Adversarial Attacks on Convolutional Siamese Signature Verification Networks

Maham Jahangir[1], Muhammad Imran Malik[1], and Faisal Shafait[1,2]

[1] School of Electrical Engineering and Computer Science (SEECS), National University of Sciences & Technology (NUST), Islamabad, Pakistan.
[2] Deep Learning Laboratory, National Center of Artificial Intelligence (NCAI), Islamabad, Pakistan

Abstract. A handwritten signature serves as an important biometric modality to identify individuals. The state-of-the-art methods for signature verification employ deep learning networks to perform the classification task. However, deep neural networks can be fooled by adversarial attacks that introduce small imperceptible perturbations to the input images. In this paper, we explore the vulnerability of signature verification systems against adversarial attacks. The state-of-the-art attacks developed by the machine learning community to fool image classifiers are unsuitable for attacking document classifiers as they are applied to the background of signature images making them quite perceptible. To overcome this challenge, we design an attack based on dictionary learning with the goal to perturb the foreground (strokes) of the signature image. The proposed method is evaluated in terms of attack success rate and imperceptibility. The experimental results on the benchmark CEDAR dataset using Siamese Deep Signet Model highlight the efficacy of the proposed approach as compared to other methods by achieving 95% and 98% attack success rates with our proposed approach.

Keywords: Adversarial Attack · Sparse Encoding · Dictionary Learning · Signature Verification.

1 Introduction

Biometric Systems are widely used to recognize individuals in legal, financial, and administrative matters [7, 15]. Handwritten signatures serve as one such biometric which are required especially during financial transactions to identify and verify an individual. The Signature verification systems can be offline (static) and online (dynamic). The offline systems identify individuals from a signature image (spatial information) containing handwriting strokes whereas the online system's recognition is based on the signature generation process (considering spatial and temporal information). Offline systems are used widely due to low cost and convenience. Moreover, there are scenarios where offline signature verification is inevitable for example during cheque transactions. The traditional systems relied on handcrafted features for signature verification but lately, most

of the research efforts on offline signature verification systems are based on deep neural networks. These systems work under two approaches a) writer independent and b) writer dependent. The writer-independent approach is generally considered more practical as the systems based on writer dependent approach need to be updated every time a new writer is registered [4]. This research article also considers writer-independent offline signature verification scenarios to evaluate the robustness of signature verification systems.

We have used SigNet: Convolutional Siamese Network [3] in this study. The available data is divided into train and test sets with a couple of image pairs such as (genuine, genuine) and (genuine, forged) labeled as, similar and dissimilar classes. Siamese networks can efficiently model such problems. Siamese networks are based on twin convolutional networks which accept two images that can either be similar or dissimilar. Since Deep Neural Networks (DNNs) are employed here for signature verification, unfortunately, DNNs are vulnerable to adversarial examples [19]. These examples are generated by imposing carefully crafted perturbations to clean input images. This research area gained quick popularity since its advent [19] and a lot of attacks have been proposed to exploit the vulnerabilities of deep neural network-based systems. However, attacking signature images is a relatively different and challenging task when compared to other fields. The vulnerability of signature verification systems against adversarial attacks has not been explored thoroughly and only a handful of research is available on the topic. In this article, we present the first attempt to particularly attack Siamese network-based signature verification system.

It should be noted that attacking verification systems is very different from attacking classification systems and presents challenges not present in classification systems. First, when a new user gets registered a new unseen class and unseen examples are introduced to the system. Second, for signature verification systems the background and foreground are clearly separated and a verification system clearly uses the foreground information (strokes) to extract features and then classify the image as genuine or forged. The state-of-the-art attacks impose perturbations on the background making them perceptible and since background information is not used by the system, therefore, the attack success rate is greatly reduced. Further, in the model used in this article, the images are inverted during pre-processing making it even harder to attack. The third problem is that most of the state-of-the-art methods specifically gradient-based methods applied to signature verification systems are white box in nature (they require full information on the training set, the model used, and parameters learned in order to attack a system). These systems are well protected by organizations and such information is unknown to attackers. So traditional white-box attack methods are not practical.

