial Attacks: Attacking deep learning models is time-limited in many real-world systems (*e.g.*, autonomous driving, face recognition). Many strong iterative optimization-based attacks [1], [3], [14] generate adversarial examples of high-quality that are hard to be defended [1], [2], but these iterative methods usually take a longer time to find proper weights/hyper-parameters for optimization, which becomes an obstacle to attacking real-time systems. For instance, it takes about one hour to generate a piece of adversarial audio with a few seconds [4].

On the other hand, accelerating adversarial attacks is a critical and non-trivial task for adversarial defenses. Adversarial Training, which trains the model with adversarial examples, has been shown as one of the most effective solutions to defend adversarial attacks [1]. However, it is time-consuming to generate adversarial examples for training using current iterative optimization-based attacks. To reduce attacking time, Taesik

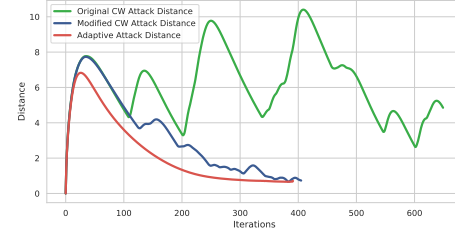


Fig. 1: ℓ_2 distance in the Attack Process: We attack an ImageNet example using three methods: original C&W attack, modified C&W attack, and our proposed Adaptive Attack and report the ℓ_2 distance of the perturbation in the attack process. Adaptive Attack achieves a small perturbation much faster than C&W attacks.

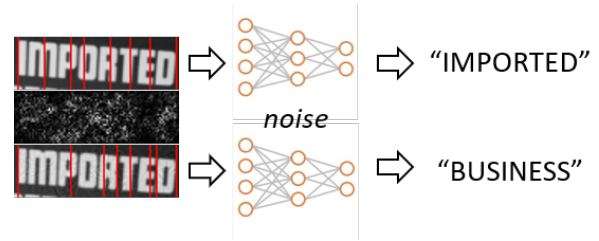


Fig. 2: An adversarial example of scene text recognition. First row: an original image ('IMPORTED'). Second row: adversarial perturbations added to the original image. Third row: the adversarial image, incorrectly predicted as 'BUSINESS'. Humans can barely recognize the difference between the two images. Red lines denote the CTC alignments.

et al. [16] trained with adversarial examples generated from C&W attack only in the first epoch, and then adopted FGSM, a faster one-step attack but with a degraded performance on perturbations, for the rest of training.

In this paper, we propose an *Adaptive Attack* to accelerate adversarial attacking leveraging multi-task learning. Generating adversarial examples is inherently a multi-task optimization problem, where we make an imperceptible change on the original samples while making the deep learning model to predict incorrectly. For classification tasks, we usually minimize distances between original examples and adversarial examples while minimizing the classification loss of adversarial examples on targeted labels.

Previous attacks simultaneously optimize two objectives, by using a weighted sum of multi-task losses. The weights of losses usually are uniformly defined [14] or manually tuned [3]. For example, C&W attack [3], a widely used adversarial attack, applies a modified binary search to find a proper weight in the attack. However, the optimal weights between two tasks are strongly dependent on tasks (e.g., image distance vs. audio distance, cross-entropy loss vs. CTC loss [7]). Researchers and practitioners have to carefully choose appropriate weights between task losses to achieve a good performance. Therefore, *it is desirable to find a better approach to learn the optimal weights automatically.*

Multi-task learning is widely studied in many machine learning tasks, which aims to improve learning efficiency and prediction accuracy by learning multiple objectives from a shared representation [12]. Recently, Kendall et al. proposed a principled approach for multi-task weightings, by combining observation (*aleatoric*) uncertainty and model (*epistemic*) uncertainty. They modeled each task in a unified Bayesian deep learning framework [11], [12] and outperformed the separately trained models for semantic segmentation, instance segmentation, and depth regression. Their solution is limited to non-sequential learning tasks (such as image classification, image segmentation), which might not directly apply to adversarial attacks on sequential learning tasks. In Section II-C, we extend this idea and further derive a novel approximate solution to sequential learning tasks.

