Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey

Naveed Akhtar and Ajmal Mian

ACKNOWLEDGEMENTS: The authors thank Nicholas Carlini (UC Berkeley) and Dimitris Tsipras (MIT) for feedback to improve the survey quality. We also acknowledge X. Huang (Uni. Liverpool), K. R. Reddy (IISC), E. Valle (UNICAMP), Y. Yoo (CLAIR) and others for providing pointers to make the survey more comprehensive. This research was supported by ARC grant DP160101458.

Abstract—Deep learning is at the heart of the current rise of artificial intelligence. In the field of Computer Vision, it has become the workhorse for applications ranging from self-driving cars to surveillance and security. Whereas deep neural networks have demonstrated phenomenal success (often beyond human capabilities) in solving complex problems, recent studies show that they are vulnerable to adversarial attacks in the form of subtle perturbations to inputs that lead a model to predict incorrect outputs. For images, such perturbations are often too small to be perceptible, yet they completely fool the deep learning models. Adversarial attacks pose a serious threat to the success of deep learning in practice. This fact has recently lead to a large influx of contributions in this direction. This article presents the first comprehensive survey on adversarial attacks on deep learning in Computer Vision. We review the works that design adversarial attacks, analyze the existence of such attacks and propose defenses against them. To emphasize that adversarial attacks are possible in practical conditions, we separately review the contributions that evaluate adversarial attacks in the real-world scenarios. Finally, drawing on the reviewed literature, we provide a broader outlook of this research direction.

Index Terms—Deep Learning, adversarial perturbation, black-box attack, white-box attack, adversarial learning, perturbation detection.

1 Introduction

Deep Learning [1] is providing major breakthroughs in solving the problems that have withstood many attempts of machine learning and artificial intelligence community in the past. As a result, it is currently being used to decipher hard scientific problems at an unprecedented scale, e.g. in reconstruction of brain circuits [2]; analysis of mutations in DNA [3]; prediction of structure-activity of potential drug molecules [4], and analyzing the particle accelerator data [5] [6]. Deep neural networks have also become the preferred choice to solve many challenging tasks in speech recognition [7] and natural language understanding [8].

In the field of Computer Vision, deep learning became the center of attention after Krizhevsky et al. [9] demonstrated the impressive performance of a Convolutional Neural Network (CNN) [10] based model on a very challenging large-scale visual recognition task [11] in 2012. A significant credit for the current popularity of deep learning can also be attributed to this seminal work. Since 2012, the Computer Vision community has made numerous valuable contributions to deep learning research, enabling it to provide solutions for the problems encountered in medical science [21] to mobile applications [181]. The recent breakthrough in artificial intelligence in the form of tabula-rasa learning of AlphaGo Zero [14] also owes a fair share to deep Residual Networks (ResNets) [147] that were originally proposed for the task of image recognition.

With the continuous improvements of deep neural network models [145], [147], [168]; open access to efficient deep learning software libraries [177], [178], [179]; and easy availability of hardware required to train complex models, deep learning is fast achieving the maturity to enter into safety and security critical applications, e.g. self driving cars [12], [182], surveillance [13], maleware detection [34], [107], drones and robotics [157], [180], and voice command recognition [7]. With the recent real-world developments like facial recognition ATM [183] and Face ID security on mobile phones [184], it is apparent that deep learning solutions, especially those originating from Computer Vision problems are about to play a major role in our daily lives.

Whereas deep learning performs a wide variety of Computer Vision tasks with remarkable accuracies, Szegedy et al. [22] first discovered an intriguing weakness of deep neural networks in the context of image classification. They showed that despite their high accuracies, modern deep networks are surprisingly susceptible to adversarial attacks in the form of small perturbations to images that remain (almost) imperceptible to human vision system. Such attacks can cause a neural network classifier to completely change its prediction about the image. Even worse, the attacked models report high confidence on the wrong prediction. Moreover, the same image perturbation can fool multiple network classifiers. The profound implications of these results triggered a wide interest of researchers in adversarial attacks and their defenses for deep learning in general.

Since the findings of Szegedy et al. [22], several interesting results have surfaced regarding adversarial attacks on deep learning in Computer Vision. For instance, in addition to the image-specific adversarial perturbations [22], Moosavi-Dezfooli et al. [16] showed the existence of 'uni-

• *N. Akhtar and A. Mian are with the School of Computer Science and Software Engineering, University of Western Australia. E-mail: {naveed.akhtar, ajmal.mian}@uwa.edu.au*

Manuscript received August 2017, revised...

