Test Time Augmentation as a Defense Against Adversarial Attacks on Online Handwriting*

Yoh Yamashita and Brian Kenji Iwana[0000−0002−5146−6818]

Graduate School of Information Science and Electrical Engineering
Kyushu University, Fukuoka, Japan
yoh.yamashita@human.ait.kyushu-u.ac.jp, iwana@ait.kyushu-u.ac.jp

Abstract. Neural networks have been shown to be weak against adversarial attacks. This study examines the effects of adversarial attacks on online handwritten characters and proposes a method to defend against such attacks. In order to make temporal neural networks more robust to adversarial attacks, we propose using Test Time Augmentation (TTA). TTA combines the predictions of transformed inputs with a trained classifier. We adapt TTA and propose its usage to make temporal neural networks more robust to adversarial attacks. The proposed method is evaluated using online handwritten characters and against four state-of-the-art adversarial attacks. We demonstrate that the nontraditional use of TTA can be used to protect against these attacks for almost no cost.

Keywords: Online Handwriting · Adversarial Attacks · Defense

1 Introduction

Online handwriting recognition is the process of converting handwritten input, typically captured using a stylus or touchscreen device, into digital text. In this way, online handwriting consists of a time series made of coordinates, strokes, or other features. This is in contrast to offline handwriting recognition which uses static images of handwritten text. Online handwriting recognition is important due to its applications in various domains, such as handwriting input on touchscreens, biometric authentication and signature verification systems, and smart whiteboards.

Notably, online handwriting recognition and temporal neural networks have a close history. Deep neural networks have had widespread successes in pattern recognition [35] and time series classification [41], including online handwriting recognition. For example, early uses of novel neural networks have been used for online handwriting in a variety of languages, such as English [10, 17], Chinese [38], Japanese [32], Arabic [30], Devanagari [21], Mongolian [42], etc. Also, online handwriting was one of the early uses of Connectionist Temporal Classification (CTC) [10]. Today, most state-of-the-art online handwriting recognition systems incorporate neural networks [8].

* This work was partially supported by MEXT-Japan (Grant No. 23K16949).

While neural networks have had a lot of successes, they have been shown to be weak against adversarial examples [39]. Adversarial examples are attacks on neural networks that aim to cause the neural network to incorrectly recognize examples. One of the most common types of attacks is to add adversarial noise or perturbations to examples. As shown in Fig. 1, the noise should be hardly perceivable to humans, but would have a large impact on the ability of the neural network. While most attacks have been designed for image recognition, temporal neural networks have also shown to be weak against adversarial attacks [7].

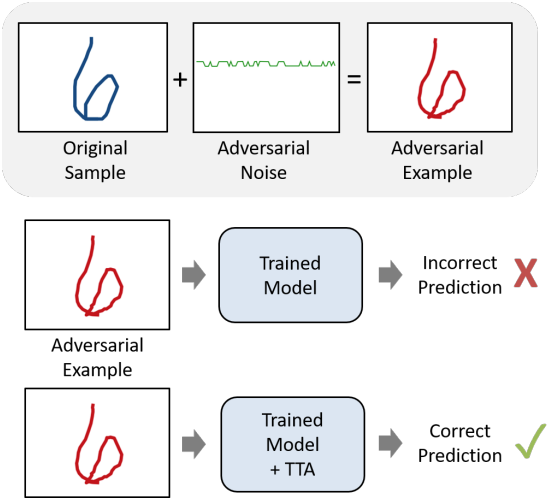

Fig. 1. The workflow of an adversarial attack and defense against it for online handwriting.

To defend against adversarial attacks, defense algorithms have been proposed. These defense algorithms aim to automatically defend against adversarial attacks on neural networks, irrespective of knowing if an attack is taking place or which attack. For example, one method of defending against attacks is the use of ensemble networks [37, 11]. However, ensembles typically require multiple trained models, which can cause limitations for resource-limited systems and cannot be applied to all networks. Therefore, there have been other methods used to add robustness to neural networks, such as Random Self-Ensembles (RSE) [27], feature squeezing [43], denoising [26], etc. Furthermore, most defense methods were designed for image recognition and little research exists for temporal neural networks [7].

In order to address adversarial attacks on online handwriting, in this paper, we propose a new method to increase the robustness of temporal neural networks. Namely, we propose a novel use of Test Time Augmentation (TTA) [36] as a defense against attacks on online handwriting. In general, TTA is a data

augmentation technique applied during the testing phase, where multiple augmented versions of a given input are generated to increase the robustness of classifying unseen or unusual data. In most TTA applications, during training time, the model is trained like normal. Only during test time is the input modified and used to create variations in the predictions. These predictions are then ensembled in the hope of mitigating overfitting by leveraging the variations to generalize the results.

