ROBUSTIFICATION OF DEEP NET CLASSIFIERS BY KEY BASED DIVERSIFIED AGGREGATION WITH PRE-FILTERING

Olga Taran, Shideh Rezaeifar, Taras Holotyak, Slava Voloshynovskiy

Department of Computer Science, University of Geneva, 7 Route de Drize, Carouge GE, Switzerland
{olga.taran, shideh.rezaeifar, taras.holotyak, svolos}@unige.ch

ABSTRACT

In this paper, we address a problem of machine learning system vulnerability to adversarial attacks. We propose and investigate a Key based Diversified Aggregation (KDA) mechanism as a defense strategy. The KDA assumes that the attacker (i) knows the architecture of classifier and the used defense strategy, (ii) has an access to the training data set but (iii) does not know the secret key. The robustness of the system is achieved by a specially designed key based randomization. The proposed randomization prevents the gradients' back propagation or the creating of a "bypass" system. The randomization is performed simultaneously in several channels and a multi-channel aggregation stabilizes the results of randomization by aggregating soft outputs from each classifier in multi-channel system. The performed experimental evaluation demonstrates a high robustness and universality of the KDA against the most efficient gradient based attacks like those proposed by N. Carlini and D. Wagner [1] and the non-gradient based sparse adversarial perturbations like OnePixel attacks [2].

Index Terms— Adversarial attacks, black / gray-box, non-gradient / gradient based attacks, defense, machine learning.

1. INTRODUCTION

Deep Neural Networks (DNN) are used to solve a wide range of problems including the classification tasks. Despite the outstanding performance and remarkable achievements, the DNN systems have recently been shown to be vulnerable to *adversarial attacks* [3]. The adversarial attacks aim at tricking a decision of DNN classifiers by introducing carefully designed perturbations to a chosen target image. These perturbations, being usually quite small in magnitude and imperceptible, can drastically change the output of the classifier. This weakness seriously questions the usage of the DNN based systems in many security- and trust-sensitive applications.

In the recent years, the number of authors reported various adversarial attacks against the DNN classifiers. The diversity of discovered attacks is quite broad but without loss of generality one can cluster all attacks into three groups [4, 5]: (1) *white-box* attacks, (2) *gray-box* attacks and (3) *black-box attacks*. The *white-box* attacks assume that the attacker has a full access to the trained model and training data. Despite a big popularity of this group of attacks, their applicability to the real-life systems is questionable. The *gray* and *black-box* scenarios are more suited to the real-life applications. The *gray-box* attacks assume that the attacker has certain knowledge about the trained model but there exist some secret elements or an access to the intermediate results is limited. The *back-box* attacks allow the attacker only to observe the output of classifier to each input without any knowledge about used architecture or possibility to observe the internal states.

The existing defense mechanisms are also quite diverse [6, 5]. However, the growing number of defenses leads to a natural invention of new and even more universal attacks. In the overwhelming majority of cases, the main interest in the adversarial attack investigation is focused on the gradient based attacks and defenses. While the non-gradient attacks and suitable defenses receive less attention but are not less dangerous and important for practice. In this respect, the goal of our paper is to investigate a new family of defense strategies that can be applied for both gradient and non-gradient based adversarial attacks in *gray* and *black-box* scenarios. We name it a Key based Diversified Aggregation (KDA) with pre-filtering. The generalized diagram of the proposed system is illustrated in Figure 1. The main idea behind the KDA is to use cryptographic principles and to create an information advantage for the defender over the attacker. A secret is shared between the training and classification stages. The secret is implemented in a form of secret key used for the randomization. The system has two levels of randomization, each of which uses its own secret keys. The classification process is diversified in several channels with own randomization targeting specific randomly selected features. To reduce the negative effect of randomization, the soft outputs of multi-channels classifiers are aggregated.