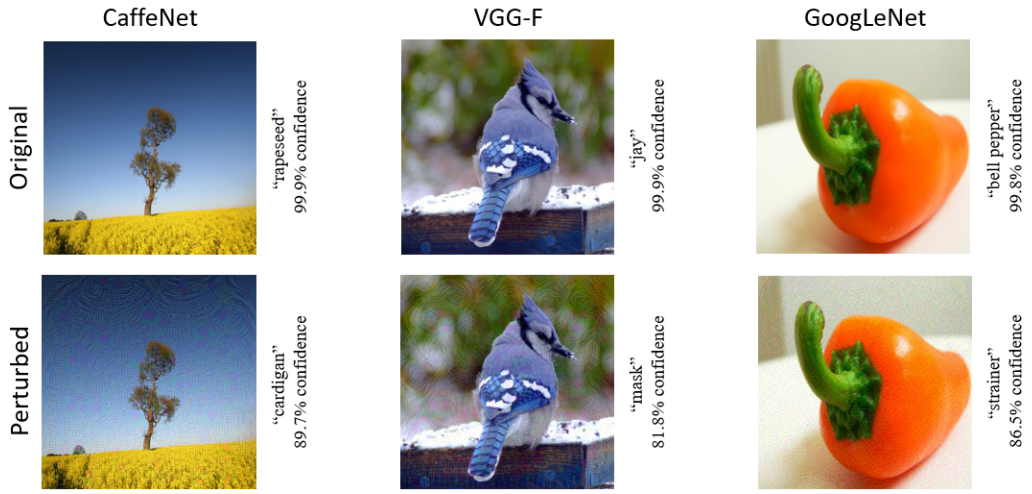


Fig. 1: Example of attacks on deep learning models with 'universal adversarial perturbations' [16]: The attacks are shown for the CaffeNet [9], VGG-F network [17] and GoogLeNet [18]. All the networks recognized the original clean images correctly with high confidence. After small perturbations were added to the images, the networks predicted wrong labels with similar high confidence. Notice that the perturbations are hardly perceptible for human vision system, however their effects on the deep learning models are catastrophic.

versal perturbations' that can fool a network classifier on *any* image (see Fig. 1 for example). Similarly, Athalye et al. [65] demonstrated that it is possible to even 3-D print real-world objects that can fool deep neural network classifiers (see Section 4.3). Keeping in view the significance of deep learning research in Computer Vision and its potential applications in the real life, this article presents the first comprehensive survey on adversarial attacks on deep learning in Computer Vision. The article is intended for a wider readership than Computer Vision community, hence it assumes only basic knowledge of deep learning and image processing. Nevertheless, it also discusses technical details of important contributions for the interested readers.

We first describe the common terms related to adversarial attacks in the parlance of Computer Vision in Section 2. In Section 3, we review the adversarial attacks for the task of image classification and beyond. A separate section is dedicated to the approaches that deal with adversarial attacks in the real-world conditions. Those approaches are reviewed in Section 4. In the literature, there are also works that mainly focus on analyzing the existence of adversarial attacks. We discuss those contributions in Section 5. The approaches that make defense against the adversarial attacks as their central topic are discussed in Section 6. In Section 7, we provide a broader outlook of the research direction based on the reviewed literature. Finally, we draw conclusion in Section 8.

2 Definitions of Terms

In this section, we describe the common technical terms used in the literature related to adversarial attacks on deep learning in Computer Vision. The remaining article also follows the same definitions of the terms.