However, instead of using TTA for data augmentation, we propose to use TTA as a method to defend against adversarial attacks. Using the trained network, we classify the online handwritten characters under four transformations, jittering, window slicing,time warping, and window warping. The results from the four transformations plus the original characters are then combined. By combining the transformations, the proposed method is able to disrupt the adversarial noise without sacrificing the accuracy of the model.

The contributions of this paper are as follows:

1. As far as we know, we are the first to demonstrate the effectiveness of adversarial attacks on online handwritten characters.
2. We propose using a novel use of TTA, as a defense against adversarial attacks.
3. The proposed method is evaluated against three defenses under four adversarial attack methods. The defenses used are Random Self-Ensemble (RSE) [27], random noise, and a median filter. The attacks used are a Carlini and Wagner attack (CW) [5], Fast Gradient Sign Method (FGSM) [9], Basic Iterative Method (BIM) [22], and Projected Gradient Descent (PGD) [29].

2 Related Work

2.1 Adversarial Attacks

Most research in adversarial attacks has been done for image recognition. In a seminal work, Szegedy et al. [39] showed that popular image benchmarks could be attacked using adversarial examples. They created adversarial examples by formulating the search for adversarial examples by minimizing the amount of perturbation required to give images the incorrect label. Similarly, DeepFool [31] perturbs images towards the hyperplane of the closest class. Other methods can use gradient information, such as the Fast Gradient Sign Method (FGSM) [9], Basic Iterative Method (BIM) [22], Carlini and Wagner attacks (CW) [5], and Projected Gradient Descent (PGD) [29]. Furthermore, there are many other methods of attacking image-based neural networks [25].

In comparison, there are fewer studies on adversarial attacks and time series recognition. Carlini et al. [4] revealed that it is possible to encode hidden commands in speech recognition systems. Fawaz et al. [7] evaluate FGSM and BIM on time series classification datasets. There also has been work in creating adversarial examples in time series classification without neural networks [33].

2.2 Defense Against Attacks

A defense algorithm is designed to improve the robustness of a neural network against an attacker. There are many ways to defend against adversarial attacks and each has varying levels of success [25]. For example, it is possible to improve robustness by training with adversarial examples [9]. However, these methods require knowledge about the attacker. Another way is to limit the effects of the adversarial perturbation such as using defensive distillation [14], feature squeezing [43], or denoising [26]. It is also possible to use ensemble and modular networks to avoid attacks trained against specific gradients [37, 11, 27, 44].

As for TTA and defense against adversarial attacks, there are a few works that use similar methods for image-based attacks. For example, Cohen and Giryes [6] propose Augmented Random Forest (ARF), which includes the use of TTA with a neural network feature extractor to make more robust Random Forest classifiers. Perez et al. [34] use a combination of horizontal flipping and different image cropping to build ensemble classifiers to increase robustness.

2.3 Adversarial Attacks and Online Handwriting

Adversarial attacks and handwriting is not a new field. However, attacks on handwriting are typically performed on image-based offline handwriting. For example, Jiang et al. [19] show that training with Chinese character images already attacked with PGD helps build models more robust to attacks. Bayram and Barner [1] propose the Efficient Combinatorial Black-box Adversarial Attack (ECoBA) for binary image classifiers, specifically for optical character recognition (OCR) systems. Attacks have also recently been performed on offline signature verification [13, 24, 18, 2].

Unlike offline handwriting, there are very few studies on attacks on online handwriting. Lopresti and Raim [28] tested the effectiveness of generative attacks on online handwriting biometric authentication. Specifically, they attacked hash-based online handwriting using a concatenation of letters to generate passphrases.

Compared to these methods, as far as we know, we are the first to use adversarial attacks on online handwritten character classification. We also are the first to construct a defense for online handwritten characters.

3 Adversarial Attacks

3.1 Threat Model

In this paper, we focus on *white-box* attacks. Namely, given neural network model $f(\mathbf{x})$, where $\mathbf{x}$ is the input, a white-box attack has full information about $f(\cdot)$, $\mathbf{x}$, the parameters of $f(\cdot)$, the gradients of $f(\mathbf{x})$, etc. The goal of the attack is to find some adversarial sample $\mathbf{x}_{adv}$ that is similar to $\mathbf{x}$ yet misclassified. Furthermore, the similarity must be within budget ϵ.

3.2 Gradient-based Adversarial Attacks

Gradient-based adversarial attacks are white-box attacks that generate adversarial samples based on the gradient change of the neural network. For the attacks in the experiments, we use the four most popular gradient-based attacks, FGSM [9], BIM [22], PGD [29], and CW [5].

