Adaptive Adversarial Attack on Scene Text Recognition

Xiaoyong Yuan, Pan He, Xiaolin Li
University of Florida

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) Strong adversarial attacks (e.g., C&W attack) 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 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 validate 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 ∼ 6× speedup compared to state-of-the-art adversarial attacks.

1 Introduction

Recent studies [6, 10, 34] 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). In the recent NIPS 2018 Adversarial Vision Challenge, each attack is requested to "process a batch of 10 images within 900s on a K80 GPU" [4]. Many strong iterative optimization-based attacks [2, 6, 23] generate adversarial examples of high-quality and hard to be defended [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 [7].

Also, accelerating adversarial attacks is a critical and non-trivial task for adversarial defenses. Adversarial Training, adding adversarial examples to the training set, has been shown the most promising solution to defend adversarial attacks [2]. However,
Figure 2: An adversarial example of scene text recognition. First row: an original image (‘IMPORTED’) from the IC13 dataset. Second row: adversarial perturbations added on 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.

it is time-consuming to generate adversarial examples for training using current iterative optimization-based attacks. To reduce attacking time, [26] 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, inspired by recent findings in multi-task learning. Generating adversarial examples is inherently a multi-task optimization problem, where we try to 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 (e.g., cross-entropy loss, CTC loss [11]) of adversarial examples on targeted labels.

Previous attacks simultaneously optimize two objectives, by using a naïve weighted sum of multi-task losses. The weights of losses usually are uniformly defined [23] or manually tuned [6]. For example, C&W attack [6] is a commonly-used attack which 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). 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, where they aim to improve learning efficiency and prediction accuracy by learning multiple objectives from a shared representation [21]. 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 [20,21] 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 3.3, we extend this idea and further derive a novel approximate solution to sequential learning tasks.

Adversarial attack on non-sequential learning: Similar to [21], 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 [6] for an image recognition task (See Section 4.1 for details).

Adversarial attack on sequential learning: Current studies on adversarial examples mainly focus on non-sequential adversarial classification problems, such as image classification [10,34], face recognition [29], reinforcement learning [22], and semantic segmentation [13]. Only very few targets at sequential learning tasks, such as speech-to-text [7] and reading comprehension [18], 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 se-
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 $\rightarrow$ coats, substitution: coat $\rightarrow$ cost, deletion: coat $\rightarrow$ cot). (ii) Each character in target labels needs well-aligned. The requirement of the alignment between input and output poses a challenge on 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 will give us more intuitive explanations of how the perturbations affect the final output.

1.1 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 which 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.

2 Background

2.1 Scene Text Recognition

Scene text recognition tasks aim at decoding a sequence of text characters from a cropped but variable-length word image.

Recent scene text recognition approaches focus on mapping the entire image to a word string, either with hand-crafted features or deep learning features. [1] proposed a subspace regression method to jointly embed both word images and their text strings into a common subspace, resulting in solving a nearest neighbor problem. [16] developed a CNN model to cast the word recognition into a multi-class classification problem. They further proposed a CNN based architecture, incorporating a conditional random field graphical model for unconstrained text recognition [15]. [2] built an RNN with HOG features and casted text recognition as a sequence labeling problem. [12] and [30] extracted rich visual features from a CNN; then a sequence labeling is carried out with LSTM and CTC. [24] incorporated attention modeling into recursive recurrent neural networks for lexicon-free optical character recognition in natural scene images. [31] and [36] extended the work of [17] to transform a distorted text region into a canonical pose suitable for recognition. In our experiment, we train a similar recognition model, based on the state-of-the-art approach, Convolutional Recurrent Neural Network (CRNN) [30].

2.2 Adversarial Examples

Adversarial examples are imperceptible to human but can easily fool deep neural networks in the testing/deploying stage. People have studied adversarial examples of machine learning models (machine learning evasion attacks) since 2012 [3].