In view of the above-mentioned problems, this research article proposes a black-box attack method to attack signature verification systems using ideas from sparse representation. Our recent work explored the idea of dictionary learning to craft sparse adversarial attacks for image classification [8]. Formally, we used the idea of sparse representation to craft adversarial images using feature

maps of an image. In this research, we have extended the idea and developed a novel approach to learn a dictionary on feature descriptor (foreground extraction) and improved sparse representation to create adversarial attacks with the goal to perturb the foreground (strokes) of the signature image. The sparse representation includes dictionary learning and sparse coding stages to generate perturbations that can be induced in the original images making them adversarial. Dictionary learning is a transformation process that transforms an image to its linear combination of basic elements called atoms. Sparse Coding is a method for learning a sparse representation of the input using dictionary learning [13]. In this paper, a novel feature descriptor approach is used to learn the dictionary and improve sparse representation quality. In this regard, we used the Grab cut algorithm [18] to extract the foreground of the signature images and then learn the dictionary. This is an attempt to learn only important and relevant information. The proposed technique is evaluated on the benchmark publicly available CEDAR Signature Dataset and is also compared with the state-of-the-art methods.

The main contributions and findings include:

1. The proposed model generates adversarial perturbations to fool signature verification systems with minimum ℓ_2-norm and maximum attack success rates of 95% and 98% respectively.
2. We introduce improved sparse representation quality by learning a dictionary on a feature descriptor (foreground extraction) rather than original unprocessed images.
3. We attacked a convolution-based Siamese network for a handwriting signature verification system not attacked before.
4. Our experiments show that attacking strokes of signature is important as attacks on the background won't produce desirable results.

The structure of the paper is as follows. Section 2 describes the related works. Section 3 details the problem, threat model, and methodology of the proposed approach. Section 4 defines the experimental protocol. Section 5 presents experimental results and analysis. Section 6 concludes the paper.

2 Related Work

Adversarial examples are manipulated input images with perturbations that fool the classifiers. The concept of adversarial attacks was introduced by Szegedy et al. [19] in 2013. Since then a lot of attacks have been proposed by the machine learning community to evaluate the robustness of deep networks. Among the pioneers is Fast Gradient Sign Method (FGSM) [5]. This is a gradient-based method that maximizes the loss of the classifier to craft adversarial examples. Later iterative methods like Deep Fool [17], Basic Iterative method (BIM) [9], and Carlini and Wagner (C&W) [2] were also introduced. Universal adversarial attacks create a single adversarial perturbation that fools the classifier with high probability and generalizes well across different neural networks [16].

Projected Gradient Descent [12] is a well-optimization method essentially similar in behavior to iterative FGSM with the difference that it initializes the input sample to a random point in the ball of interest. On the other hand, Boundary Attack [1] is one of the decision-based attacks which follows the decision boundary between adversarial and non-adversarial examples using a simple rejection sampling algorithm.

In the context of adversarial attacks against signature verification systems Hafemann [6] explored the vulnerability of these systems against adversarial attacks. They attacked the system using existing adversarial attacks, like FGSM and C&W and presented two types of threats to these systems hence two types of attacks. Type: I, where an adversary manipulates a genuine signature to be misclassified by the system (False Rejection). Type: II where a forged signature is manipulated to be classified as genuine by the systems (False Acceptance). The authors point out that Type: I attacks are easy to generate as compared to Type: II. These perturbations were introduced on the background of the images making them quite perceptible and requiring perfect knowledge of the system under attack which is not practical. In another research, Li et al. [10] proposed a gradient-free black-box attack against signature verification systems by restricting the area of perturbations to the region of strokes. Their attack method is not applicable to binary images as the perturbation intensity of each pixel is not continuously adjustable. Therefore, selecting optimal pixels for perturbations will not be possible.

To the best of the authors' knowledge, these two research articles explored the vulnerability of signature verification systems against adversarial attacks. This area still needs a lot of exploration and presents great room for improvement. None of the above-mentioned researchers tested their proposed methods on Siamese Networks. Attacking Siamese networks is much more challenging than other classification systems. It is evident from the results section that state-of-the-art attack methods couldn't attack these networks efficiently. The attack success rates of the state-of-the-art are quite low when compared with literature where they showed good performance while they attacked other signature verification systems. Siamese Networks are widely used and acquired state-of-the-art performance on signature verification systems. That is why they have gained fast-growing popularity in signature verification systems. These systems serve the rightful purpose of comparing the images and then identifying them as genuine or forged based on their similarity or dissimilarity. Therefore, in this research, we explored the vulnerability of the Convolutional Siamese Networks against adversarial attacks. We designed a black-box attack (information on the training set, the model used by a verifier, and parameters learned are not required) based on the sparse representation of foreground features of images. The experimental results prove the efficacy of the proposed method.