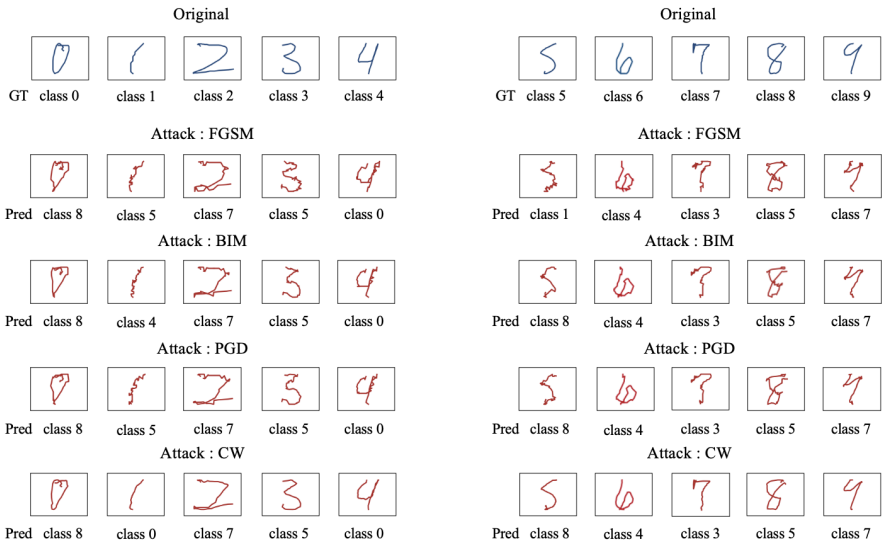

Fig. 2. Example comparison of the original characters and the attacked characters.

Figure 2 shows an example of online handwritten characters before the attacks and compares them to after each type of attack. The data after the FGSM, BIM, and PGD attacks are all perturbed in such a way that adversarial noise is added to each element of the input. On the other hand, the CW attack attempts to perturb the data so it is as close to the original data as possible. Nonetheless, in each case, the differences between the attacked data and the original is hardly noticeable to a human, but gives false predictions by a classifier.

Fast Gradient Sign Method (FGSM) Proposed by Goodfellow et al. [9], FGSM uses the gradient direction to create the adversarial noise. Namely, the adversarial sample $\mathbf{x}_{fgsm}$ is:

$$\mathbf{x}_{fgsm} = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, y)), \tag{1}$$

where $\mathcal{L}(\mathbf{x}, y)$ is the loss between the prediction of $\mathbf{x}$ and the prediction y, $\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, y)$ is the gradient of the loss, and $\text{sign}(\cdot)$ returns the positive or negative sign. The adversarial perturbation added to the original input $\mathbf{x}$ is defined as the constant ϵ in the direction of the gradient.

Basic Iterative Method (BIM) BIM [22] is an extension of FGSM that uses an iterative approach. Instead of perturbing at the constant ϵ in the one-step approach of FGSM, BIM adds adversarial perturbations α repeatedly, or:

$$\mathbf{x}_0 = \mathbf{x} \tag{2}$$

$$\mathbf{x}_i = (\mathbf{x}_{i-1} + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}_{i-1}} \mathcal{L}(\mathbf{x}_{i-1}, y))) \tag{3}$$

$$\mathbf{x}_{bim} = \mathbf{x}_I, \tag{4}$$

where I is the number of iterations.

Projected Gradient Descent (PGD) In principle, PGD [29] is similar to BIM, with multi-step perturbations. PGD generally applies clipping so that the attack area is within the original image area, making it possible to create examples that are more difficult to distinguish, or:

$$\mathbf{x}_i = \text{clip}_{\mathbf{x},\epsilon}(\mathbf{x}_{i-1} + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}_{i-1}} \mathcal{L}(\mathbf{x}_{i-1}, y))), \tag{5}$$

where $\text{clip}_{\mathbf{x},\epsilon}(\cdot)$ is a function ensures that the perturbations stay within ϵ.

Calini and Wagner Attack (CW) The CW attack proposed by Calini and Wagner [5] solves the optimization problem in the following equation with respect to perturbations to obtain stronger perturbations than FGSM and BIM, or:

$$\text{argmin}_{\mathbf{x}} \left|\mathbf{x} - \mathbf{x}_0\right|_2 + c \cdot f(\mathbf{x}), \tag{6}$$

where c is a positive constant and $f(\mathbf{x})$ is the objective function of the attack. The objective function is expressed in the form:

$$f(\mathbf{x}) = \max\left(\log\left(1 + \exp\left(z_i(\mathbf{x})\right)\right) - t_i, -\kappa\right), \tag{7}$$

where $z_i(\mathbf{x})$ is the score that the input $\mathbf{x}$ is output by the classifier i, t_i is the score for the classifier i's correct answer class, and κ is the limit value.