- *Adversarial example/image* is a modified version of a clean image that is intentionally perturbed (e.g. by adding noise) to confuse/fool a machine learning technique, such as deep neural networks.
- *Adversarial perturbation* is the noise that is added to the clean image to make it an adversarial example.
- *Adversarial training* uses adversarial images besides the clean images to train machine learning models.
- *Adversary* more commonly refers to the agent who creates an adversarial example. However, in some cases the example itself is also called adversary.
- *Black-box attacks* feed a targeted model with the adversarial examples (during testing) that are generated without the knowledge of that model. In some instances, it is assumed that the adversary has a limited knowledge of the model (e.g. its training procedure and/or its architecture) but definitely does not know about the model parameters. In other instances, using any information about the target model is referred to as 'semi-black-box' attack. We use the former convention in this article.
- *Detector* is a mechanism to (only) detect if an image is an adversarial example.
- *Fooling ratio/rate* indicates the percentage of images on which a trained model changes its prediction label after the images are perturbed.
- *One-shot/one-step methods* generate an adversarial perturbation by performing a single step computation, e.g. computing gradient of model loss once. The opposite are iterative methods that perform the same computation multiple times to get a single perturbation. The latter are often computationally expensive.
- *Quasi-imperceptible* perturbations impair images very slightly for human perception.
- *Rectifier* modifies an adversarial example to restore the prediction of the targeted model to its prediction on the clean version of the same example.
- *Targeted attacks* fool a model into falsely predicting a specific label for the adversarial image. They are

opposite to the *non-targeted* attacks in which the predicted label of the adversarial image is irrelevant, as long as it is not the correct label.
- *Threat model* refers to the types of potential attacks considered by an approach, e.g. black-box attack.
- *Transferability* refers to the ability of an adversarial example to remain effective even for the models other than the one used to generate it.
- *Universal perturbation* is able to fool a given model on 'any' image with high probability. Note that, universality refers to the property of a perturbation being 'image-agnostic' as opposed to having good transferability.
- *White-box attacks* assume the complete knowledge of the targeted model, including its parameter values, architecture, training method, and in some cases its training data as well.

3 Adversarial Attacks

In this section, we review the body of literature in Computer Vision that introduces methods for adversarial attacks on deep learning. The reviewed literature mainly deals with the art of fooling the deep neural networks in 'laboratory settings', where approaches are developed for the typical Computer Vision tasks, e.g. recognition, and their effectiveness is demonstrated using standard datasets, e.g. MNIST [10]. The techniques that focus on attacking deep learning in the real-world conditions are separately reviewed in Section 4. However, it should be noted that the approaches reviewed in this section form the basis of the real-world attacks, and almost each one of them has the potential to significantly affect deep learning in practice. Our division is based on the evaluation conditions of the attacks in the original contributions.

The review in this section is mainly organized in chronological order, with few exceptions to maintain the flow of discussion. To provide technical understanding of the core concepts to the reader, we also go into technical details of the popular approaches and some representative techniques of the emerging directions in this area. Other methods are discussed briefly. We refer to the original papers for the details on those techniques. This section is divided into two parts. In part 3.1, we review the methods that attack deep neural networks performing the most common task in Computer Vision, i.e. classification/recognition. Approaches that are predominantly designed to attack deep learning beyond this task are discussed in part 3.2.

3.1 Attacks for classification

3.1.1 Box-constrained L-BFGS

Szegedy et al. [22] first demonstrated the existence of small perturbations to the images, such that the perturbed images could fool deep learning models into misclassification. Let $\mathbf{I}_c \in \mathbb{R}^m$ denote a vectorized clean image - the subscript 'c' emphasizes that the image is clean. To compute an additive perturbation $\boldsymbol{\rho} \in \mathbb{R}^m$ that would distort the image very slightly to fool the network, Szegedy et al. proposed to solve the following problem:

$$\min_{\boldsymbol{\rho}} ||\boldsymbol{\rho}||_2 \;\; \text{s.t. } \mathcal{C}(\mathbf{I}_c + \boldsymbol{\rho}) = \ell;\; \mathbf{I}_c + \boldsymbol{\rho} \in [0,1]^m, \qquad (1)$$

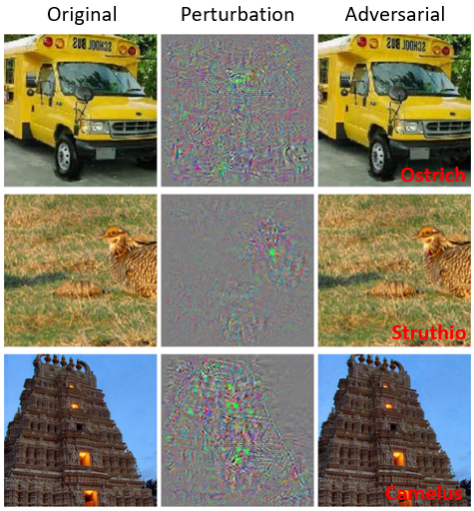


Fig. 2: Illustration of adversarial examples generated using [22] for AlexNet [9]. The perturbations are magnified 10x for better visualization (values shifted by 128 and clamped). The predicted labels of adversarial examples are also shown.

