INTRODUCING K-ANONYMITY PRINCIPLES TO ADVERSARIAL ATTACKS FOR PRIVACY PROTECTION IN IMAGE CLASSIFICATION PROBLEMS

Vasileios Mygdalis, Anastasios Tefas and Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, 54124, Greece
Email: {mygdalisv,tefas,pitas}@csd.auth.gr

ABSTRACT

The network output activation values for a given input can be employed to produce a sorted ranking. Adversarial attacks typically generate the least amount of perturbation required to change the classifier label. In that sense, generated adversarial attack perturbation only affects the output in the 1st sorted ranking position. We argue that meaningful information about the adversarial examples i.e., their original labels, is still encoded in the network output ranking and could potentially be extracted, using rule-based reasoning. To this end, we introduce a novel adversarial attack methodology inspired by the K-anonymity principles, that generates adversarial examples that are not only misclassified, but their output sorted ranking spreads uniformly along K different positions. Any additional perturbation arising from the strength of the proposed objectives, is regularized by a visual similarity-based term. Experimental results denote that the proposed approach achieves the optimization goals inspired by K-anonymity with reduced perturbation as well.

***Index Terms*—** K-anonymity, Adversarial Attacks

1. INTRODUCTION

Adversarial attacks in classification tasks aim to generate the minimum amount of perturbation required to be added to the inputs, in order to fool a trained task classifier. Several classification models based on deep-learning, including Convolutional Neural Networks (CNNs), have been found to be vulnerable to adversarial attacks [1, 2]. Furthermore, recent studies [3, 4] have shown that carefully crafted adversarial examples may deceive various classification methods at the same time, ranging from similar deep architectures to even totally different classification methods, e.g., Support Vector Machines or Random Forests, rendering adversarial attacks a strong weapon for adversaries to by-pass automated classification systems.

This work has received funding from the European Unions European Union Horizon 2020 research and innovation programme under grant agreement 951911 (AI4Media). This publication reflects only the authors views. The European Commission is not responsible for any use that may be made of the information it contains.**Author pre-print version.**

Despite the obvious potential application for malicious purposes, adversarial attacks may also be used in benevolent applications as well, notably to protect against automated private data analysis by automatic recognition systems, that are typically used by service providers in social media [5]. For example, adversarial attacks have been employed to disable known automatic face detection/recognition algorithms applied on visual data uploaded by social media users [6], while at the same time, not hiding the person identities to human viewers, since they are imperceptible to the human eye. Up to date, methods that adhere to privacy protection principles [7] typically emphasize on protecting private data against every possible recognizer (including humans), thus minimizing data utility for human viewers. On the other hand, methods that fool image classification systems using adversarial attacks [8] do not consider privacy protection constraints.

Motivated by the potential applications in privacy protection against automated image classification, we propose novel optimization objectives for adversarial attacks in order to fool deep neural network classifiers in a privacy preserving manner, while generating humanly imperceptible perturbations at the same time. Inspired by K-anonymity principles [9, 10], the proposed optimization conditions assure that the initial identities of the crafted adversarial examples are not only misclassified by the neural network decision function, but they are also uniformly spread along K different ranked output positions. In addition, human imperceptibility is maintained by emphasizing on minimizing the introduced perturbation by our adversarial attack. To this end, we add a visual similarity term to the optimization constraints, i.e., the Complex Wavelet transform variant of the Structural Similarity Index Measure (CW-SSIM), implemented as an additional loss function [11], that guides the adversarial attack towards image pixel value modifications having minimal impact on the perceived image quality.

The rest of the paper is structured as follows. Influential adversarial attacks that consist the state of the art in this field are described in Section 2. Section 3 provides the insights and methodological description of the proposed method. Section 4 includes the experimental evaluation and finally, conclusions are drawn in Section 5.