4 Test Time Augmentation as a Defense

4.1 Test Time Augmentation (TTA)

TTA is a technique used to enhance the performance of machine learning models [36]. The idea of TTA is to reduce overfitting aggregating predictions under different transformations during test time. In this way, during testing, multiple transformations are applied to a single test sample, and the average of their prediction is used. By considering multiple perspectives of the input data, TTA can provide more reliable predictions, especially in cases where there is a large variation in the samples.

There are many advantages of using TTA. Some of the reasons include:

- **Ease of use**. It only requires implementing augmentation methods and no change of the underlying model.
- **Cost-effectiveness**. TTA provides a cost-effective way to improve model performance without the need for retraining or modifying the model architecture. Instead, it leverages existing trained models and applies data augmentation techniques at the test time, making it computationally efficient and easy to integrate into existing workflows.
- **Flexibility**. TTA can be used with any model. Because the transformations are performed on the input, the ensemble can work with any classifier.
- **Mitigation of overfitting**. By introducing different augmented versions of the test data, the classifier has a reduced risk of using memorized patterns present in the training data.
- **Increased performance**. TTA can be seen as a form of ensemble learning, where predictions from multiple augmented versions of the test data are combined to obtain a final prediction. Ensemble methods often lead to improved performance compared to individual models, as they leverage diverse perspectives and sources of information.

4.2 Proposed Use of TTA

In order to add robustness to trained models and protect against adversarial attacks, TTA is used during inference time, as shown in Fig. 3. Under the threat model, we aim to protect against adversarial attacks, yet maintain a high accuracy on non-attacked data. The idea is that it should not be possible to know if the network is being attacked or not. In general, white-box attacks have knowledge of the trained network and use specific and intentional adversarial noise to exploit the gradients of the network. Accordingly, by applying transformations on the inputs, the goal of the proposed method is to disrupt the adversarial noise.

Transformations As shown in Fig. 3, in the proposed method, four transformations are performed in order to generate diverse predictions via TTA. The following transformation methods are used on input sequence $\mathbf{x} = x_1, \ldots, x_t, \ldots, x_T$ with T number of time steps and where x_t is a two-dimensional element of the handwritten character representing a spatial coordinate.

- **Jittering** [3]. Random noise is added to each element from a Gaussian distribution, or:

$$\mathbf{x}_{\text{jittering}} = x_1 + \epsilon_1, \ldots, x_t + \epsilon_t, \ldots, x_T + \epsilon_T, \tag{8}$$

  where $\epsilon \sim \mathcal{N}(\mu, \sigma^2)$. We add Gaussian noise with a mean $\mu = 0$ and standard deviation $\sigma = 0.03$ to the original time series.
- **Window Slicing** [23]. In window slicing, a window of 90% of the original time series is randomly chosen, or:

$$\mathbf{x}_{\text{window slicing}} = x_i, \ldots, x_t, \ldots, x_{.9T+i}, \tag{9}$$

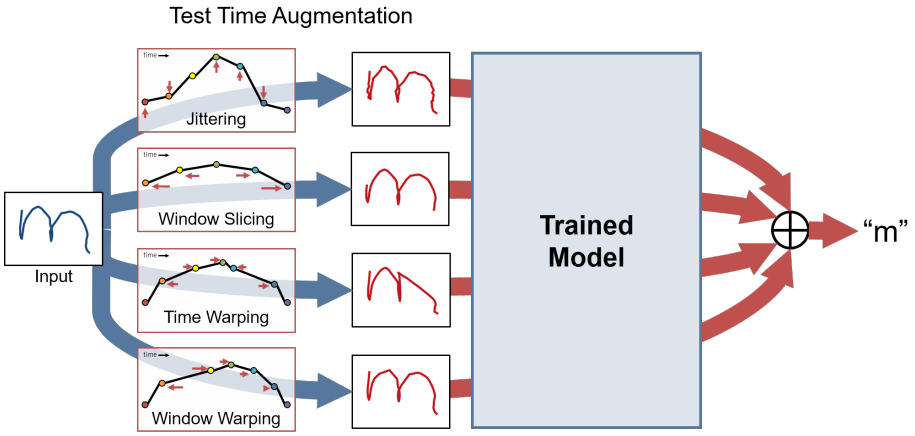

Fig. 3. The proposed use of TTA for defense against adversarial attacks

where $.9T$ is 90% of the time series and i is a random start point given $1 \leq i \leq .1T$. Our implementation interpolates this back to the original size to fit the classifier.