Adversarial attack on non-sequential learning: We assume that two tasks in adversarial attacks (minimizing the classification loss and adversarial perturbations) follow probabilistic models. Our task is reformulated as attacking non-sequential classification problem by adaptively balancing tasks and tuning the weights. We refer to our proposed method as *Adaptive Attack*. Figure 1 illustrates that Adaptive Attack achieves a small adversarial perturbation much faster compared with strong and commonly-used attacks, C&W attack [2] for an image recognition task.

Adversarial attack on sequential learning: Current studies on adversarial examples mainly focus on non-sequential adversarial classification problems, such as image classification [6], [21], face recognition [18], reinforcement learning [13], and semantic segmentation [8]. Only very few studies target sequential learning tasks, such as speech-to-text [4] and reading comprehension [9], not to mention the analysis of sequential adversarial examples. *Applying Adaptive Attack to sequential learning tasks is non-trivial, due to the specific objective function and sequential properties.*

Take adversarial attacks on text recognition tasks as an example. The differences between non-sequential and sequential adversarial examples involve: i) The output of a sequential model is a varied-length label, instead of a single label. The non-sequential attacks (such as object classification model) only involve the *substitution* operation (e.g., modify the original class label), while the sequential attacks consider three operations: *insertion*, *substitution*, and *deletion* (e.g., insertion: coat → coats, substitution: coat → cost, deletion: coat →

cot); ii) Each character in target labels needs well-aligned. The requirement of the alignment between input and output poses a challenge in generating adversarial examples; iii) Sequential models usually leverage recurrent neural networks, where the internal feature representation involves more sequential context than those in convolutional neural networks.

Inspired by these observations, we conduct attacks on scene text recognition tasks. The scene text recognition is naturally a sequential learning task, which is closely related to standard classification tasks in computer vision. Measuring hidden features in the object classification task or speech recognition task is difficult, with more uncertainty on model interpretability. Adversarial attacks by modifying each character in the text image give us more intuitive explanations of how the perturbations affect the final output.

A. Contributions

We propose a novel *Adaptive Attack* that directly learns multi-task weightings without manually searching hyper-parameters, different from all previous adversarial attacks. *Adaptive Attack* is a general method that can be applied to the current iterative optimization-based attacks for both non-sequential and sequential tasks. Especially for the scene text recognition attack, *Adaptive Attack* accelerates adversarial attacking by three to six times. Also, we successfully attack a scene text recognition system with over 99.9% success rate. To the best of our knowledge, it is the first work to generate adversarial examples on a scene text recognition system.

II. ADAPTIVE ADVERSARIAL ATTACK

In this section, we present details on our proposed *Adaptive Attack* approach, which is inspired by recent findings in multi-task learning. The multi-task learning concerns the problem of optimizing a model with respect to multiple objectives [12]. The naive approach would be a linear combination of the losses for each task:

$$\mathcal{L} = \sum_i \lambda_i \mathcal{L}_i. \quad (1)$$

However, the model performance is extremely sensitive to weight selection, λ_i . It is also expensive to tune these weight hyper-parameters manually. In Bayesian modeling, we model these weight hyper-parameters using task-dependent uncertainty (homoscedastic uncertainty), which captures the relative important confidence between tasks, reflecting the uncertainty inherent to our multiple objectives. *Adaptive Attack* treats each task as a Gaussian distribution, where the mean is given by the model output, with an observation noise scalar σ . We will show how to relate σ to the relative weight of each loss. Our proposed *Adaptive Attack* generates adversarial examples on both non-sequential and sequential classification tasks, which generalizes the idea of multi-task learning in [12].

A. Threat Model

We assume that the adversary has access to the scene text recognition system, including the architecture and parameters of the recognition model. This type of attack is referred to

as “White-Box Attack.” We do not consider the “Black-Box Attack” in this paper, which assumes the adversary has no access to the trained neural network model. Prior work has shown that adversarial examples generated by “White-Box Attack” can be transferred to attack black-box services due to the transferability of adversarial examples [17]. Approximating the gradients [5] is another option for “Black-Box Attack.”