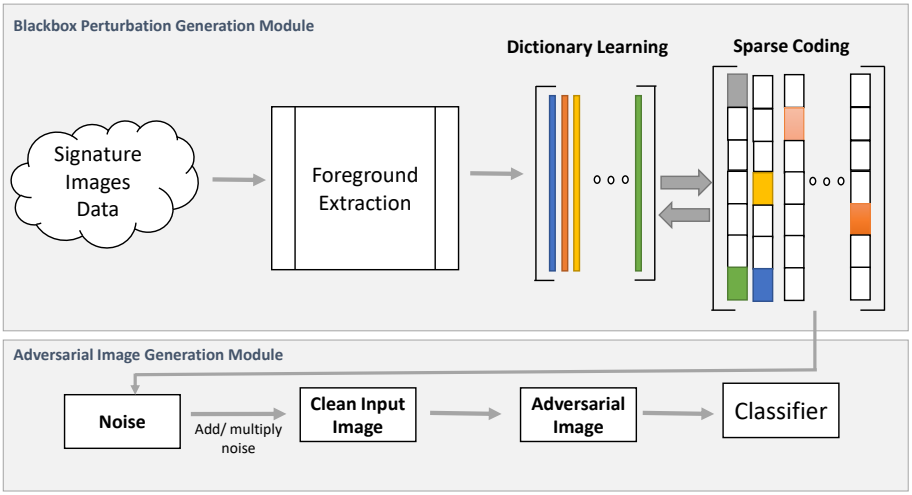


Fig. 1. General Framework of the Proposed Attack Model

3 Methodology

This section explains our methodology in detail as illustrated in Figure 1. The first step in the proposed approach is to extract the foreground from the signature images. These images are fed to a dictionary learning algorithm to learn sparse representation. The sparse representation is then used as a perturbation to manipulate the original input image to fool the classifier. Below we discuss the Siamese Network under attack, followed by the problem statement, foreground extraction, sparse representation, and adversarial image generation.

3.1 Siamese Network

In this research, we evaluated the robustness of the Siamese network named: Signet [3] against adversarial attacks. The Siamese networks are very popular among signature verification systems and to the best of our knowledge are not yet studied for robustness against adversarial attacks. One of the reasons behind their popularity is their ability to learn from minimum data. They need only a few images to make better predictions and data is not abundant in various problems including signature verification [14]. The Siamese networks are based on twin CNN architectures with shared weights joined at the output by a loss function. The goal is to find similarities between the two images. They learn a feature space when similar observations are placed in proximity and are used to evaluate whether a given signature is genuine or forged. This is achieved by exposing the network to both similar and dissimilar pairs and the network maximizes the Euclidean distance between dissimilar pairs whereas minimizes

the distance between similar pairs. The popular loss function used by Siamese networks is contrastive loss and is defined as follows:

$$L(a, b, y) = \alpha(1 - y)D_w^2 + \beta y \max(0, m - D_w)^2 \quad \text{where} \quad a, b \in X \quad (1)$$

a and b are input samples that belong to the set X. They can be genuine signatures or forged entries in the system. y is a binary indicator that indicates whether the given two signatures belong to the same class or not. α and β are two constants whereas, m indicates the margin i.e. 1 in this case. $D_w = \| f(a; w_1) - f(b; w_2) \|_2$. It is the Euclidean distance computed in feature space, f is a function that maps a signature image to its real vector space through CNN whereas, w_1 and w_2 are learned weights of that particular layer of the network. The training of Siamese networks involves pairwise learning so the classifier won't output probabilities of the prediction but the distance from each class. We have reported this distance in our experiments of the proposed approach as well as for the state-of-the-art methods. The threshold of 0.5 is selected to determine if the output of the Siamese network is the same or not.

3.2 Problem

A typical Siamese-based offline signature verification model under attack is depicted in Figure 2. The model takes signature images as input. These signature images can be genuine – by authentic users or can be forgeries – entered into the system by a skilled forger. The forgers generate signature images that resemble original images from the same user in an attempt to fool the system. Since the system is trained on skilled forgeries as well, Signature verification systems successfully recognize the forgeries. However, these systems are still vulnerable to two main threats. First, an original authentic signature image can be modified in a way that system rejects the original image that is **Type: I, False Rejection (FR)**. The second form of attack is the one in which the forged signature images are modified in a way that gets accepted by the system termed as **Type: II, False Acceptance (FA)**. Some previous researchers consider that the second type of adversarial attack is harder to generate [6, 10] as compared to the first one. However, in the case of Siamese networks, our experiments show that Type: I attacks are harder to generate. In this paper, we considered both of these adversarial attacks for evaluation purposes. Adversarial examples are images similar to the true data distribution but fool the system. These images are generated by adding small perturbations to the original data. If we denote X as input space and a function $F(X)$ maps these input to a label Y then the adversarial examples X_{adv} that are visually similar to clean samples X_{org} but fools the classifier that is $F(X_{\text{adv}} \neq Y)$. In the case of the Siamese network