The vulnerability to adversarial examples becomes one of the major obstacles for applying deep neural networks in many safety-critical scenarios. Most existing studies on adversarial examples focus on computer vision related tasks: image classification [34], face recognition [29], reinforcement learning [14], and semantic segmentation [13]. Only a few studies have
been devoted to sequential domains, such as speech recognition [7] and reading comprehension [18]. Iterative attacks have been prevalent in adversarial attacks due to high success rates and small perturbations: iteratively searching and updating the perturbations based on the gradient of the output of the victim model. In contrast, one-time attacks update the adversarial perturbations only once and are usually used in real-time systems. For example, [14] and [22] leveraged a one-time attack, Fast Gradient Sign Method (FGSM) [10], to attack reinforcement learning systems due to the requirement of quick response to the real-time input. However, it is easy to detect/defend one-time attacks.

The magnitude of perturbations is measured by $\ell_p$ norm ($p = 0, 1, \infty$). A few studies use other measurements (e.g., SSIM [28], spatial transformation [9]). In this paper, we only consider $\ell_2$ norm, which can be comparable to most of the current work.

To our best knowledge, we are the first group to successfully generate sequential adversarial examples on a scene text recognition task.

3 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 [21]. The naive approach would be a linear combination of the losses for each task:

$$\mathcal{L} = \sum_i \lambda_i \mathcal{L}_i.$$  

(1)

However, the model performance is extremely sensitive to weight selection, $\lambda_i$. It is also costly 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 $\sigma$. We will show how to relate $\sigma$ 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 [21].

3.1 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 [27]. Approximating the gradients [8] is another option for “Black-Box Attack.”

3.2 Basic Attack

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

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

(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 [6], we transform the function $\mathcal{F}$ to a differentiable function, $CTC_{\text{Loss}}$. 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:

$$\begin{align*}
\min_{x'} \quad & CTC_{\text{Loss}}(\tanh(w), l') + \lambda \mathcal{D}(x, \tanh(w)) ,
\end{align*}$$  

(3)
where $CTCLoss(\cdot, \cdot)$ denotes the CTC loss of the classifier $\mathcal{F}$. $\lambda$ is a task and data dependent hyperparameter to balance the importance of being adversarial and close to the original image. People usually search for a proper $\lambda$ uniformly (log scale). In the experiment, we follow a modified binary search between $\lambda = 0.01$ and $\lambda = 1000$, starting from $\lambda = 0.1$. For each $\lambda$, we run 2,000 iterations of gradient descent searching using Adam. We adopt an early-stop strategy to avoid unnecessary computation.

### 3.3 Adaptive Attack

It is time-consuming to search for $\lambda$ manually and find an optimal parameter $\lambda$ since $\lambda$ largely depends on individual tasks. In our experiment (Section 4.2.1), we compare the performance of Basic Attack with fixed $\lambda$ 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 $\lambda$ that achieves a high success rate, small iterations, and small perturbations simultaneously. For modified binary searching of $\lambda$, as long as we conduct enough searching steps, it can always find a proper $\lambda$ 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 $CTCLoss(\cdot, \cdot)$ and Euclidean loss $D(\cdot, \cdot)$ are viewed as a classification task and a regression task respectively. Thus generating adversarial examples (Equation 4) 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:

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

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

$$
\Pr(x'|x, l') = \frac{\Pr(x, l'|x') \Pr(x')}{\Pr(x, l')} 
\propto \Pr(x, l'|x') = \Pr(x|x') \Pr(l'|x')
$$

We define original input $x$ as a random variable which follows Gaussian distribution with mean $x'$ and noise scale $\lambda_1$:

$$
\Pr(x|x') = \mathcal{N}(x', \lambda_1^2),
$$

$$
\log \Pr(x|x') \propto -\|x - x'\|_2^2 - \log \lambda_1^2. \tag{6}
$$

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

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

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 $\lambda_2$:

$$
\Pr(l'|x', \lambda_2) = \text{Softmax}(\frac{f(x')}{\lambda_2}),
$$

$$
\log \Pr(l' = c|x', \lambda_2) = \frac{f_c(x')}{\lambda_2^2} - \log \sum_{c' \neq c} \exp \frac{f_{c'}(x')}{\lambda_2^2}. \tag{8}
$$