– **Time Warping** [40]. Time warping distorts the time axis based on a randomly smooth warping curve generated by a cubic spline with four knots of random size ($\mu = 1, \sigma = 0.2$), or:

$$\mathbf{x}_{\text{timewarping}} = x_{\phi(1)}, \ldots, x_{\phi(t)}, \ldots, x_{\phi(T)}, \tag{10}$$

where $\phi(\cdot)$ is the random warping function applied to the time steps based on the smooth curve with perturbations in time. Notably, this method does not modify the spatial coordinates of the handwriting, only the time steps at which the coordinates occur.

– **Window Warping** [23]. Window Warping is a variation of time warping that uses a random window with a random $2\times$ or $0.5\times$ multiplier to warp the time steps. The window size is 10% of the original time series length. Similar to Window Slicing, the length is resampled to work with the fixed-sized neural network.

Ensembling It should be noted that there are different ways to ensemble the predictions from each augmentation. For example, it is possible to use voting, averaging, and summing. For our proposed method, we use the sum of the predictions from each of the augmentations, or:

$$\hat{y} = \underset{c \in C}{\text{argmax}} \sum_{n}^{N} P(c|\mathbf{x}_n), \tag{11}$$

where $\hat{y}$ is the prediction, c is a class in C, $\mathbf{x}_n$ is an input online handwritten character under augmentation method n of N number of total methods, and $P(c|\mathbf{x}_n)$ is the probability of the class, i.e. the post-softmax prediction.

5 Experimental Results

5.1 Datasets

To evaluate the proposed method, experiments were conducted on three online handwriting datasets from the International Unipen Foundation [12]. The datasets consist of numerical digits (Unipen 1A), uppercase alphabet (Unipen 1B), and lowercase alphabet (Unipen 1C). For pre-processing, the characters were normalized to be 50 time steps. We split the datasets into a training set of about 11,000 patterns, a test of 1,300 patterns for the digits, and 1,250 patterns for the alphabets. However, the purpose of splitting the data is not to obtain state-of-the-art results, but to set a good model to attack and defend against.

5.2 Architecture and Training Settings

For the experiments, we use a temporal CNN as the backbone for the defense methods. The temporal CNN has four 1D convolutions, each with Batch Normalization [15], a rectified linear unit (ReLU) activation, and is followed by max pooling. In the first block, 64 filters are used, and the subsequent blocks have 128 filters. After the convolutional layers, two fully connected layers are used. The first fully connected layer has 512 nodes with ReLU activation and the second is the output layer with the number of nodes equal to the number of classes and softmax activation. Between the two fully connected layers, dropout with 0.5 probability is used.

To train all of the networks, Adam optimizer [20] is used with an initial learning rate of 0.001. The network is trained for 10,000 iterations with a batch size of 256. A single CNN is trained and five test sets, an unmodified test set, one with FGSM, one with BIM, one with PGD, and one with CW are used. The trained CNN is used with the defense methods only during test time.

5.3 Adversarial Attack Settings

For the experiments, the network is trained with the training set, and the test set is attacked with adversarial noise. All of the experiments use the same trained model with different attacks on the test set. The FGSM, BIM, PGD, and CW attacks were used as adversarial attack methods. For FGSM and BIM, maximum distortion $\epsilon = 0.2$ is used and for BIM and PGD, step size $\alpha = 0.05$ for $I = 10$ iterations is used.

5.4 Evaluated Defenses

We compare the proposed usage of TTA against other defenses to evaluate the robustness of the defense. The following models were used in the evaluation:

- **No Defense**. This is the baseline to show how effective the adversarial attack is on the trained model.
- **Random Noise**. In an attempt to cancel or obfuscate the adversarial noise, random Gaussian noise can be added to the input [37]. For the noise, a standard deviation of $\sigma = 0.15$ is used.
- **Median Filter**. Another way to quell adversarial noise is the use of a median filter. The median filter smooths the trajectories of the characters. In the experiment, the filter size was set to 3 due to the adversarial attacks using element-wise noise.
- **Random Self-Ensemble (RSE)** [27]. RSE is an image-based defense against adversarial attacks. RSE adds a noise layer to every convolutional layer in a CNN. The noise layer allows for different predictions similar to TTA. We adapted the RSE to be used for time series, namely online handwritten characters. For a fair comparison, we use a self-ensemble with five networks. For the experiments, a standard deviation of $\sigma = 0.15$ was used.
- **Test Time Augmentation (Proposed)**. The input is transformed using the aforementioned transformations, jittering, window slicing, time warping, and window warping, and test time predicted using the same network. The predictions are combined as proposed. The parameters used in the experiments are standard parameters for data augmentation, suggested by Iwana and Uchida [16].