The main contribution of this paper is twofold:

- A multi-channel classification architecture with the KDA mechanism as an universal defense strategy

S. Voloshynovskiy is a corresponding author. The work was supported by the SNF project No. 200021_182063.

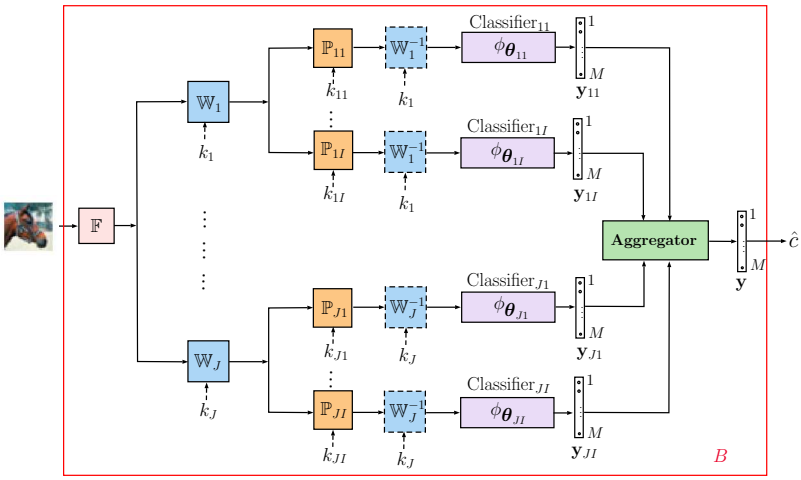

Fig. 1: Generalized diagram of the proposed multi-channel KDA with pre-filtering.

against the gradient and non-gradient based *gray* and *black-box* attacks.

- An investigation of the efficiency of the proposed approach on the well-known gradient and non-gradient based adversarial attacks.

The rest of paper is organized as follows. Section 2 introduces a new multi-channel classification architecture with the KDA. The efficient key-based data independent transformation is proposed in Section 3. Section 4 presents the empirical results obtained for the proposed algorithm. Finally, Section 5 concludes the paper.

2. CLASSIFICATION ALGORITHM WITH KEY BASED DIVERSIFIED AGGREGATION

The generalized diagram of the proposed algorithm is shown in Figure 1 and it consists of five main blocks:

1. *Pre-filtering* $\mathbb{F}$ that has an optional character. The goal of this block is to remove high magnitude outliers in the input images introduced by the attacker, if any. One can choose a broad range of pre-filtering algorithms from a simple local mean filter to more complex algorithms as, for example, BM3D [7] or based on DNN systems [8].

2. The input signal mapping into a *transform domain* via $\mathbb{W}_j$, $1 \leq j \leq J$. In general, the transform $\mathbb{W}_j$ can be any linear mapper like a random projection or belong to the family of orthonormal transformations ($\mathbb{W}_j \mathbb{W}_j^T = \mathbb{I}$) like DFT (discrete Fourier transform), DCT (discrete cosines transform), DWT (discrete wavelet transform), etc. Moreover, $\mathbb{W}_j$ can also be a learnable transform or even a deep encoder. However, to avoid any key-leakage from the trained transforms, we use the data independent transform $\mathbb{W}_j$ in this paper. Thus, the transform $\mathbb{W}_j$ is generated from a secret key k_j. Along this line, one can also envision, for example, the DCT transform with the key defined sampling in the transform domain. We will detail below the properties of this transform.

3. *Data independent processing* $\mathbb{P}_{ji}$, $1 \leq i \leq I$ serves as a defense against gradient back propagation to the direct domain. As simple examples of such kind of processing one can mention a lossy sampling $\mathbb{P}_{ji} \in \{0,1\}^{l \times n}$ $l < n$ of the input signal of length n as considered in [9] or a lossless permutation $\mathbb{P}_{ji} \in \{0,1\}^{n \times n}$ similar to [6]. The sub-block sign flipping $\mathbb{P}_{ji} \in \{-1,0,+1\}^{n \times n}$ presents an additional option. It should be pointed out that to make the *data independent processing* irreversible for the attacker, it is preferable to use the block $\mathbb{P}_{ji}$ based on a secret key k_{ji}.

4. *Classification block* can be represented by any family of classifiers. We consider a DNN based family.

5. *Aggregation block* can be any operation ranging from a simple summation to learnable operators or special aggregation networks adapted to the data or to a particular adversarial attack. We focus on additive aggregation to demonstrate the power of a simple strategy leaving the investigation of more complex aggregations to our future work.

As shown in Figure 1, in the proposed architecture the principal blocks are organized in a parallel multi-channel structure that can be followed by one or several *aggregation blocks*. The final decision is made based on the aggregated result. The rejection option can naturally be also envisioned.