2. ADVERSARIAL ATTACKS

Let $\mathcal{S} = \{\mathcal{X}, \mathcal{Y}\}$ be an image classification dataset that has been employed to train a classifier f with trainable parameters $\boldsymbol{\theta}$. That is, for any given sample $\mathbf{x} \in \mathcal{X}$, the neural network classifier is able to recover the true label $y \in \mathcal{Y}$ by its decision function e.g., $f(\mathbf{x}; \boldsymbol{\theta}) = y$, where $f(\mathbf{x}; \boldsymbol{\theta})$ contains the network output activation values for a given $\mathbf{x}$. Adversarial attacks typically generate the necessary perturbation as a mapping to a space of similar characteristics i.e., $\mathcal{X} \mapsto \tilde{\mathcal{X}}$ such that the ability of the classifier to map to the correct label is disabled:

$$f(\tilde{\mathbf{x}}; \boldsymbol{\theta}) \neq y,$$

where $\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{p}$ is an adversarial example, and $\mathbf{p}$ is the introduced perturbation vector. The minimum amount of perturbation is determined by optimizing some objective function, e.g., minimizing the L_2 norm $\|\mathbf{p}\|_2$.

One of the most well-known methods to this end is the L-BFGS attack [1]. Assuming access to the outputs of a continuous loss function L_f, associated with the classifier function f to be deceived, the adversary selects a target label $t \neq y \in \mathcal{Y}$, for the adversarial example $\tilde{\mathbf{x}}$. Then, the minimum perturbation is determined in an iterative fashion:

$$\min_{\mathbf{p}}: \quad c\|\mathbf{p}\|_2 + L_f(f(\tilde{\mathbf{x}}; \boldsymbol{\theta}), t), \tag{1}$$

until the minimum $\mathbf{p}$ that satisfies $f(\tilde{\mathbf{x}}; \boldsymbol{\theta})) = t$ is obtained (or approximated for non-convex loss functions L_f). The parameter $c > 0$ controls the amount of perturbation introduced per iteration step and is empirically set using line search.

Fast Gradient Sign [2] is a significantly faster alternative that estimates the perturbation $\mathbf{p}$ in a single gradient update step, along the direction of the gradient sign at each image pixel:

$$\mathbf{p} = c \cdot \text{sign}(\nabla L_f(f(\mathbf{x}; \boldsymbol{\theta}), t)), \tag{2}$$

at the expense of producing more noisy examples than L-BFGS.

DeepFool [12] is an un-targeted adversarial attack method that produces adversarial examples with less perturbation than L-BFGS or Fast Gradient Sign, by approximating the decision boundaries of deep neural networks with linear/affine classifiers, of the form $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$. The minimum perturbation $\mathbf{p}$ required to change the classifier label is estimated by the orthogonal projection of the sample $\mathbf{x}$ to the closest decision boundary, namely $\mathbf{p} = -\frac{g(\mathbf{x})}{\|\mathbf{w}\|^2}\mathbf{w}$. An iterative optimization algorithm estimates this perturbation, as follows:

$$\begin{aligned} &\min_{\mathbf{p}}: \quad \|\mathbf{p}\|_2^2 \\ &\text{s. t.}: \quad g(\tilde{\mathbf{x}}) - \nabla g(\tilde{\mathbf{x}})^T\mathbf{p} = 0, \end{aligned} \tag{3}$$

until the perturbation $\mathbf{p}$ changes the classifier label. A similar approach can be employed for deriving the multiclass case [12], as well.

The Carlini-Wagner (C & W) attack [13] is perhaps one of the most powerful targeted attacks up to date, that have been found to be effective against Defensive Distillation [14], as well as a number of other defenses [15]. This method is the generalization of the L-BFGS attack, having investigated different combinations of loss functions, suitable image data mappings for avoiding limitations of the box constraint $\mathbf{x}, \tilde{\mathbf{x}} \in [0, 1]$ and various gradient descend optimization algorithms.

Finally, we should also mention that other adversarial attack types have been proposed, such as the Jacobian-based Saliency Map Attack [16], or even attacks that modify only a single image pixel [17]. The reader is referred to the review papers [18, 19, 20] for more information.

3. K-ANONYMITY ATTACK AGAINST DEEP NEURAL NETWORKS

K-anonymity is a generic privacy protection concept that suggests that the maximum probability of identifying an individual in a specific set must be lower than $1/K$ [9, 10]. In order to quantify privacy protection introduced by adversarial attacks according to the K-anonymity principles, one might employ the class identification probabilities (e.g., classification rate) of a set of adversarial examples $\tilde{\mathcal{X}}$ produced by some attack against the classifier decision function $f(\tilde{\mathbf{x}}; \boldsymbol{\theta}) = y$ that have been attacked. However, such a definition does not examine the overall neural network output activation values.