5.5 Results

Accuracy in the no-attack case is shown in Table 1. In the ideal case, the defense method should not degrade the accuracy of the unmodified test set. The experimental results show that the accuracy remains comparable to that of the CNN under normal conditions for all datasets. This is a good result because, in the problem set, it is unknown whether an attack is taking place or not. Thus, it is important to maintain the accuracy on normal data. It should be noted, that RSE, the defense designed to protect against adversarial attacks, has the lowest accuracy of all the defenses. This means sacrificing the accuracy on clean handwriting for the robustness on attacked handwriting.

Next, the FGSM, BIM, PGD, and CW attacks were performed, and the results are shown in Table 2. For this table, a robust method would close the gap between the accuracy of the No Defense case under attack and the accuracy in Table 1. For all attacks, all datasets show an improvement in accuracy with the proposed method over the No Defense.

The CW attack is the most difficult attack in this study because it is the most misrecognized method despite having the smallest amount of noise. However, the proposed method shows significant improvement over other comparative metrics against the CW attack.

Table 1. Accuracy (%) Without Attacks

Defense	Unipen 1A	Unipen 1B	Unipen 1C
No Defense	99.4	98.4	97.5
RSE	97.3	93.4	94.5
Random Noise	99.2	98.3	97.5
Median Filter	99.3	98.3	**97.6**
Proposed	**99.5**	**98.5**	97.4

5.6 Ablation

The proposed method utilizes four augmentations. We investigated their contributions in terms of accuracy and robustness against CW attacks. Table 3 shows how the proposed method performs when removing one of the augmentation methods. It shows that the removal of any of the augmentation methods does not have a significant effect on the accuracy when there is no attack. Furthermore, it demonstrates that all four augmentation methods are important for defending against the CW attack. Additionally, the three augmentations excluding jittering show high contributions to the defense. Window warping in particular protects against CW the most.

5.7 Analysis

Here, we examine the classifications while under a CW attack. A confusion matrix during the attack is depicted in Fig. 4. As evident from this figure, almost all samples are mispredicted due to the attack the attack, with the diagonal being near zero. One interesting thing about the figure is that it reveals which handwritten characters are easily confused. For example, the handwritten "0"s and "5"s get changed to an "8" and "3"s and "8"s get changed to "5"s. This is interesting because intuitively the time series representation of these numbers are similar.

Next, confusion matrices for the defense methods RSE and TTA are shown in Fig. 5. While RSE does help recover the accuracy, this figure demonstrates the instances where the proposed method performed better. The proposed TTA performed much better in instances such as "3" and "8" being misclassified as "5."

6 Conclusion

In this paper, we propose a defense mechanism against adversarial attacks on online handwritten characters using unconventional use of TTA. We employ TTA not as a means of data augmentation, but as a defense mechanism against adversarial attacks. Based on the experimental results, it is confirmed that the robustness of temporal neural networks against adversarial attacks on online

Table 2. Accuracy (%) Under Adversarial Attacks

Under a FGSM Attack			
Defense	Unipen 1A	Unipen 1B	Unipen 1C
No Defense	69.5	75.5	65.2
RSE	69.0	72.7	66.1
Random Noise	69.8	75.3	64.9
Median Filter	71.0	76.2	66.6
Proposed	**72.9**	**77.2**	**70.5**
Under a BIM Attack			
Defense	Unipen 1A	Unipen 1B	Unipen 1C
No Defense	44.5	53.7	39.4
RSE	**56.9**	58.7	**52.8**
Random Noise	44.9	54.7	40.6
Median Filter	48.3	58.0	44.0
Proposed	53.8	**63.4**	52.6
Under a PGD Attack			
Defense	Unipen 1A	Unipen 1B	Unipen 1C
No Defense	43.7	54.7	40.4
RSE	**55.6**	60.1	52.4
Random Noise	44.6	54.6	41.5
Median Filter	47.8	58.9	44.5
Proposed	53.9	**65.7**	**54.3**
Under a CW Attack			
Defense	Unipen 1A	Unipen 1B	Unipen 1C
No Defense	0.38	0.97	1.62
RSE	65.8	56.8	67.6
Random Noise	51.0	44.4	49.6
Median Filter	68.0	70.4	68.9
Proposed	**85.6**	**81.8**	**86.2**

Table 3. Ablation of the Proposed Method Under a CW Attack

Without Attack			
Defense	Unipen 1A	Unipen 1B	Unipen 1C
Proposed	**99.5**	**98.5**	97.4
without jittering	99.4	98.1	97.3
without window slicing	**99.5**	98.4	97.4
without time warping	**99.5**	98.2	**97.6**
without window warping	**99.5**	98.1	**97.6**
Under a CW Attack			
Defense	Unipen 1A	Unipen 1B	Unipen 1C
Proposed	**85.6**	**81.8**	**86.2**
without jittering	84.6	81.3	84.6
without window slicing	80.0	75.7	80.7
without time warping	78.6	74.5	81.2
without window warping	76.4	71.0	78.4