It should be pointed out that the access to the intermediate results inside the considered system provides the attacker a possibility to use the full system as a *white-box*. The attacker can discover the secret keys k_j and/or k_{ji}, make the the system end-to-end differentiable using the Backward Pass Differentiable Approximation technique [10] or via replacing the key based blocks by the bypass mappers. Therefore, it is important to restrict the access of the attacker to the intermediate results within the block B (see Figure 1). That satisfies our assumption about *gray* and *black-box* attacks. Additionally, it is in the accordance with the Kerckhoffs's cryptographic prin-

(a) sub-bands (b) original (c) V (d) H (e) D

Fig. 2: Local key based sign flipping in the DCT sub-bands.

ciple when we assume that the algorithm and architecture are known to the attacker besides the used secret key.

The training of the described classification architecture can be performed as follows:

$$(\hat{\boldsymbol{\vartheta}}, \{\hat{\boldsymbol{\theta}}_{\boldsymbol{ji}}\}) = \underset{\boldsymbol{\vartheta}, \{\boldsymbol{\theta}_{ji}\}}{\arg\min} \sum_{t=1}^{T} \sum_{j=1}^{J} \sum_{i=1}^{I_j} \mathcal{L}(\mathbf{y}_t, A_{\boldsymbol{\vartheta}}(\phi_{\boldsymbol{\theta}_{ji}}(f(\mathbf{x}_t)))), \tag{1}$$

with:

$$f(\mathbf{x}_t) = \mathbb{W}_j^{-1} \mathbb{P}_{ji} \mathbb{W}_j \, \mathbb{F}(\mathbf{x}_t),$$

where $\mathcal{L}$ is a classification loss, $\mathbf{y}_t$ is a vectorized class label of the sample $\mathbf{x}_t$, $A_{\boldsymbol{\vartheta}}$ corresponds to the aggregation operator with parameters $\boldsymbol{\vartheta}$, $\phi_{\boldsymbol{\theta}_{ji}}$ is the ith classifier of the jth channel, $\boldsymbol{\theta}$ denotes the parameters of the classifier, T equals to the number of training samples, J is the total number of channels and I_j equals to the number of classifiers per channel j that we will keep fixed and equals to I for all channels.

3. RANDOMIZATION USING KEY BASED SIGN FLIPPING IN THE DCT DOMAIN

The core element of the defense in the proposed multi-channel architecture shown in Figure 1 is a data independent processing $\mathbb{P}$ in a transform domain $\mathbb{W}$.

In our implementation, we use the DCT as a $\mathbb{W}$ and the local sign flipping $\mathbb{P}_{ji} \in \{-1, 0, 1\}^{n \times n}$ based on the individual secret key k_{ji} for each classifier. The term *local* means that the processing is done only in some sub-band or block of the input signal. In general, the signal can be split into overlapping or non-overlapping sub-bands of different sizes and different positions that are kept in secret. In our experiments for the simplicity and interpretability, we split the signal in the DCT domain into four non-overlapping fixed sub-bands of the same size denoted as: (*L*) top left that represents the low frequencies of the image, (*V*) vertical, (*H*) horizontal and (*D*) diagonal sub-bands as illustrated in Figure 2a. The key based sign flipping is applied independently in *V*, *H* and *D* sub-bands keeping all other sub-bands unchanged. The effects of such processing after the inverse DCT transform are perceptually almost unnoticeable and exemplified in Figure 2c - 2e.

The corresponding multi-channel architecture is illustrated in Figure 3. For simplicity, as an aggregation operator A we use a simple summation. For the pre-filtering $\mathbb{F}$ we use a filter based on a difference of the point of interest in the

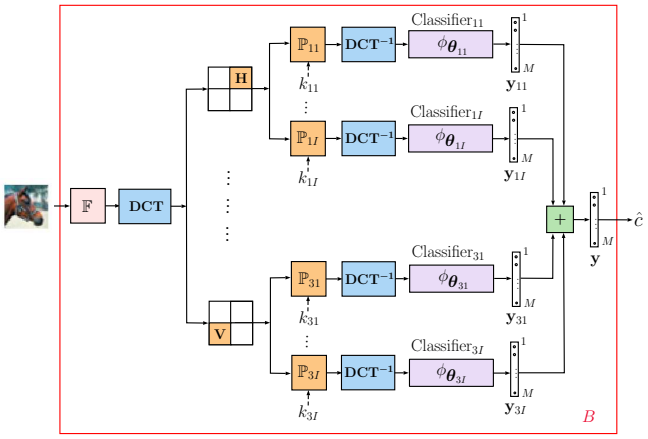

Fig. 3: Classification with the local DCT sign flipping.