where 'ℓ' denotes the label of the image and $\mathcal{C}(.)$ is the deep neural network classifier. The authors proposed to solve (1) for its non-trivial solution where 'ℓ' is different from the original label of $\mathbf{I}_c$. In that case, (1) becomes a hard problem, hence an approximate solution is sought using a box-constrained L-BFGS [20]. This is done by finding the minimum $c > 0$ for which the minimizer $\boldsymbol{\rho}$ of the following problem satisfies the condition $\mathcal{C}(\mathbf{I}_c + \boldsymbol{\rho}) = \ell$:

$$\min_{\boldsymbol{\rho}} \;\; c|\boldsymbol{\rho}| + \mathcal{L}(\mathbf{I}_c + \boldsymbol{\rho}, \ell) \;\; s.t.\, \mathbf{I}_c + \boldsymbol{\rho} \in [0,1]^m, \qquad (2)$$

where $\mathcal{L}(.,.)$ computes the loss of the classifier. We note that (2) results in the exact solution for a classifier that has a convex loss function. However, for deep neural networks, this is generally not the case. The computed perturbation is simply added to the image to make it an adversarial example.

As shown in Fig. 2, the above method is able to compute perturbations that when added to clean images fool a neural network, but the adversarial images appear similar to the clean images to the human vision system. It was observed by Szegedy et al. that the perturbations computed for one neural network were also able to fool multiple networks. These astonishing results identified a blind-spot in deep learning. At the time of this discovery the Computer Vision community was fast adapting to the impression that deep learning features define the space where perceptual distances are well approximated by the Euclidean distances. Hence, these contradictory results triggered a wide interest of researchers in adversarial attacks on deep learning in Computer Vision.

3.1.2 Fast Gradient Sign Method (FGSM)

It was observed by Szegedy et al. [22] that the robustness of deep neural networks against the adversarial examples could be improved by adversarial training. To enable effective adversarial training, Goodfellow et al. [23] developed

a method to efficiently compute an adversarial perturbation for a given image by solving the following problem:

$$\boldsymbol{\rho} = \epsilon \ \text{sign}\left(\nabla \mathcal{J}(\boldsymbol{\theta}, \mathbf{I}_c, \ell)\right), \tag{3}$$

where $\nabla \mathcal{J}(.,.,.)$ computes the gradient of the cost function around the current value of the model parameters $\boldsymbol{\theta}$ w.r.t. $\mathbf{I}_c$, sign$(.)$ denotes the sign function and ϵ is a small scalar value that restricts the norm of the perturbation. The method for solving (3) was termed 'Fast Gradient Sign Method' (FGSM) in the original work.

Interestingly, the adversarial examples generated by FGSM exploit the 'linearity' of deep network models in the higher dimensional space whereas such models were commonly thought to be highly non-linear at that time. Goodfellow et al. [23] hypothesized that the designs of modern deep neural networks that (intentionally) encourage linear behavior for computational gains, also make them susceptible to cheap analytical perturbations. In the related literature, this idea is often referred to as the 'linearity hypothesis', which is substantiated by the FGSM approach.

Kurakin et al. [80] noted that on the popular large-scale image recognition data set ImageNet [11], the top-1 error rate on the adversarial examples generated by FGSM is around $63-69\%$ for $\epsilon \in [2, 32]$. The authors also proposed a 'one-step target class' variation of the FGSM where instead of using the true label ℓ of the image in (3), they used the label ℓ_{target} of the least likely class predicted by the network for $\mathbf{I}_c$. The computed perturbation is then subtracted from the original image to make it an adversarial example. For a neural network with cross-entropy loss, doing so maximizes the probability that the network predicts ℓ_{target} as the label of the adversarial example. It is suggested, that a random class can also be used as the target class for fooling the network, however it may lead to less interesting fooling, e.g. misclassification of one breed of dog as another dog breed. The authors also demonstrated that adversarial training improves robustness of deep neural networks against the attacks generated by FGSM and its proposed variants.