B. Basic Attack

Give an input image x , the ground-truth sequential label $l = \{l_0, l_1, \dots, l_T\}$, a targeted sequential label $l' = \{l'_0, l'_1, \dots, l'_T\}$ ($l \neq l'$), and a scene text recognition model \mathcal{F} , generating adversarial examples can be defined as the following optimization problem:

$$\begin{aligned} & \underset{x'}{\text{minimize}} \quad \mathcal{D}(x, x') \\ & \text{s.t.} \quad \mathcal{F}(x') = l', \\ & \quad \mathcal{F}(x) = l, \\ & \quad x' \in [-1, 1]^n, \end{aligned} \quad (2)$$

where x' is the modified adversarial image. $x' \in [-1, 1]^n$ ensures a valid input of x' . $\mathcal{D}(\cdot)$ denotes the distance between the original image and the adversarial image.

Following C&W Attack [3], we transform the function \mathcal{F} to a differentiable function, CTC_{Loss} . To remove the constraint of validation on new input x' , we introduce a new variable w to replace x' , where $x' = \tanh(w)$. The new optimization problem is given by:

$$\underset{w}{\text{minimize}} \quad CTC_{Loss}(\tanh(w), l') + \lambda \mathcal{D}(x, \tanh(w)), \quad (3)$$

where $CTC_{Loss}(\cdot, \cdot)$ denotes the CTC loss of the classifier \mathcal{F} . λ is a task and data dependent hyper-parameter to balance the importance of being adversarial and close to the original image. People usually search for a proper λ uniformly (log scale). In the experiments, we follow a modified binary search between $\lambda = 0.01$ and $\lambda = 1000$, starting from $\lambda = 0.1$. For each λ , we run 2,000 iterations of gradient descent searching using Adam. We adopt an early-stop strategy to avoid unnecessary computation.

C. Adaptive Attack

It is time-consuming to search for λ manually and find an optimal parameter λ because λ largely depends on individual tasks. In our experiments (Section III-B1), we compare the performance of Basic Attack with fixed λ values, regarding the success rate of attacks, the iterations of gradient search, and the magnitude of perturbations (Figure 3). We cannot find a proper value of fixed λ that achieves a high success rate, small iterations, and small perturbations simultaneously. For modified binary searching of λ , as long as we conduct enough searching steps, it can always find a proper λ to achieve a high success rate and small perturbations. However, it takes a much longer time, which makes adversarial examples hardly applicable to real-time systems.

We propose an adaptive search method. *Adaptive Attack*, to generate adversarial examples. From Equation 3, CTC loss $CTC_{Loss}(\cdot, \cdot)$ and Euclidean loss $\mathcal{D}(\cdot, \cdot)$ are viewed as a classification task and a regression task, respectively. Thus generating adversarial examples (Equation 3) becomes solving a multi-task problem with two objectives.

Non-sequential Classification Tasks: We optimize the adversarial images based on maximizing the Gaussian likelihood with uncertainty:

$$\text{maximize} \quad \Pr(x'|x, l'). \quad (4)$$

We assume that $\Pr(x')$ and $\Pr(x, l')$ are constant values, and x and l' are independent variables, then

$$\begin{aligned} \Pr(x'|x, l') &= \frac{\Pr(x, l'|x') \Pr(x')}{\Pr(x, l')} \\ &\propto \Pr(x, l'|x') = \Pr(x|x') \Pr(l'|x') \end{aligned} \quad (5)$$

We define original input x as a random variable which follows Gaussian distribution with mean x' and noise scale λ_1 :

$$\begin{aligned} \Pr(x|x') &= \mathcal{N}(x', \lambda_1^2), \\ \log \Pr(x|x') &\propto -\frac{\|x - x'\|_2^2}{2\lambda_1^2} - \log \lambda_1^2. \end{aligned} \quad (6)$$