$$L(a, b_{\text{adv}}) \neq y \quad (2)$$

where,

$$b_{\text{adv}} = b + \epsilon p \quad \text{and} \quad d(b_{\text{adv}}, b) < \epsilon \quad (3)$$

where, ϵ is the magnitude of perturbation p added in the image. The distance d between original signature image b and adversarial image b_{adv} should be minimum.

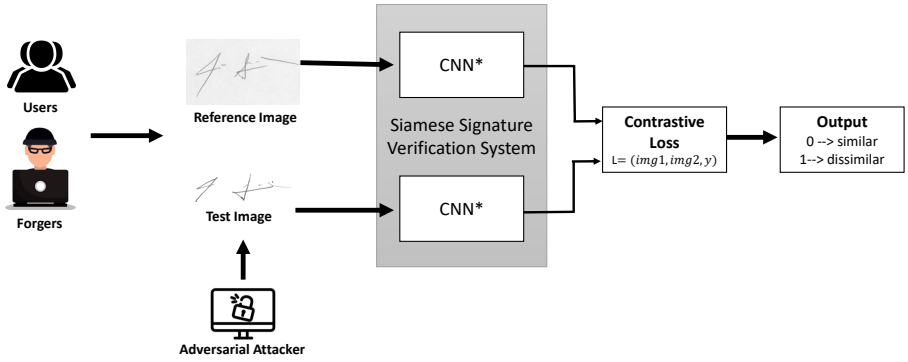


Fig. 2. Siamese Network-based Signature Verification System and Threat Model

3.3 Foreground Extraction

The first step in our proposed approach is to extract the foreground of the signature image. The foreground contains the signature strokes. Our goal is to learn a dictionary on these strokes as they are the only important and relevant information that we need from the signature image. Background doesn't hold any detail in signature verification systems. In order to learn specific features we intend to learn the dictionary on the foreground of the image rather than the full image. The background pixels are changed to 0-pixel value whereas, the foreground to 1. Let pixels covering the foreground be denoted as F_d and that of the background as B_d.

$$X^{'} = F_d + B_d \quad \text{where,} \quad F_d = 1 \ \text{ and } \ B_d = 0 \tag{4}$$

For the above-mentioned purpose, we used the GrabCut algorithm [18] to extract the foreground of the image which can be used to learn the dictionary and its corresponding sparse representation. It is a graph cuts-based image segmentation method. It uses a Gaussian mixture model to separate the background and the target object.

3.4 Sparse Representation (Dictionary Learning and Sparse Coding)

The next step is to learn the sparse representation of the processed images from the last section. The foreground extraction serves as an important feature descriptor to improve the quality of learned representations. The goal is to improve

the feature descriptor of the signature images by keeping specific and minimal information. Sparse coding is an encoding process where a sparse representation of input images is learned using a linear combination of basic elements. these elements are called atoms and they combine to form a dictionary. Let $X^{'}$ denote the foreground extracted images from the previous step. A transformation operator to learn sparse representation is applied to it and denoted as $T(X^{'})$. The optimization function to learn dictionary and sparse representation proposed by Mairal et al. [13] and is given as

$$T(X^{'}) = D\alpha \tag{5}$$

$$\min_{D,\alpha} \frac{1}{2} \parallel x^{'} - D\alpha \parallel_2^2 + \lambda \parallel \alpha \parallel_1 \quad \text{s.t.} \parallel D_k \parallel_2 = 1 \ \ \forall \ \ k \in [0, n] \tag{6}$$

were, $x^{'}$ is the pre-processed signature image and λ is a regularization parameter, α is the sparse representation, D is the dictionary learned, and n is the number of dictionary atoms. The algorithm explaining the steps of this section is listed in Algorithm: 1.