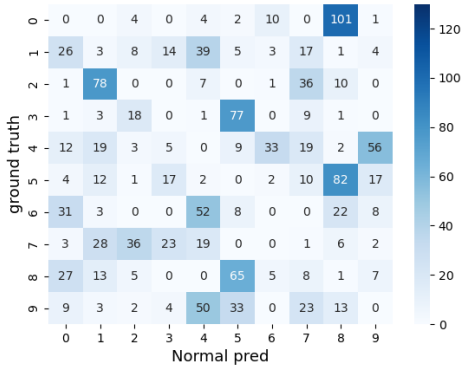

Fig. 4. Confusion matrices of the regular model with no defense during a CW attack.

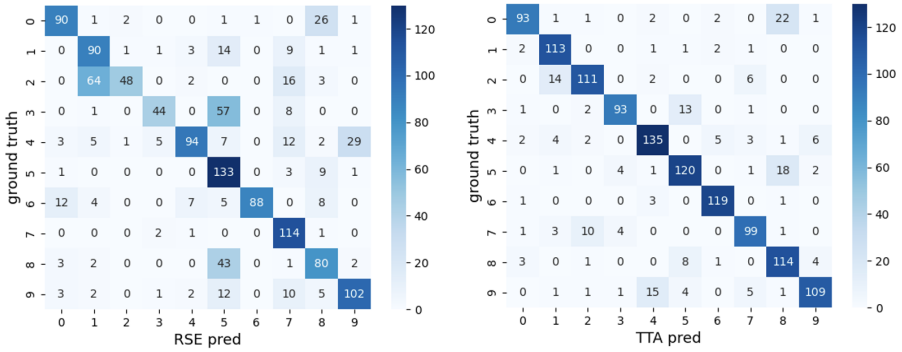

Fig. 5. Confusion matrices of RSE and the proposed method during CW attacks.

handwritten characters can be enhanced by TTA. In particular, significant performance improvements were demonstrated, especially against the most challenging CW attacks.

References

1. Bayram, S., Barner, K.: A black-box attack on optical character recognition systems. In: CVMI. pp. 221–231 (2023)
2. Bird, J.J., Naser, A., Lotfi, A.: Writer-independent signature verification; evaluation of robotic and generative adversarial attacks. Information Sciences **633**, 170–181 (2023)
3. Bishop, C.M.: Training with noise is equivalent to tikhonov regularization. Neural computation **7**(1), 108–116 (1995)

4. Carlini, N., Mishra, P., Vaidya, T., Zhang, Y., Sherr, M., Shields, C., Wagner, D., Zhou, W.: Hidden voice commands. In: USENIX. pp. 513–530 (2016)
5. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: IEEE SP. pp. 39–57. Ieee (2017)
6. Cohen, G., Giryes, R.: Simple post-training robustness using test time augmentations and random forest. In: WACV. pp. 3996–4006 (2024)
7. Fawaz, H.I., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A.: Adversarial attacks on deep neural networks for time series classification. In: IJCNN (2019). https://doi.org/10.1109/ijcnn.2019.8851936
8. Ghosh, T., Sen, S., Obaidullah, S., Santosh, K., Roy, K., Pal, U.: Advances in online handwritten recognition in the last decades. Computer Science Review **46**, 100515 (2022). https://doi.org/10.1016/j.cosrev.2022.100515
9. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
10. Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., Schmidhuber, J.: A novel connectionist system for unconstrained handwriting recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(5), 855–868 (2009)
11. Guo, C., Rana, M., Cisse, M., Van Der Maaten, L.: Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117 (2017)
12. Guyon, I., Schomaker, L., Plamondon, R., Liberman, M., Janet, S.: Unipen project of on-line data exchange and recognizer benchmarks. In: ICPR. https://doi.org/10.1109/icpr.1994.576870, http://dx.doi.org/10.1109/icpr.1994.576870
13. Hafemann, L.G., Sabourin, R., Oliveira, L.S.: Characterizing and evaluating adversarial examples for offline handwritten signature verification. IEEE Transactions on Information Forensics and Security **14**(8), 2153–2166 (2019)
14. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 **2**(7) (2015)
15. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML. pp. 448–456 (2015)
16. Iwana, B.K., Uchida, S.: An empirical survey of data augmentation for time series classification with neural networks. PLOS ONE (2021). https://doi.org/10.1371/journal.pone.0254841
17. Iwana, B.K., Frinken, V., Uchida, S.: Dtw-nn: A novel neural network for time series recognition using dynamic alignment between inputs and weights. Knowledge-Based Systems **188**, 104971 (2020). https://doi.org/10.1016/j.knosys.2019.104971
18. Jahangir, M., Malik, M.I., Shafait, F.: Adversarial attacks on convolutional siamese signature verification networks. In: ICDAR. pp. 350–365 (2023)
19. Jiang, G., Qian, Z., Wang, Q.F., Wei, Y., Huang, K.: Adversarial attack and defence on handwritten chinese character recognition. Journal of Physics: Conference Series **2278**(1), 012023 (2022). https://doi.org/10.1088/1742-6596/2278/1/012023
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
21. Kubatur, S., Sid-Ahmed, M., Ahmadi, M.: A neural network approach to online devanagari handwritten character recognition. In: HPCS (2012). https://doi.org/10.1109/hpcsim.2012.6266913
22. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236 (2016)
23. Le Guennec, A., Malinowski, S., Tavenard, R.: Data augmentation for time series classification using convolutional neural networks. In: IWAATD (2016)