center of the window with the median value in the window of size 3×3 around this point. If the magnitude of difference exceeds a specified threshold, the pixel is considered to be corrupted by the adversary and its value is replaced by a mean value computed in the window or otherwise, it is kept intact. Finally, each classifier $\phi_{\boldsymbol{\theta}_{ji}}$ is trained independently as:

$$\hat{\boldsymbol{\theta}}_{ji} = \underset{\boldsymbol{\theta}_{ji}}{\arg\min} \sum_{t=1}^{T} \mathcal{L}(\mathbf{y}_t, \phi_{\boldsymbol{\theta}_{ji}}(\mathbb{W}^{-1} \mathbb{P}_{ji} \mathbb{W} \, \mathbb{F}(\mathbf{x}_t))), \tag{2}$$

to ensure the best recognition in each channel under the introduced perturbation.

4. EXPERIMENTAL RESULTS

4.1. Setup

The efficiency of the proposed multi-channel architecture diversified and randomized by the key based sign flipping in the DCT domain against the adversarial attacks was tested for two scenarios:

1. *Gray-box* gradient based attack. As a gradient based attack we use the attack proposed in [1]. This attack is among the most efficient attacks against many proposed defense strategies. Further it will be referred to as C&W. In our experiment we use the C&W attacks based on ℓ_2, ℓ_0 and ℓ_∞ norms.

2. *Black-box* non-gradient based attack. As a non-gradient based attack we use the OnePixel attack proposed in [2] that uses a Differential Evolution (DE) optimisation algorithm [11] for the attack generation. The DE algorithm doesn't require the objective function to be differentiable or known but instead it observes the output of the classifier used as a black box. The OnePixel attack aims at perturbing limited number of pixels in the input image. In our experiments, we use this attack to perturb 1, 3 and 5 pixels.

For the fair comparison, the gradient based attacks were tested on the classifier with the architecture identical to those

Attack type	Vanilla	$J \cdot I$		
		3	6	9
Original	21	21.2	19.6	19.4
C&W ℓ_2	100	22.42	21.3	21.04
C&W ℓ_0	100	24.58	23.52	23.03
C&W ℓ_∞	100	22.8	21.39	21.21

Table 1: Classification error (%) on the first 1000 test samples against the *gray-box* gradient-based attacks.

tested in [1]. The non-gradient based attacks were tested for the classifiers based on the VGG16 [12] an ResNet18 [13] architectures used in [11].

All experiments have been done on the CIFAR-10 dataset [14] that presents a particular interest as a data set with the images close to natural ones. The CIFAR-10 consists of 60000 colour images of size 32×32 (50000 train and 10000 test) with 10 classes. Due to the fact that the attack generation process is sufficiently slow for all considered attacks the experimental results were obtained on the first 1000 test samples.

4.2. Empirical results and analysis

The results obtained for the gradient based attacks in the *gray-box* scenario are given in Table 1. The results obtained for the non-gradient based attacks in the *black-box* scenario are shown in Table 2. In both cases the column "vanilla" corresponds to the accuracy of the original classifier without any defense. The row "original" corresponds to the use of non-attacked original data.

The analysis of the obtained results for the *gray-box* gradient based attacks and the original non-attacked data demonstrates that the use of the proposed multi-channel architecture allows to improve the classification accuracy of vanilla classifier. This is quite remarkable by itself since it shows that the multi-channel processing with the aggregation does not degrade the performance due to the introduced randomization in contrast to many defense strategies based on gradient obfuscation or detection and rejection of attacked data mechanisms. In the case of attacked data, the C&W attacks achieves the 100% classification error on the vanilla undefended classifier thus showing a complete vulnerability of this deep classifier. At the same time, the use of the multi-channel architecture based on the same type of classifier with the proposed defense strategy improves the classification accuracy to the similar level of the vanilla classifier on the original non-attacked data. In the worst case of C&W ℓ_0 attack, the classification error is only about 2% higher than on the original data.