Algorithm 1: Adversarial Dictionary Learning

Input: $X^{'} \rightarrow$ Set of pre-processed original signature images;
Result: D $\rightarrow$ Learned Dictionary , $T(X^{'}) \rightarrow$ Sparse representation
$D \rightarrow$ Initial Dictionary ;
$OMP \rightarrow$ Orthogonal Matching Pursuit() ;
$k \rightarrow$ Sparsity ;
$n \rightarrow$ no. of atoms ;
for $t = 1$ *to iterations* **do**
  $T(X^{'}) \leftarrow OMP(D, X^{'})$;
  Dictionary Update Stage;
  $D = \min_{D,\alpha} \frac{1}{2} \parallel x^{'} - D\alpha \parallel_2^2 + \lambda \parallel \alpha \parallel_1 \qquad s.t. \parallel D_k \parallel_2 = 1 \ \ \forall \ \ k \in [0, n]$;
Return D **Return** $T(X^{'})$

3.5 Tuned adversarial signature image generation

This is the final stage where an adversarial image is generated. A dictionary of perturbations is learned and saved by the dictionary learning algorithm as discussed above. These perturbations have a different effect on the attack success rate. So in this step, the adversarial signature image is tuned for all the available perturbations. The perturbations that maximize the loss of the classifier and achieve the highest attack success rate are selected. The complete process of adversarial image generation involving all sections is defined step by step in Algorithm 2. The first step is to extract the foreground of signature images. For Type: I attack the forged samples of images are used to learn the dictionary

whereas, genuine samples in the case of Type: II attacks. Next, we learn the dictionary and compute sparse representation. This sparse representation is basically our noise/perturbation to be used to manipulate the original image. As we discussed earlier, contrary to some findings in literature the Type: I attack was much more challenging than anticipated in the case of Siamese networks. With reference to Siamese networks, the additive noise model couldn't attack the genuine image to be declared as forged by the classifier. Therefore, inspired by recent work on multiplicative noises [11] we multiplied the noise perturbation with the original image to craft our adversarial example. The experimental results prove the effectiveness of multiplicative noise over additive. Detailed analysis of multiplicative and additive noises for the Type: I attack is discussed in Section 5.

Algorithm 2: Tuned Adversarial Signature Image Generation

Result: $X_{adv} \rightarrow$ Tuned Adversarial Image
Input: $X_{org} \rightarrow$ legitimate source input image;
$X_{forg} \rightarrow$ skilled forged signature input image;
$\epsilon \rightarrow$ magnitude of noise ;
$L \rightarrow$ classifier's loss;
if $attack = type : I$ **then**
  $X^{'} = Grabcut(X_{forg})$;
else
  $X^{'} = Grabcut(X_{org})$;
$T(X^{'}) = DictLearningAlgo(X^{'})$;
$P = T(X^{'})$;
for $i < size(X_{org})$ **do**
  if $attack = type : I$ **then**
    $\max_{L(X_{org}, X_{adv}, Y)} X_{adv\,i} = X_{org\,i} * \epsilon P_i$;
  else
    $\min_{L(X_{org}, X_{adv}, Y)} X_{adv\,i} = X_{org\,i} + \epsilon P_i$;
Return X_{adv}

4 Experimental Protocol

The experimental design and detail to evaluate the proposed methodology are discussed in this section.

4.1 Dataset

We conducted the experiments on the widely used benchmark signatures dataset, CEDAR signature Database[3]. We have used this dataset as it is quite well-

[3] http://www.cedar.buffalo.edu/NIJ/data/signatures.rar

Table 1. Attributes of CEDAR dataset used in the experiments to define training and test splits. Note that the splits were carefully done in a way that the users in dictionary learning, training Siamese network, and testing were mutually exclusive.

Attributes	Count
Number of users	55
Users in the training set	28
Users in the test set	12
Users to train the dictionary	15
Genuine signatures per user	24
Forgeries per user	24

known and used by almost all the articles we reviewed during this research. Moreover, it contains signatures of 55 users from different ethnic and professional backgrounds. Each user signed 24 genuine signatures with a difference of 24 minutes in between. Forgers copied the signatures of 3 genuine users, 8 times each. Hence, each user has 24 genuine and 24 forged signatures. A total of $55 \times 24 = 1320$ genuine and 1320 forged signatures are available in this dataset. The total number is $1320 \times (2) = 2640$. These images are available in grayscale mode.