More specifically, according to the perspectives of Label Ranking [21] and Multi-Label Classification [22], the network output activation values contain an underlying strict ordered ranking over the finite label set $\mathcal{Y} = \{\ell_i, \ldots, \ell_C\}$, where C is the total number of classes supported by the model and $\ell_i \succ_{\mathbf{x}} \ell_j$ denotes that for a given data example $\mathbf{x}$, label ℓ_i is a more probable output classification label than label ℓ_j. For simplicity reasons, we denote the output ranking with a vector function $\mathbf{r}(\mathbf{x}) = [r_{\mathbf{x}}(1), \ldots, r_{\mathbf{x}}(C)]^T$, such that the output classification label of sample $\mathbf{x}$ by the deep neural network model is given in the 1st ranking position $r_{\mathbf{x}}(1)$. We argue that the ranking obtained for any sample $\mathbf{x}$ may encode underlying data properties, that may expose information about the class of interest (e.g., we assume that in most cases, the true label of misclassified samples may be obtained by $r_{\mathbf{x}}(2)$).

Taking the above into consideration, we design an adversarial attack methodology in order to preserve anonymity (hide the true label) of every sample $\mathbf{x} \in \mathcal{X}$ against the neural network according to the K-anonymity constraints, taking into consideration the whole network output layer. To this end, we argue that the appropriate mapping $\mathcal{X} \mapsto \tilde{\mathcal{X}}$ should achieve two conditions:

$$r_{\tilde{\mathbf{x}}}(1) \neq y, \quad \forall \tilde{\mathbf{x}} \in \tilde{\mathcal{X}}, \tag{4}$$

$$p(i) = \begin{cases} P\left(r_{\tilde{\mathbf{x}}}(i) = y\right) \leq 1/K & ,\forall i \in \{1, \ldots, C\} \\ 0 & ,\text{otherwise}, \end{cases} \tag{5}$$

where $p(\cdot)$ is the probability mass function of its argument, containing the probability of a retrieving the true label of the adversarial example $\tilde{\mathbf{x}}$ in the $i-$th position of the output ranking, e.g., $P\left(r_{\tilde{\mathbf{x}}}(1)\right)$ is equal to the classification rate of the model.

K is a hyperparameter denoting the $K-$anonymity protection level, e.g., 5-Anonymity. Condition (4) is the adversarial attack objective, i.e., disabling correct classification. Condition (5) is the novel K-anonymity objective, that achieves anonymity in adversarial examples along the whole network output. Without it, we argue that the network may still be used to classify adversarial examples by exploiting rule-based reasoning, e.g., by exploiting an adversarial example detector in the system. For instance, if an adversarial attack has been detected by some adversarial attack detection method [23], the example may still be classified correctly using the same network, only using the output of e.g., the 2nd ranking position $r_{\tilde{\mathbf{x}}}(2)$ instead of the 1st $r_{\tilde{\mathbf{x}}}(1)$. Condition (5) guarantees that such simplified reasoning rules are impossible to be devised for the adversarial examples, according to the K-anonymity concepts.

Let the dataset $\mathcal{X}$ consisting of exactly N samples on which we would like to fool the task classifier. We assume that the network has 100% accuracy in this dataset, i.e., $r_{\mathbf{x}}(1) = y$ and for every $\mathbf{x} \in \mathcal{X}$. The aim of the proposed adversarial attack is to generate a set $\tilde{\mathcal{X}}$ of N adversarial examples that satisfies the constraints (4) and (5). To this end, we demand that the true labels of exactly K data groups, each containing N/K samples of $\tilde{\mathcal{X}}$, are not retrieved in at least $k \in \mathcal{K} = \{2, \ldots, K+1\}$ sorted ranking positions, relevant to K. Assuming $K = 5$, then 5 such data groups must be formed, demanding that the true labels of the first group are not retrieved in ranking positions $r_{\mathbf{x}}(1), r_{\mathbf{x}}(2)$, while the labels in the 5-th group are not retrieved in any position of $r_{\mathbf{x}}(i), \forall i \leq 6$.

Furthermore, since the proposed methodology is designed to achieve more difficult constraints when compared to standard adversarial attacks, it should be expected that increased perturbation may be generated to the crafted adversarial examples, as a consequence. To counteract this effect, we introduce a visual similarilty term to our proposed optimization problem, namely the CW-SSIM loss function [11] $s(\mathbf{x}, \tilde{\mathbf{x}})$ between the initial samples and the crafted adversarial examples, guiding the optimization problem towards solutions that regulate the amount of perturbation generated by the adversarial attack.