For a classification task, we apply a classification likelihood to the output:

$$\Pr(l'|x') = \text{Softmax}(f(x')), \quad (7)$$

where $f(\cdot)$ denotes the output of the neural network before softmax layer. We use a squashed version of model output with a positive scalar λ_2 :

$$\begin{aligned} \Pr(l'|x', \lambda_2) &= \text{Softmax}\left(\frac{f(x')}{\lambda_2}\right), \\ \log \Pr(l' = c|x', \lambda_2) &= \frac{f_c(x')}{\lambda_2^2} - \log \sum_{c' \neq c} \exp \frac{f_{c'}(x')}{\lambda_2^2}. \end{aligned} \quad (8)$$

Then we define the joint loss (optimized objective) as follows:

$$\mathcal{L} = -\log \Pr(x, l'|x') = -\log \Pr(x|x') - \log \Pr(l'|x'). \quad (9)$$

According to [12], we can derive the optimization problem (Equation 4,5) by minimizing:

$$\mathcal{L} \propto \frac{\mathcal{L}_1(x, x')}{2\lambda_1^2} + \frac{\mathcal{L}_2(x', l')}{\lambda_2^2} + \log \lambda_1^2 + \log \lambda_2^2, \quad (10)$$

where $\mathcal{L}_1(x, x') = \|x - x'\|_2^2$ denoting the squared Euclidean distance and $\mathcal{L}_2(x', l')$ denotes the cross-entropy loss. We adaptively generate adversarial examples for non-sequential classification tasks without manually tuning λ . In practice, we replace $\log \lambda_i^2$ with η_i to avoid numerical instability.

Sequential Classification Tasks: Given an input sequence X , the network will output a probability distribution over the output domain for each frame. The probability of a given path π can be written as:

$$\Pr(\pi|x) = \prod_{t=1}^T y_{\pi_t}^t \quad (11)$$

We define \mathcal{B} as the conditional probability of a given labeling l . Let CTC_{Loss} be the negative log of probabilities of all the paths given l :

$$CTC_{Loss}(x, l) = -\log \sum_{\pi \in \mathcal{B}^{-1}(l)} \Pr(\pi|x). \quad (12)$$

The CTC_{Loss} of an adversarial example x' with targeted label l' is:

$$CTC_{Loss}(x', l') = -\log \sum_{\pi \in \mathcal{B}^{-1}(l')} \Pr(\pi|x'). \quad (13)$$

We consider paths with probability greater than or equal to a small constant c : $\Pr(\pi|x') = \prod_{t=1}^T y_{\pi_t}^t \geq c$. We first assume that only two valid paths satisfy this constraint: π_1, π_2 .

$$\begin{aligned} -\log \Pr(l'|x') &\approx -\log (\Pr(\pi_1|x') + \Pr(\pi_2|x')) \\ &\leq -\frac{1}{2} (\log \Pr(\pi_1|x') + \log \Pr(\pi_2|x')) - \log 2 \\ &\quad (\text{Jensen's inequality}) \\ &= -\frac{1}{2} \log \prod_{t=1}^T \Pr(y_{\pi_{1t}}^t) - \frac{1}{2} \log \prod_{t=1}^T \Pr(y_{\pi_{2t}}^t) - \log 2 \\ &= -\frac{1}{2} \sum_{t=1}^T (\log \Pr(y_{\pi_{1t}}^t) + \log \Pr(y_{\pi_{2t}}^t)) - \log 2 \\ &\approx \frac{1}{2} \sum_{t=1}^T \left(-\frac{A_{1,t}}{\lambda_1^2} + \log \lambda_1^2 - \frac{A_{2,t}}{\lambda_2^2} + \log \lambda_2^2 \right) - \log 2 \\ &= -\frac{\sum_{t=1}^T A_{1,t} + A_{2,t}}{2\lambda_1^2} + T \log \lambda_1^2 - \log 2, \end{aligned} \quad (14)$$

where $A_{i,t} = \log \text{Softmax}(y_{\pi_{it}}^t, f(x'))$, $i = 1, 2$.