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

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

According to [21], we can derive the optimization problem (Equation 10) by minimizing:

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

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 $\lambda$. In practice, we replace $\log \lambda_1^2$ with $\eta_1$ 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 $\pi$ can be written as:

$$
\Pr(\pi|x) = \prod_{t=1}^{T} \tilde{y}_{\pi_t}^t. \tag{11}
$$
We define $B$ as the conditional probability of a given labeling $l$. Let $CTC_{Loss}$ be the negative log of probabilities of all the paths given $l$:

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

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

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

We consider paths with probability greater than or equal to a small constant $c$: $Pr(\pi | x') = \prod_{t=1}^{T} y_{l_{t}}^{x'} \geq c$. We first assume that only two valid paths satisfy this constraint: $\pi_{1}, \pi_{2}$.

$$-\log Pr(l'|x') \approx -\log (Pr(\pi_{1}|x') + Pr(\pi_{2}|x')) \leq -\frac{1}{2} (\log Pr(\pi_{1}|x') + \log Pr(\pi_{2}|x')) - \log 2 \quad \text{(Jensen's inequality)}$$

$$= -\frac{1}{2} \log \prod_{i=1}^{T} Pr(y_{l_{t}}^{x'}) - \frac{1}{2} \log \prod_{i=1}^{T} Pr(y_{l_{t}}^{x'}) - \log 2$$

$$= -\frac{1}{2} \sum_{i=1}^{T} \left( \frac{A_{1,i}}{\lambda_{2}} + \log \lambda_{2}^{2} - \frac{A_{2,i}}{\lambda_{2}} + \log \lambda_{2}^{2} \right) - \log 2 \quad \text{(Similar to Eq. (8))}$$

$$= -\sum_{t=1}^{T} \frac{A_{1,t} + A_{2,t}}{2\lambda_{2}^{2}} + T \log \lambda_{2}^{2} - \log 2, \quad (14)$$

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

$CTC_{Loss}$ with two valid paths $\pi_{1}, \pi_{2}$ becomes:

$$CTC_{Loss} \approx -\log (Pr(\pi_{1}|x') + Pr(\pi_{2}|x'))$$

$$= -\log Pr(\pi_{1}|x') - \log Pr(\pi_{2}|x') - \log \frac{Pr(\pi_{1}|x') + Pr(\pi_{2}|x')}{Pr(\pi_{1}|x') Pr(\pi_{2}|x')}$$

$$\geq -\log Pr(\pi_{1}|x') - \log Pr(\pi_{2}|x') - \log 2 \frac{2}{c}. \quad (15)$$

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

$$-\log Pr(l'|x') \leq \frac{CTC_{Loss} + \log \frac{2}{c}}{2\lambda_{2}^{2}} - \log 2,$$

$$\mathcal{L} \leq \frac{\mathcal{L}_{1}(x, x')}{2\lambda_{1}^{2}} + \frac{CTC_{Loss}(x', l')}{2\lambda_{2}^{2}} + \log \lambda_{1}^{2} + T \log \lambda_{2}^{2} + \frac{1}{\lambda_{2}^{2}} - \log 2. \quad (16)$$

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

$$\mathcal{L} \leq \frac{\mathcal{L}_{1}(x, x')}{2\lambda_{1}^{2}} + \frac{CTC_{Loss}(x', l')}{n\lambda_{2}^{2}} + \log \lambda_{1}^{2} + T \log \lambda_{2}^{2} + \frac{1}{\lambda_{2}^{2}}.$$ \quad (17)

From our observation, 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 $\mathcal{L}$:

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

### 4 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. Then we 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.