Maintaining the assumptions of white-box attacks i.e., access to a continuous loss function L_f associated with f, we propose the following optimization problem:

$$\min_{\mathbf{p}}: \quad \|\mathbf{p}\|_2 + (1 - s(\mathbf{x}, \tilde{\mathbf{x}})) + \sum_{i=2}^{k} L_f(f(\tilde{\mathbf{x}}; \boldsymbol{\theta}), r_{\mathbf{x}}(i)), \quad (6)$$

until the ranking obtained for $\tilde{\mathbf{x}}$ by the neural network architecture satisfies the constraint $r_{\tilde{x}}(i) \succ_{\tilde{\mathbf{x}}} y, \forall i \in \mathcal{K}$. In fact, instead of the using the ranked label positions, any k randomly selected target labels $\ell_i \neq y \in \mathcal{Y}$ could be employed in the proposed method, as well, without violating the K-anonymity constraints. However, it should also be noted that the variable $k \in \mathcal{K}$ must be set to different value for every N/K sample groups, (e.g., $k = 2$ for group 1, $k = K + 1$ for group K). If we set the variable k equal to some specific value of $\mathcal{K}$ for every of the N adversarial examples to be crafted (e.g., $k = 5$), then the K-Anonymity constraints will be violated, since the probability of retrieving the true label of $\tilde{\mathbf{x}}$ will violate the constraint (5), since $P(r_{\tilde{x}}(5) = y) > 1/K$.

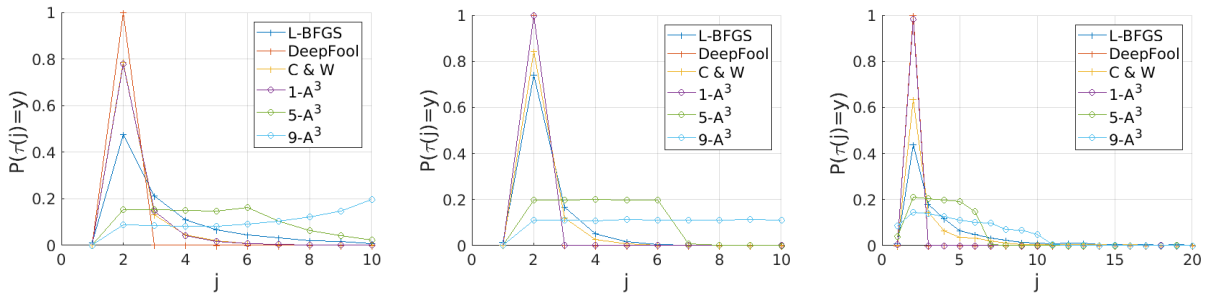
As can be observed in (6), for a given $K = 1$ (i.e., $k = 2$ for all training data) and by omitting the visual similarity term, the proposed method degenerates to the standard L-BFGS method [1]. Therefore, the proposed method can be viewed as a generalization of L-BFGS that respects and supports the K-anonymity constraints, for the multi-class classification case.

4. EXPERIMENTS

In order to evaluate the performance of the proposed method, we have employed the MNIST (digit classification), CIFAR-10 (object recognition) and Yale (face recognition) datasets. A CNN architecture, namely the LeNet5 (MNIST-LeNet) [24] was trained from scratch in MNIST dataset. In CIFAR-10 dataset, we have trained the MobileNetV2 architecture [25] (CIFAR-10-MobileNetV2). In Yale dataset, we fine-tuned a 9-Layer LightCNN architecture [26], that had been pre-trained using more than 1.5M facial images from CelebA dataset (Yale-LightCNN), totaling 3 architecture-dataset combinations. All conducted experiments were implemented using PyTorch.

The proposed method was employed to attack each architecture for different values of $K = 1, 5, 9$. For comparison reasons, we have also employed the L-BFGS [1], DeepFool [12] and the C & W [13] attack with L_2 distance. All methods were implemented using their default parameter settings. During the optimization process, we have slightly tuned the learning rate parameter of the optimizers, while keeping their settings equal for all competing methods. The same target labels were assigned to L-BFGS and C & W attacks. We have employed these methods to generate adversarial datasets $\tilde{\mathcal{X}}$ by modifying the training samples of each dataset, so that all samples were classified correctly by the task classifier.