CTC_{Loss} with two valid paths π_1, π_2 can be derived from:

$$\begin{aligned} CTC_{Loss} &\approx -\log (\Pr(\pi_1|x') + \Pr(\pi_2|x')) \\ &= -\log \Pr(\pi_1|x') - \log \Pr(\pi_2|x') - \log \frac{\Pr(\pi_1|x') + \Pr(\pi_2|x')}{\Pr(\pi_1|x') \Pr(\pi_2|x')} \\ &\geq -\log \Pr(\pi_1|x') - \log \Pr(\pi_2|x') - \log \frac{2}{c}. \end{aligned} \quad (15)$$

Combining Equation 14 and 15, we have an upper bound of $-\log \Pr(l'|x')$ and the joint loss \mathcal{L} :

$$\begin{aligned} -\log \Pr(l'|x') &\leq \frac{CTC_{Loss} + \log \frac{2}{c}}{2\lambda_1^2} + T \log \lambda_1^2 - \log 2, \\ \mathcal{L} &\leq \frac{\mathcal{L}_1(x, x')}{2\lambda_1^2} + \frac{CTC_{Loss}(x', l')}{2\lambda_2^2} \\ &\quad + \log \lambda_1^2 + T \log \lambda_2^2 + \frac{1}{\lambda_2^2} - \log 2. \end{aligned} \quad (16)$$

To extend the number of valid paths from 2 to an arbitrary number n , the joint loss \mathcal{L} satisfies:

$$\begin{aligned} \mathcal{L} &\leq \frac{\mathcal{L}_1(x, x')}{2\lambda_1^2} + \frac{CTC_{Loss}(x', l')}{n\lambda_2^2} + \log \lambda_1^2 + T \log \lambda_2^2 \\ &\quad + \frac{\log n - (n-1) \log c}{n\lambda_2^2}. \end{aligned} \quad (17)$$

Methods	Search Step	Success Rate	Distance	Iteration
Basic Attack	3	100%	1.957	199.298
Basic Attack	5	100%	0.689	342.993
Adaptive Attack	1	100%	0.517	253.088

TABLE I: Performance Comparison between Basic Attack and Adaptive Attack on the ImageNet dataset.

From our observation, the CTC loss always reduces very fast (Figure 5). Thus we can use a small number of valid paths to generate adversarial examples. From our experiments, it works well when $n < 50$. We use $n = 2$ to report our results. Thus, we can generate sequential adversarial examples by minimizing the upper bound of \mathcal{L} :

$$\frac{\mathcal{L}_1(x, x')}{\lambda_1^2} + \frac{CTC_{Loss}(x', l')}{\lambda_2^2} + \log \lambda_1^2 + T \log \lambda_2^2 + \frac{1}{\lambda_2^2}. \quad (18)$$

III. EXPERIMENTS AND ANALYSIS

In this section, we first evaluate the performance of Adaptive Attack for the non-sequential classification task comparing Adaptive Attack with C&W Attack. We then focus on the performance of Adaptive Attack for the sequential task. We attack a scene text recognition model as our use case. Besides, we investigate the generated images and the image changes during the attack for the sequential classification task.

A. Non-Sequential Attack

We evaluate non-sequential attacks for the image classification task. We use a pre-trained Inception V3 model [20] as the victim model, which achieves 22.55% top-1 error rate and 6.44% top-5 error rate in the ImageNet recognition challenge.

We compare Adaptive Attack with C&W attack. The original C&W attack achieves small perturbations but takes more iterations (See an example in Figure 1). To make a fair comparison, we modify the C&W attack by applying a more strict early-stop strategy - we stop attacking when the objective function (Equation 4) does not decrease in the past k iterations. We will use the modified C&W attack as our baseline (referred to as Basic Attack) in the paper. We set early-stop $k = 20$ for Basic Attack and $k = 1$ for Adaptive Attack, due to a more smooth objective function in Adaptive Attack.