In the case of *black-box* non-gradient based attacks, the use of the KDA improves the classification accuracy of the vanilla classifier similar to the previous case. In contrast to the *gray-box* scenario, the *black-box* attacks are not so harmful against the vanilla classifier. In the case of VGG16, the classification error of the vanilla classifier is about 60-80%.

Attack type	Vanilla	$J \cdot I$		
		3	6	9
VGG16				
Original	10.7	11	9.2	8.9
OnePixel $p = 1$	58.04	11	9.5	8.7
OnePixel $p = 3$	72.13	10.9	8.9	8.3
OnePixel $p = 5$	79.02	12.1	9.3	9.1
ResNet18				
Original	9.5	11.1	9.1	7.8
OnePixel $p = 1$	36.96	11.5	9	7.7
OnePixel $p = 3$	49.85	11.5	9.1	7.8
OnePixel $p = 5$	59.74	11.7	9.2	7.8

Table 2: Classification error (%) on the first 1000 test samples against the *black-box* non-gradient based attacks.

In the case of ResNet18, it is about 35-60%. For both classifiers the increase of the number of perturbed pixels (p) leads to the increase of classification error. The use of the proposed defense mechanism based on the KDA architecture allows to decrease the classification error to the level of classification on the original data or in other words it diminished the effect of these attacks.

In summary, one can conclude that the obtained results indicate that the KDA architecture with the proposed defense strategy demonstrates the high robustness to the gradient and non-gradient based attacks in the *gray* and *black-box* scenarios. Moreover, it allows to improve the classification accuracy of the vanilla classifiers. Finally, it should be pointed out that in all cases the increase of the number of classification channels and *data independent processing* $\mathbb{P}_{ij}$ leads to improving the classification accuracy. However, a trade-off between the further decrease of the classification error and the increase of the complexity of the algorithm should be carefully addressed that goes beyond the scope of this paper.

5. CONCLUSIONS

In this paper, we considered the defense mechanism against the gradient and non-gradient based *gray* and *black-box* attacks. The proposed mechanism is based on the multi-channel architecture with the randomization and the aggregation of classification scores. It is remarkable that the architecture of the defense is not tailored for each class of attacks and is uniformly used for both attacks. It is also interesting to note that the diversified classification with the aggregation of the outputs of classifiers allows not only to withstand the attacks but it also improves the accuracy of vanilla classifier. It is also important to remark that the proposed approach is compliant with the cryptographic principles when the defender has an information advantage over the attacker. In our future research, we plan to extend the aggregation mechanism to more complex learnable strategies instead of used summation.

6. REFERENCES

[1] Nicholas Carlini and David Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 39–57.

[2] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, 2019.

[3] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2015.

[4] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Siwei Li, Li Chen, Michael E Kounavis, and Duen Horng Chau, "Shield: Fast, practical defense and vaccination for deep learning using jpeg compression," *arXiv preprint arXiv:1802.06816*, 2018.

[5] Naveed Akhtar and Ajmal Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *arXiv preprint arXiv:1801.00553*, 2018.

[6] Olga Taran, Shideh Rezaeifar, and Slava Voloshynovskiy, "Bridging machine learning and cryptography in defence against adversarial attacks," in *Workshop on Objectionable Content and Misinformation (WOCM), ECCV2018*, Munich, Germany, September 2018.

[7] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on image processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[8] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of machine learning research*, vol. 11, no. Dec, pp. 3371–3408, 2010.

[9] Zhipeng Chen, Benedetta Tondi, Xiaolong Li, Rongrong Ni, Yao Zhao, and Mauro Barni, "Secure detection of image manipulation by means of random feature selection," *CoRR*, vol. abs/1802.00573, 2018.

[10] Anish Athalye, Nicholas Carlini, and David Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning*, Jennifer Dy and Andreas Krause, Eds., Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018, vol. 80 of *Proceedings of Machine Learning Research*, pp. 274–283, PMLR.

[11] Rainer Storn and Kenneth Price, "Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces," *Journal of global optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[12] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton, "The cifar-10 dataset," *online: http://www. cs. toronto. edu/kriz/cifar. html*, 2014.