The datasets obtained by each method were evaluated in terms of satisfying K-anonymity principles, as have been defined by equation (5) of this paper. That is, we have tried to retrieve the original dataset labels using the architecture $\boldsymbol{\theta}$ to obtain ranked label outputs for each adversarial sample. We have determined the probability mass functions $p(i)$ for obtaining the ground truth label at the $i-$th ranking position, plotted in Figure 1. As can be observed, the datasets obtained by employing L-BFGS, DeepFool, C & W, and the variant



(a) MNIST (b) CIFAR (c) Yale

Fig. 1: Probability mass functions of recovering the original labels y in the j-th sorted ranking position $P(r_{\tilde{\mathbf{x}}}(j) = y)$ generated by L-BFGS, DeepFool, C & W, and the proposed methods. L-BFGS, DeepFool, C & W, and the variant 1-A^3 of the proposed method do not satisfy the K-anonymity requirements, since in most cases, the original label y can be recovered by retrieving the label ranked 2nd.

Table 1: Introduced perturbation of competing methods

Dataset	MNIST			CIFAR10			Yale		
Method/Metric	ASR%	SSIM	MSE$\times 10^3$	ASR%	MS-SSIM	MSE$\times 10^5$	ASR%	SSIM	MSE$\times 10^4$
L-BFGS	99.98	73.75	2.42	98.53	99.84	5.35	99.23	93.89	5.97
DeepFool	99.92	80.35	1.57	99.74	99.99	2.04	100	97.87	1.71
C & W	99.93	80.12	1.45	99.96	99.99	3.65	99.94	94.21	5.16
Proposed ($K = 1$)	100	**84.64**	**1.87**	99.95	**99.99**	**1.65**	99.43	**98.05**	**1.59**
Proposed ($K = 5$)	100	72.72	5.37	99.77	99.99	4.96	96.26	95.17	7.52
Proposed ($K = 9$)	100	61.69	15.15	99.85	99.98	8.08	91.34	93.17	4.36

of the proposed method with $K = 1$ do not satisfy the K-anonymity requirements, since $P(r_{\tilde{\mathbf{x}}}(2) = y) > 1/K$ for every $K > 1$. On the other hand, the probabilities of the adversarial examples crafted by the proposed method with $K = 5$ and $K = 9$, satisfy $P(r_{\tilde{\mathbf{x}}}(j) = y) \leq 1/5$ for $i = 2, \ldots, 6$ and $P(r_{\tilde{\mathbf{x}}}(j) = y) \leq 1/9$ for $j = 2, \ldots, 10$ respectively in almost every case, as the output distribution is almost uniform, and the discrete probabilities in each ranking position do not exceed $1/K$.

In addition, all adversarial attack methods were evaluated in terms of the introduced perturbation. As evaluation metrics, we have computed the Adversarial Attack success rate (ASR%), the average Mean Squared Error (MSE)$= \|\mathbf{x} - \tilde{\mathbf{x}}\|_2^2$ and average Structural Similarity [11] (SSIM), scaled from $[0, 1]$ to $[0, 100]$, for better visibility purposes, for the adversarial datasets generated by each method. For CIFAR-10-MobileNet, we have reported the Multi-Scale Structural Similarity (MS-SSIM) instead of SSIM, which is more suitable variant for RGB images. Higher MSE values denote that the adversarial examples contain increased perturbation, and high SSIM values denote that the adversarial example $\tilde{\mathbf{x}}$ appears visually similar with the training example $\mathbf{x}$. A visual comparison of the introduced perturbation for Yale dataset is shown in Figure 2.

Results of the evaluation are drawn on Table 1. The proposed method with $K = 1$ generated the least amount of perturbation and generated the most visually similar examples, in the light of the selected evaluation metrics, especially when compared to the standard L-BFGS attack, attributed to employing the CW-SSIM loss in its optimization process. The introduced perturbation of the proposed method for $K = 5$ and $K = 9$ was increased when compared to when $K = 1$. This was expected due to the demand of adherence to the K-anonymity requirements. However, it should be noted that the visual similarity of the adversarial examples generated by the proposed methods for $K = 5$ and $K = 9$ with the original images, have been found to be very close, or even increased in some cases, when compared to the similarity of adversarial examples generated by L-BFGS with the original images. This effect should also be attributed to the exploitation of the CW-SSIM loss function.