We evaluate Adaptive Attack on the first 2,000 images in the validation set of ImageNet. We randomly assign a new label on an image and report success if the attack generates an adversarial image predicted with this label. Table I reports the success rate, average ℓ_2 distance of perturbation, and average attacking iterations for modified C&W attack (Basic Attack) and Adaptive Attack. Search step denotes the number of λ manually searched in the Basic Attack.

Our results show that Adaptive Attack reaches small perturbations within a much less time compared to Basic Attack.

B. Sequential Attack

We evaluate sequential attacks on a scene text recognition model. We compare the performance of Basic Attack and propose Adaptive Attack on three standard benchmarks. We then analyze the sequential attacks on a simulated sequential MNIST dataset.

Methods	IC13			SVT			IIIT5K		
	Success Rate	Distance	Iteration	Success Rate	Distance	Iteration	Success Rate	Distance	Iteration
Basic0.1	99.90%	3.57	1621.88	99.69%	3.59	1470.29	99%	2.90	7127.92
Basic1	88.55%	1.75	526.92	91.55%	1.67	518.99	95.39%	1.77	3606.26
Basic10	53.38%	0.44	179.75	68.23%	0.47	172.82	39.12%	0.39	1395.99
BasicBinary3	100.00%	1.64	1531.84	100.00%	1.17	1442.86	100.00%	2.01	4097.52
BasicBinary5	100.00%	1.64	1706.18	100.00%	1.15	1616.35	100.00%	1.96	5055.21
BasicBinary10	100.00%	1.58	2138.86	100.00%	1.11	1993.47	100.00%	1.94	6811.86
Adaptive	100.00%	2.15	480.28	100.00%	1.26	529.90	99.96%	2.68	682.48

TABLE II: Performance Comparison between Basic Attack and Adaptive Attack on Three Scene Text Recognition Benchmarks.

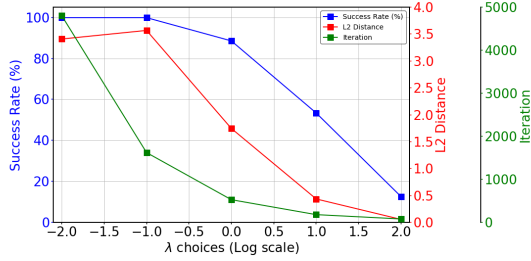


Fig. 3: **Basic Attack with fixed λ values on the IC13 dataset.** None of the λ values can well balance CTC loss (success rate curve in blue), distance (ℓ_2 distance curve in red), and optimizing time (iteration curve in green).

1) *Basic Attack vs. Adaptive Attack:* We conducted experiments on three standard benchmarks for cropped word image recognition: the Street View Text dataset (SVT) [22], the ICDAR 2013 dataset (IC13) [10], and the IIIT 5K-word dataset (IIIT5K) [15]. We train an end-to-end deep learning model using Pytorch, based on the state-of-the-art scene text recognition approach, Convolutional Recurrent Neural Network (CRNN) [19]¹.

We compare Basic Attack and Adaptive Attack on these benchmarks. The targeted sequential label is set as the common word with the same length as the original one. Figure 3 shows the performance of Basic Attack with fixed λ values. We run gradient descent searches for 10,000 iterations. Early stopping is adopted to avoid unnecessary computation. We only calculate distances and iterations of the successful attacks to avoid the extremely large values when the attack fails. The results show that when we use large λ values (1, 10, 100), it fails to generate adversarial examples in most cases. For small λ values (0.1, 0.01), although Basic Attack successfully generates adversarial images, it spends a much longer time and brings a larger magnitude of perturbations.

We then compare the performance of Adaptive Attack and Basic Attack using a fixed λ or a modified binary search. Table II lists our results. *Basic0.1*, *Basic1*, and *Basic10* denote Basic Attack with fixed λ values: 0.1, 1, and 10, respectively. *BasicBinary3*, *BasicBinary5*, *BasicBinary10* denote Basic Attack with 3, 5, and 10 steps of binary searching. We set the initial λ as 0.1, which is the best λ according to the results of fixed λ values (Figure 3).