We divided the dataset into training and test sets as shown in Table: 1. The system is trained and tested using signatures from 40 users with a train test split of 70% : 30%. We also reserved some signature images which were not part of the training or testing of the model. This allows us to define a black-box attack scenario to evaluate our approach where the attacker has no access to the training or test data or the model used by the signature verification system. The remaining signatures from 15 users are used to simulate the environment where an attacker has a dataset of his own with some genuine signatures by users and the respective forgeries. These images are used to train the dictionary and learn sparse representations. These sparse representations are added as perturbations to the test set of the dataset to create adversarial examples.

4.2 Pre-processing and Performance of Signet-Siamese Network

The model is trained and tested as per the guidelines outlined in the paper [3]. The same pre-processing steps are employed. The publicly available implementation of the model architecture is used to carry out the training[4]. The images are resized to a fixed size (155×220) and then inverted to get a black background with pixel values: 0. Finally, all the images are normalized. The detail on the Siamese network has been provided in Section 3.1. We trained the network for 80 epochs. The training loss equal to 0.3 and accuracy of 85% are calculated respectively. The test loss and accuracy were 0.015 and 97% respectively.

[4] https://github.com/AtharvaKalsekar/SigNet/

4.3 Metrics

The contrastive loss of the classifier, attack success rate, and mean and median ℓ_2-norm are calculated during experimentation. The attack success rate defines the number of genuine signatures that failed to pass through the system and the number of forged signatures that successfully passed through the system. The ℓ_2-norm is a standard method to compute the length of a vector in Euclidean space. We use it to find the similarity between two images. Here it is the squared distance between the adversarial and original clean image. A lower distance means that the two images appear the same and the noise in adversarial images is imperceptible. We have calculated the mean and median values of ℓ_2-norm.

4.4 State-of-the-art Adversarial Attacks

We compared our approach with state-of-the-art methods. The adversarial robustness toolbox[5] was used to conduct experiments for the state-of-the-art. We evaluated the proposed systems against Fast Gradient Sign Method (FGSM) [5], Basic Iterative Method (BIM [9], Projected Gradient Descent (PGD) [12], and Boundary Attack Method [1]. These are all baseline attack methods that achieved state-of-the-art attack success rates in traditional image classification systems. These systems are gradient-based evasion attacks that are white-box in nature (where the attacker has access to the training or test data or the model used by a signature verification system). Epsilon ϵ refers to the magnitude of noise introduced to the original clean image to create an adversarial image. Our proposed method relies on a very small magnitude of noise in order to attack the system. The other state-of-the-art methods don't attack the system at all if the ϵ is kept very low. Therefore, we cannot test the system for the same values of ϵ. We have used $\epsilon = 0.3$ for the state-of-the-art to conduct the experiments.

5 Results and Discussion

This section explains the results reported when the proposed approach is applied to the CEDAR signature dataset and compared with the state-of-the-art methods. Moreover, the effect of perturbations on strokes of signatures images is discussed with reference figures and examples.

5.1 Type: I Attack (False Rejection)

In this attack, perturbation is applied to genuine signatures images such that the system fails to verify the image as genuine. Contrary to the popular opinion in the literature where attacking genuine signatures(Type: I) is argued to be an easy task, we found the Type: I attack to be equally challenging as that of Type: II specifically in the case of Siamese networks. Since the model pre-processes the image where the background is black and the signature strokes are white. This

[5] https://adversarial-robustness-toolbox.readthedocs.io/en/latest/

makes it hard to add noise to the strokes. The background noise fails to attack the system. This is evident from results tabulated in Table 2. Only the proposed method is able to attack successfully with a success rate of 95% and with the lowest ℓ_2-norm value of 0.09. The first row of Figure 3 illustrates the example images generated through our proposed approach as well as the state-of-the-art. It can be clearly seen that almost all baseline methods attack the background of the image, therefore, their attack success rates are very low, and ℓ_2-norm is quite high.

Table 2. The magnitude of noise ϵ, Loss of Classifier (higher the value more successful the attack is), Attack Success Rate, Mean and median ℓ_2-norm (lower the value more imperceptible the attack is) values reported for Type: I attack for our proposed method and state-of-the-art.