24. Li, H., Li, H., Zhang, H., Yuan, W.: Black-box attack against handwritten signature verification with region-restricted adversarial perturbations. Pattern Recognition **111**, 107689 (2021). https://doi.org/10.1016/j.patcog.2020.107689
25. Liang, H., He, E., Zhao, Y., Jia, Z., Li, H.: Adversarial attack and defense: A survey. Electronics **11**(8), 1283 (2022). https://doi.org/10.3390/electronics11081283
26. Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., Zhu, J.: Defense against adversarial attacks using high-level representation guided denoiser. In: CVPR. pp. 1778–1787 (2018)
27. Liu, X., Cheng, M., Zhang, H., Hsieh, C.J.: Towards robust neural networks via random self-ensemble. In: ECCV. pp. 381–397 (2018). https://doi.org/10.1007/978-3-030-01234-2_23
28. Lopresti, D.P., Raim, J.D.: The effectiveness of generative attacks on an online handwriting biometric. In: ICAVBPA. pp. 1090–1099 (2005)
29. Madry, A., Makelov, A., Schmidt, ., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
30. Mezghani, N., Mitiche, A., Cheriet, M.: On-line recognition of handwritten arabic characters using a kohonen neural network. In: IWFHR. https://doi.org/10.1109/iwfhr.2002.1030958
31. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: DeepFool: A simple and accurate method to fool deep neural networks. In: CVPR (2016). https://doi.org/10.1109/cvpr.2016.282
32. Nguyen, H.T., Nguyen, C.T., Nakagawa, M.: Online japanese handwriting recognizers using recurrent neural networks. In: ICFHR (2018). https://doi.org/10.1109/icfhr-2018.2018.00082
33. Oregi, I., Ser, J.D., Perez, A., Lozano, J.A.: Adversarial sample crafting for time series classification with elastic similarity measures. In: IDC. pp. 26–39 (2018). https://doi.org/10.1007/978-3-319-99626-4_3
34. Pérez, J.C., Alfarra, M., Jeanneret, G., Rueda, L., Thabet, A., Ghanem, B., Arbeláez, P.: Enhancing adversarial robustness via test-time transformation ensembling. In: ICCV. pp. 81–91 (2021)
35. Schmidhuber, J.: Deep learning in neural networks: An overview. Neural networks **61**, 85–117 (2015)
36. Shanmugam, D., Blalock, D., Balakrishnan, G., Guttag, J.: When and why test-time augmentation works. arXiv preprint arXiv:2011.11156 (2020)
37. Strauss, T., Hanselmann, M., Junginger, A., Ulmer, H.: Ensemble methods as a defense to adversarial perturbations against deep neural networks. arXiv preprint arXiv:1709.03423 (2017)
38. Sun, L., Su, T., Liu, C., Wang, R.: Deep lstm networks for online chinese handwriting recognition. In: ICFHR (2016). https://doi.org/10.1109/icfhr.2016.0059
39. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
40. Um, T.T., Pfister, F.M.J., Pichler, D., Endo, S., Lang, M., Hirche, S., Fietzek, U., Kulić, D.: Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In: ACM ICMI. pp. 216–220 (2017). https://doi.org/10.1145/3136755.3136817
41. Wang, Z., Yan, W., Oates, T.: Time series classification from scratch with deep neural networks: A strong baseline. In: IJCNN. pp. 1578–1585 (2017). https://doi.org/10.1109/ijcnn.2017.7966039

42. Wei, W., Guanglai, G.: Online handwriting mongolia words recognition with recurrent neural networks. In: ICCSCIT. IEEE (2009). https://doi.org/10.1109/iccit.2009.197
43. Xu, W., Evans, D., Qi, Y.: Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155 (2017)
44. Yu, Y., Yu, P., Li, W.: Auxblocks: defense adversarial examples via auxiliary blocks. In: IJCNN. pp. 1–8 (2019)