¹we refer to the implementation on github.com/bgshih/crnn and modify the kernel size in the pooling layers for better alignment.

From Table II, we observe that both Adaptive Attack and Basic Attack with a modified binary search can successfully generate adversarial examples on the scene text recognition model. Basic Attack with fixed λ values cannot achieve both success rate and low distance of perturbations. Adaptive Attack conducts attacks much faster ($3 \sim 6\times$) than Basic Attack with a modified binary search. Although Basic Attack achieves smaller perturbations than Adaptive Attack, it is reasonable for binary search method to have a finer tuning on the λ if the initial λ value is properly set and the number of binary search iterations is large enough.

C. Analysis of Sequential Attack

To dig into the phenomenon of sequential adversarial examples, we generate adversarial examples on a simple sequential classification task. We first simulate a sequential digit dataset by concatenating digit images in the MNIST dataset and fit them into a 32x100 pixel box. The training and test sequential digits are generated from MNIST training and test sets. We refer to this dataset as *SeqMNIST*. The first row in Figure 4 illustrates an example ('24500'). We then trained our model with *SeqMNIST*.

We analyze three types of common adversarial operations on targeted sequential labels: insertion, substitution, and deletion. We perform these operations on one digit and remain the rest unchanged. We also include another operation that inserts a repeated digit (e.g., '24500' \rightarrow '245500'). We perform an adversarial attack on the *SeqMNIST* dataset and exam 100 adversarial images for each operation.

CTC alignments in the adversarial examples: We observe that most CTC alignments are stable against adversarial examples (Figure 4). When added perturbations on the images, only the CTC alignments surrounding the targeted labels will be changed. We investigate the four operations: (1) insertion, (2) repeated insertion, (3) substitution, and (4) deletion.

Visualizing the attack process: We analyze the process of generating adversarial examples and focus on the change of the CTC alignment and the perturbation distance. Figure 5 visualizes the attack process of a *SeqMNIST* sample. The adversarial perturbations are firstly added to the edges of digits surrounding the targeted positions, (the new positions based on various adversarial operations) where the gradients are larger than the others. The CTC alignments will be aligned to the targeted positions very quickly (in most cases, less than 100 iterations). After the CTC alignments get stable, the alignments will not be changed. The later changes

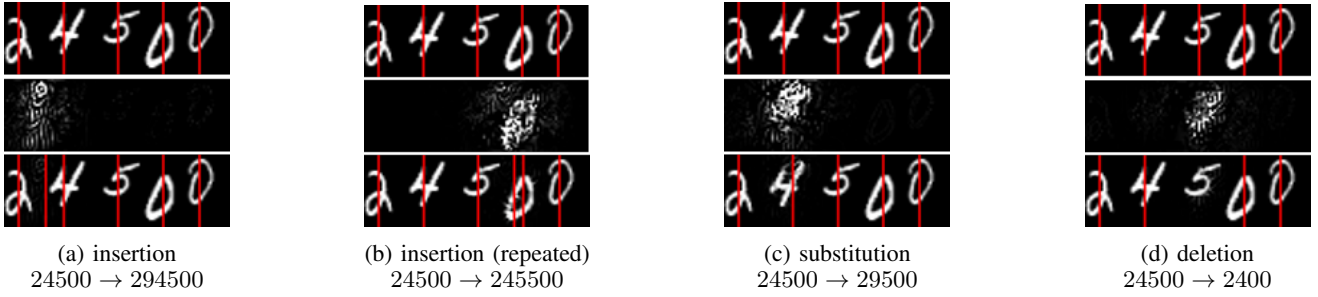


Fig. 4: **Four types of adversarial attacks on the SeqMNIST dataset.** The images in the first row are the same original images. After perturbing (amplified by $10\times$, second row), we generate the adversarial images (third row), which can be misclassified as different labels.