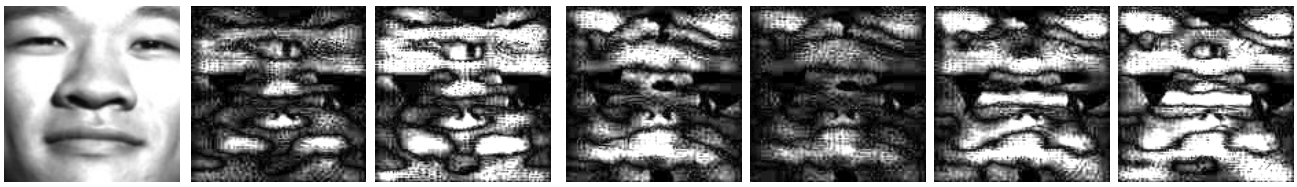
5. CONCLUSION

In this paper, we presented a method that incorporates the K-anonymity requirements to the white-box adversarial attack optimization problem. An important limitation of the proposed method is that it requires explicit knowledge about the target architecture it targets to fool. Future work may focus towards extending the proposed method to black-box adversarial attack methodologies, or study its performance across other modalities (e.g., sound). In addition, it could include studying the robustness of models trained with examples derived by the proposed adversarial attack, via an adversarial training scheme.

6. REFERENCES

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[2] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial exam-



(a) X (b) L-BFGS (c) C&W (d) DeepFool (e) K-1-A^3 (f) K-5-A^3 (g) K-9-A^3

Fig. 2: Adversarial examples on Yale Dataset. The original image is shown on the left (a). Columns (b)-(g) depict the perturbation generated by each method (magnified by a scale of 10).

ples," in *International Conference on Learning Representations*, 2015.

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

[4] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard, "Universal adversarial perturbations," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 86–94.

[5] Bogdan Batrinca and Philip C. Treleaven, "Social media analytics: a survey of techniques, tools and platforms," *AI & SOCIETY*, vol. 30, no. 1, pp. 89–116, Feb 2015.

[6] Yujia Liu, Weiming Zhang, and Nenghai Yu, "Protecting privacy in shared photos via adversarial examples based stealth," *Security and Communication Networks*, vol. 2017, 2017.

[7] Ivan Sikiric, Tomislav Hrkac, Karla Zoran Kalafatic, et al., "I know that person: Generative full body and face de-identification of people in images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 15–24.

[8] Efstathios Chatzikyriakidis, Christos Papaioannidis, and Ioannis Pitas, "Adversarial face de-identification," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 684–688.

[9] Latanya Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[10] Vasileios Mygdalis, Anastasios Tefas, and Ioannis Pitas, "K-anonymity inspired adversarial attack and multiple one-class classification defense," *Neural Networks*, vol. 124, pp. 296–307, 2020.

[11] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[12] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2574–2582.

[13] 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.

[14] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2016, pp. 582–597.

[15] 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.

[16] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami, "The limitations of deep learning in adversarial settings," in *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*. IEEE, 2016, pp. 372–387.

[17] Jiawei Su, Danilo Vasconcellos Vargas, and Sakurai Kouichi, "One pixel attack for fooling deep neural networks," *arXiv preprint arXiv:1710.08864*, 2017.

[18] Xiaoyong Yuan, Pan He, Qile Zhu, Rajendra Rana Bhat, and Xiaolin Li, "Adversarial examples: Attacks and defenses for deep learning," *arXiv preprint arXiv:1712.07107*, 2017.

[19] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay, "Adversarial attacks and defences: A survey," *arXiv preprint arXiv:1810.00069*, 2018.

[20] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018.

[21] Johannes Fürnkranz, Eyke Hüllermeier, Eneldo Loza Mencía, and Klaus Brinker, "Multilabel classification via calibrated label ranking," *Machine Learning*, vol. 73, no. 2, pp. 133–153, Nov 2008.

[22] Jianqing Zhu, Shengcai Liao, Zhen Lei, and Stan Z Li, "Multi-label convolutional neural network based pedestrian attribute classification," *Image and Vision Computing*, vol. 58, pp. 224–229, 2017.

[23] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff, "On detecting adversarial perturbations," *arXiv preprint arXiv:1702.04267*, 2017.

[24] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[25] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," *arXiv preprint arXiv:1801.04381*, 2018.

[26] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan, "A light cnn for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.