Method	**Epsilon(ϵ)**	**Loss**	**Attack Succ. (%)**	**Mean ℓ_2-norm**	**Median ℓ_2-norm**
FGSM [5]	0.3	0.19	29	0.37	0.37
BIM [9]	0.3	0.05	8	0.13	0.12
PGD [12]	0.3	0.04	7	0.13	0.12
Boundary Attack [1]	-	0.01	2	0.42	0.43
Proposed	**0.002**	**1.50**	**95**	**0.09**	**0.09**

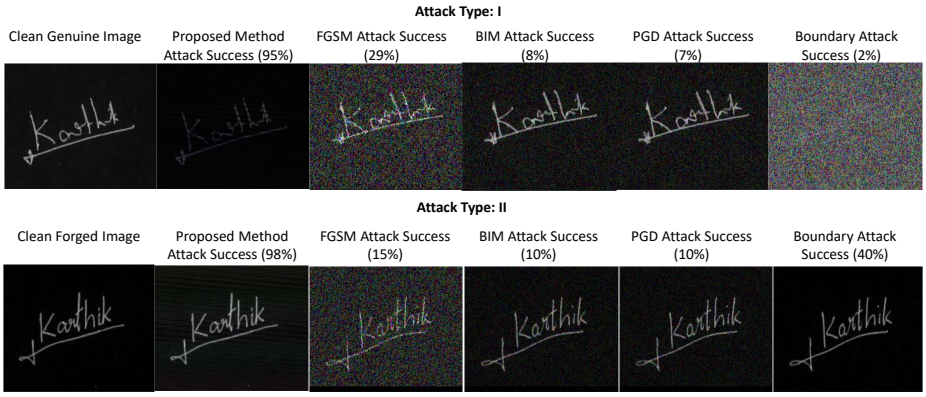


Fig. 3. Clean and Adversarial Image examples from the results of experiments reported in Table:2 and Table:3 for Type-I and Type-II attacks

5.2 Type: II Attack (False Acceptance)

In this attack, the perturbation is applied to forged signature images such that the system accepts them during the verification and classifies them as genuine

which was previously declared forged by the system. The results for this type are tabulated in Table 3. The proposed approach successfully attacks the system with an attack success rate of 98% using a very low magnitude of noise $\epsilon = 0.0004$. The ℓ_2-norm is also the lowest among all baseline methods which is 0.07. The second row of Figure 3 illustrates the example images of the proposed method and all other methods. Again the other methods fail to attack the system significantly as they attack the background of the image except for the Boundary Attack. Nevertheless, its attack success rate is still very low (attack success rate of 40%, and the ℓ_2-norm of 0.17) compared to the proposed method (attack success rate of 98%, and the ℓ_2-norm of 0.07).

Table 3. The magnitude of noise ϵ, Loss of Classifier (lower the value more successful the attack is), Attack Success Rate, Mean and median ℓ_2-norm (lower the value more imperceptible the attack is) values reported for Type: II attack for our proposed method and state-of-the-art.

Method	**Epsilon(ϵ)**	**Loss**	**Attack Succ. (%)**	**Mean ℓ_2-norm**	**Median ℓ_2-norm**
FGSM	0.3	0.88	15	0.36	0.36
BIM	0.3	1.27	10	0.12	0.12
PGD	0.3	1.27	10	0.12	0.12
Boundary Attack	-	0.92	40	0.17	0.17
Proposed Method	0.0004	0.01	98	0.07	0.07

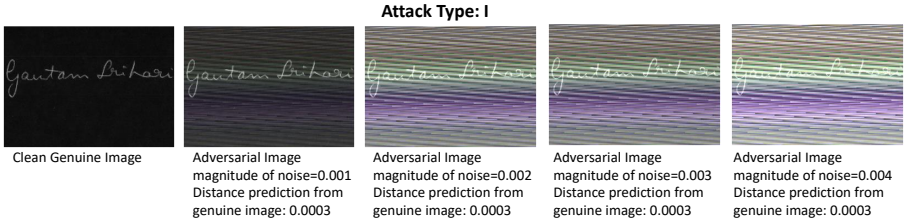
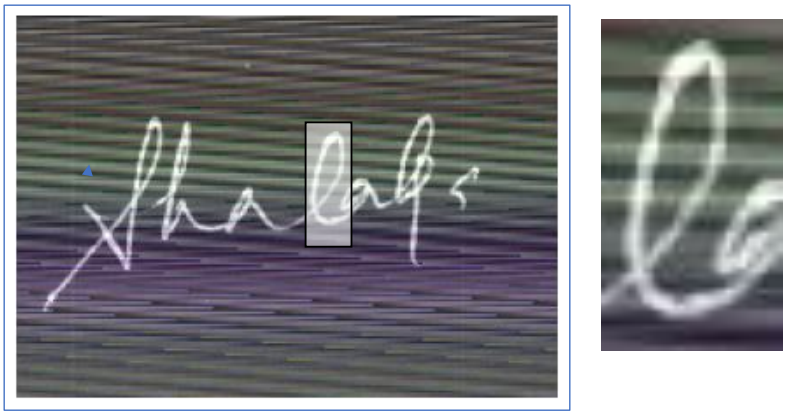