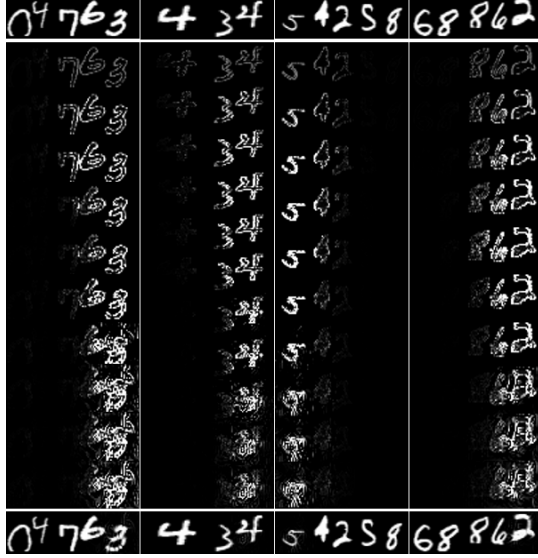


Fig. 5: **Attack Process.** We demonstrate four adversarial examples on the SeqMNIST dataset, including insertion ('04763' → '047673'), insertion repeated ('434' → '4334'), substitution ('54258' → '94258'), and deletion ('68862' → '6886'). First row: original images. Last row: adversarial images. The middle rows: adversarial perturbations (amplified by $10\times$) added on the original images in the order of iterations.

mainly minimize the adversarial perturbations while keeping the prediction unchanged. In a few iterations, the adversarial perturbations look similar to the targeted digits. In the final iteration, it will be minimized to an 'abstract' version of targeted digits, which can not be recognized by humans. Based on previous observations, CTC alignments only change in the first few iterations and appear close to the targeted (inserted/substituted/ deleted) positions. After that, the adversarial attack will focus on minimizing the magnitude of perturbations.

IV. CONCLUSIONS

We have proposed a novel approach to learn multi-task weights without manually tuning the hyper-parameters. The proposed Adaptive Attack method substantially speeds up the process of adversarial attacks for both non-sequential and sequential tasks. Our proposed attacks successfully fooled the scene text recognition system with over 99.9% success rate on three standard benchmark datasets. We will investigate defense

mechanisms against *Adaptive Attack* for both non-sequential and sequential tasks in our future work.

REFERENCES

- [1] A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, pages 274–283, 2018.
- [2] N. Carlini and D. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. *AISEC*, 2017.
- [3] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy*, 2017.
- [4] N. Carlini and D. Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *IEEE Security and Privacy Workshops*, 2018.
- [5] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26, 2017.
- [6] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.
- [7] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, pages 369–376. ACM, 2006.
- [8] J. Hendrik Metzen, M. Chaithanya Kumar, T. Brox, and V. Fischer. Universal adversarial perturbations against semantic image segmentation. In *CVPR*, pages 2755–2764, 2017.
- [9] R. Jia and P. Liang. Adversarial examples for evaluating reading comprehension systems. In *EMNLP*, 2017.
- [10] D. Karatzas, F. Shafait, and S. e. a. Uchida. Icdar 2013 robust reading competition. In *ICDAR*, pages 1484–1493. IEEE, 2013.
- [11] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS*, pages 5574–5584, 2017.
- [12] A. Kendall, Y. Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018.
- [13] J. Kos and D. Song. Delving into adversarial attacks on deep policies. *ICLR Workshop*, 2017.
- [14] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *ICLR Workshop*, 2017.
- [15] A. Mishra, K. Alahari, and C. V. Jawahar. Scene text recognition using higher order language priors. In *BMVC*, 2012.
- [16] T. Na, J. H. Ko, and S. Mukhopadhyay. Cascade adversarial machine learning regularized with a unified embedding. *ICLR*, 2018.
- [17] N. Papernot, P. D. McDaniel, and I. J. Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.
- [18] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *CCS*, pages 1528–1540. ACM, 2016.
- [19] B. Shi, X. Bai, and C. Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on PAMI*, 39(11):2298–2304, 2017.
- [20] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.
- [21] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.
- [22] K. Wang, B. Babenko, and S. Belongie. End-to-end scene text recognition. In *ICCV*, 2011.
- [23] X. Yuan, P. He, Q. Zhu, and X. Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 2019.