#### 4.1 Non-Sequential Attack

We evaluate non-sequential attacks for the image classification task. We use a pre-trained Inception V3 model [33] as the victim model, which achieves 22.55% top-1 error rate and 6.44% top-5 error rate in 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 4). 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 1 reports the success rate, average $\ell_2$ distance of perturbation, and average attacking iterations for modified C&W attack (Basic Attack) and Adaptive Attack. Search step denotes the number of $\lambda$ manually searched in the Basic Attack.

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

4.2 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.

4.2.1 Basic Attack vs. Adaptive Attack

We conducted experiments on three standard benchmarks for cropped word image recognition: the Street View Text dataset (SVT) \[^{35}\], the ICDAR 2013 dataset (IC13) \[^{19}\], and the IIIT 5K-word dataset (IIIT5K) \[^{25}\]. 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) \[^{30}\].

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 $\lambda$ 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 $\lambda$ values (1, 10, 100), it fails to generate adversarial examples in most cases. For small $\lambda$ 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 $\lambda$ or a modified binary search. Table 2 lists our results. Basic0.1, Basic1, and Basic10 denote Basic Attack with fixed $\lambda$ 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 $\lambda$ as 0.1, which is the best $\lambda$ according to the results of fixed $\lambda$ values (Figure 3).

From Table 2 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 $\lambda$ 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 $\lambda$ if the initial $\lambda$ value is prop-

![Figure 3: Basic Attack with fixed $\lambda$ values. We generate adversarial examples on the IC13 dataset using Basic Attack. None of the $\lambda$ values can well balance CTC loss (success rate curve in blue), distance ($\ell_2$ distance curve in red), and optimizing time (iteration curve in green).](image)

\[^{1}\text{we refer to the implementation on } \texttt{github.com/bgshih/crnn} \text{ and modify the kernel size in the pooling layers for better alignment.}\]


| 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 1: Performance Comparison between Basic Attack and Adaptive Attack on ImageNet

| 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     | 1476.29  | 99%       | 2.90    | 7127.92  |
| Basic1         | 88.55% | 1.75     | 526.92   | 91.55%     | 1.67     | 518.99   | 95.39%    | 1.77    | 3066.26  |
| 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%    | 1.77     | 1442.86  | 100.00%   | 2.01    | 4097.52  |
| BasicBinary5   | 100.00%| 1.64     | 1706.18  | 100.00%    | 1.58     | 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 2: Performance Comparison between Basic Attack and Adaptive Attack on Three Scene Text Recognition Benchmarks

![Image of adversarial attacks on SeqMNIST](image)

(a) insertion 24500 → 294500  
(b) insertion (repeated) 24500 → 245500  
(c) substitution 24500 → 29500  
(d) deletion 24500 → 2400

Figure 4: Four types of Adversarial attacks on the SeqMNIST dataset: insertion, insertion (repeated), substitution, and deletion. The images in the first row are the same original images. After perturbing (amplified by 10×, second row), we generate the adversarial images (third row), which can be misclassified as different labels. Red lines illustrate the corresponding CTC alignments of original labels and targeted labels. There are 25 alignment positions each representing 4 pixels in width.

4.3 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 which inserts a repeated digit (e.g., ‘24500’ → ‘245500’). We perform an adversarial attack on the SeqMNIST dataset and examine 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
Figure 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×) added on the original images in the order of iterations.

added perturbations on the images, only the CTC alignments surrounding the targeted labels will be changed. We investigate the following four operations. (i) **Insertion**: When inserting a digit into the targeted label, we find that the added perturbations usually appear in the middle of two adjacent digits. Sometimes, the neighboring pixels are ‘borrowed’ by the new digit that is close to one side for self-construction, which costs smaller perturbations. However, it is not the case for repeated insertions. (ii) **Insertion (repeated)**: The added perturbations usually appear far from the repeated digit. For instance, the new ‘5’ is close to ‘0’ and far from ‘5’ (Figure 4). This is a more efficient and optimal solution to attack CTC alignments. (iii) **Substitution**: When substituting a label, the CTC alignments change slightly in the position of substitution. The rest of the targeted labels remain in the same positions. (iv) **Deletion**: When we delete the targeted digit, the remaining CTC alignments barely change their positions. It also requires the least magnitude of perturbations.