Fig. 4. Effect of magnitude of noise on the prediction of the model in case of Type: I attack

5.3 Effect of magnitude of noise on the prediction of the model on genuine signatures images

As discussed above, Type: I attack, which were generally considered as an easy target [6, 10], have been proven challenging while attacking the Siamese network. Figure 4 illustrates the effect of the increasing magnitude of noise on the genuine signatures. It can be seen that even increasing the magnitude of noise causes no effect on the prediction of the model. It still declares the image as genuine. This



(a) Signature adversarial image with additive noise –minimum to no perturbations on strokes

(b) Signature adversarial image with multiplicative noise – perturbations on strokes

Fig. 5. Additive and Multiplicative Noise Adversarial Example Images with same values of epsilon and their effect on strokes of the signature image

is because the strokes of the images remain intact and the model only used the information of strokes to learn features and classify them.

5.4 Effect of multiplicative and additive noise on genuine signatures

We have used multiplicative noise in the case of the Type: I attack for the proposed method. Figure 5 illustrates how multiplicative noise attacks the strokes of the signature image while additive noise just disrupts the background. We have shown a zoomed version of the portion of the stroke to illustrate our point. This is why we chose multiplicative noise rather than popular additive noise to craft our adversarial examples.

6 Conclusion

In this research, we attacked a convolutional Siamese signature verification network using sparse representation and dictionary learning. A novel algorithm to learn a dictionary from an important feature descriptor that extracts foreground is proposed. The attack proposed is black-box in nature that doesn't require information about the signature verification model used, its weights, or training or test data. The experimental results show that our proposed method outperforms all the baseline methods and achieves attack success rates of 95% and 98% for Type: I and Type: II adversarial attacks, respectively.
In the future, we will test our proposed method with more datasets and evaluate its performance for transferability across other deep networks. We shall also evaluate our proposed approach against defense methods. The improvement of sparse representation quality in terms of improved feature descriptors should be studied too.

References

1. Brendel, W., Rauber, J., Bethge, M.: Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. arXiv preprint arXiv:1712.04248 (2017)
2. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy. pp. 39–57. IEEE (2017)
3. Dey, S., Dutta, A., Toledo, J.I., Ghosh, S.K., Lladós, J., Pal, U.: SigNet: Convolutional siamese network for writer independent offline signature verification. CoRR **abs/1707.02131** (2017), http://arxiv.org/abs/1707.02131
4. Diaz, M., Ferrer, M.A., Impedovo, D., Malik, M.I., Pirlo, G., Plamondon, R.: A perspective analysis of handwritten signature technology. ACM Computing Surveys **51**(6), 1–39 (2019)
5. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
6. 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)

7. Hameed, M.M., Ahmad, R., Kiah, M.L.M., Murtaza, G.: Machine learning-based offline signature verification systems: A systematic review. Signal Processing: Image Communication **93**, 116139 (2021)
8. Jahangir, M., Shafait, F.: Adversarial attack using sparse representation of feature maps. IEEE Access **10**, 120724–120734 (2022)
9. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial examples in the physical world. In: Artificial Intelligence Safety and Security, pp. 99–112. Chapman and Hall/CRC (2018)
10. 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)
11. Lo, S.Y., Patel, V.M.: Multav: Multiplicative adversarial videos. In: 2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance. pp. 1–6. IEEE (2021)
12. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
13. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online dictionary learning for sparse coding. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 689–696 (2009)
14. Malik, J., Elhayek, A., Ahmed, S., Shafait, F., Malik, M.I., Stricker, D.: 3DAirSig: A framework for enabling in-air signatures using a multi-modal depth sensor. Sensors **18**(11), 3872 (2018)
15. Malik, M.I., Liwicki, M., Dengel, A.: Part-based automatic system in comparison to human experts for forensic signature verification. In: 2013 12th International Conference on Document Analysis and Recognition. pp. 872–876. IEEE (2013)
16. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1765–1773 (2017)
17. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2574–2582 (2016)
18. Rother, C., Kolmogorov, V., Blake, A.: GrabCut interactive foreground extraction using iterated graph cuts. ACM Transactions on Graphics **23**(3), 309–314 (2004)
19. 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)