Figure 5 visualizes the attack process of a SeqMNIST sample. We find that 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.

5 Conclusions

We proposed a novel approach to learn multi-task weights without manually tuning the hyperparameters. The proposed Adaptive Attack method substantially speeds up the process of adversarial attacks for both non-sequential and sequential tasks. We successfully attacked a popular scene text recognition system with over 99.9% success rate on three standard benchmark datasets. Our future work will investigate i) defense mechanisms using Adaptive Attack, ii) defense mechanism for sequential tasks.

References

[1] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. *IEEE TPAMI*, 36(12):2552–2566, 2014.

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

[3] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *arXiv preprint arXiv:1712.03141*, 2017.

[4] Wieland Brendel, Jonas Rauber, Alexey Kurakin, Nicolas Papernot, Behar Velij, Marcel Salathé, Sharada P Mohanty, and Matthias Bethge. Adversarial vision challenge. *arXiv preprint arXiv:1808.01976*, 2018.

[5] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. *AISEC*, 2017.

[6] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *S&P*, pages 39–57. IEEE, 2017.

[7] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. *arXiv preprint arXiv:1801.01944*, 2018.
[8] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. *arXiv preprint arXiv:1708.03999*, 2017.

[9] Logan Engstrom, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling cnns with simple transformations. *arXiv*, 2017.

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[11] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, pages 369–376. ACM, 2006.

[12] Pan He, Weilin Huang, Yu Qiao, Chen Change Loy, and Xiaoou Tang. Reading scene text in deep convolutional sequences. *AAAI*, 2016.

[13] Jan Hendrik Metzen, Mumma Chaitanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. In *CVPR*, pages 2755–2764, 2017.

[14] Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017.

[15] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep structured output learning for unconstrained text recognition. *arXiv*, 2014.

[16] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv*, 2014.

[17] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015.

[18] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *EMNLP*, 2017.

[19] Dimosthenis Karatzas, Faisal Shafait, and Seiichi et. al. Uchida. Icdar 2013 robust reading competition. In *ICDAR*, pages 1484–1493. IEEE, 2013.

[20] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS*, pages 5574–5584, 2017.

[21] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018.

[22] Jernej Kos and Dawn Song. Delving into adversarial attacks on deep policies. *ICLR Workshop*, 2017.

[23] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[24] Chen-Yu Lee and Simon Osindero. Recursive recurrent nets with attention modeling for ocr in the wild. In *CVPR*, 2016.

[25] A. Mishra, K. Alahari, and C. V. Jawahar. Scene text recognition using higher order language priors. In *BMVC*, 2012.

[26] Taesik Na, Jong Hwan Ko, and Saibal Mukhopadhyay. Cascade adversarial machine learning regularized with a unified embedding. *ICLR*, 2018.

[27] Nicolas Papernot, Patrick D. McDaniel, and Ian J. Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

[28] Andras Rozsa, Ethan M Rudd, and Terrance E Boult. Adversarial diversity and hard positive generation. In *CVPR Workshops*, pages 25–32, 2016.

[29] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *CCS*, pages 1528–1540. ACM, 2016.

[30] Baoguang Shi, Xiang Bai, and Cong 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.

[31] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. In *CVPR*, 2016.

[32] Bolan Su and Shijian Lu. Accurate scene text recognition based on recurrent neural network. In *ACCV*, pages 35–48. Springer, 2014.

[33] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.
[34] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.

[35] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In ICCV, 2011.

[36] Kwan-Yee K. Wong Zhizhong Su Wei Liu, Chaofeng Chen and Junyu Han. Star-net: A spatial attention residue network for scene text recognition. In BMVC, 2016.