The FGSM perturbs an image to increase the loss of the classifier on the resulting image. The sign function ensures that the magnitude of the loss is maximized, while ϵ essentially restricts the ℓ_∞-norm of the perturbation. Miyato et al. [103] proposed a closely related method to compute the perturbation as follows

$$\boldsymbol{\rho} = \epsilon \frac{\nabla \mathcal{J}(\boldsymbol{\theta}, \mathbf{I}_c, \ell)}{||\nabla \mathcal{J}(\boldsymbol{\theta}, \mathbf{I}_c, \ell)||_2}. \tag{4}$$

In the above equation, the computed gradient is normalized with its ℓ_2-norm. Kurakin et al. [80] referred to this technique as 'Fast Gradient L_2' method and also proposed an alternative of using the ℓ_∞-norm for normalization, and referred to the resulting technique as 'Fast Gradient L_∞' method. Broadly speaking, all of these methods are seen as 'one-step' or 'one-shot' methods in the literature related to adversarial attacks in Computer Vision.

3.1.3 Basic & Least-Likely-Class Iterative Methods

The one-step methods perturb images by taking a single large step in the direction that increases the loss of the classifier (i.e. one-step gradient ascent). An intuitive extension of this idea is to iteratively take multiple small steps while adjusting the direction after each step. The Basic Iterative Method (BIM) [35] does exactly that, and iteratively computes the following:

$$\mathbf{I}_{\boldsymbol{\rho}}^{i+1} = \text{Clip}_{\epsilon}\left\{\mathbf{I}_{\boldsymbol{\rho}}^{i} + \alpha \ \text{sign}(\nabla \mathcal{J}(\boldsymbol{\theta}, \mathbf{I}_{\boldsymbol{\rho}}^{i}, \ell)\right\}, \tag{5}$$

where $\mathbf{I}_{\boldsymbol{\rho}}^{i}$ denotes the perturbed image at the i^{th} iteration, $\text{Clip}_{\epsilon}\{.\}$ clips (the values of the pixels of) the image in its argument at ϵ and α determines the step size (normally, $\alpha = 1$). The BIM algorithm starts with $\mathbf{I}_{\boldsymbol{\rho}}^{0} = \mathbf{I}_c$ and runs for the number of iterations determined by the formula $\lfloor \min(\epsilon + 4, 1.25\epsilon) \rfloor$. Madry et al. [55] pointed out that BIM is equivalent to (the ℓ_∞ version of) Projected Gradient Descent (PGD), a standard convex optimization method.

Similar to extending the FGSM to its 'one-step target class' variation, Kurakin et al. [35] also extended BIM to Iterative Least-likely Class Method (ILCM). In that case, the label ℓ of the image in (5) is replaced by the target label ℓ_{target} of the least likely class predicted by the classifier. The adversarial examples generated by the ILCM method has been shown to seriously affect the classification accuracy of a modern deep architecture Inception v3 [145], even for very small values of ϵ, e.g. < 16.

3.1.4 Jacobian-based Saliency Map Attack (JSMA)

In the literature, it is more common to generate adversarial examples by restricting ℓ_∞ or ℓ_2-norms of the perturbations to make them imperceptible for humans. However, Papernot et al. [60] also created an adversarial attack by restricting the ℓ_0-norm of the perturbations. Physically, it means that the goal is to modify only a few pixels in the image instead of perturbing the whole image to fool the classifier. The crux of their algorithm to generate the desired adversarial image can be understood as follows. The algorithm modifies pixels of the clean image one at a time and monitors the effects of the change on the resulting classification. The monitoring is performed by computing a saliency map using the gradients of the outputs of the network layers. In this map, a larger value indicates a higher likelihood of fooling the network to predict ℓ_{target} as the label of the modified image instead of the original label ℓ. Thus, the algorithm performs targeted fooling. Once the map has been computed, the algorithm chooses the pixel that is most effective to fool the network and alters it. This process is repeated until either the maximum number of allowable pixels are altered in the adversarial image or the fooling succeeds.

3.1.5 One Pixel Attack

An extreme case for the adversarial attack is when only one pixel in the image is changed to fool the classifier. Interestingly, Su et al. [68] claimed successful fooling of three different network models on 70.97% of the tested images by changing just one pixel per image. They also reported that the average confidence of the networks on the wrong labels was found to be 97.47%. We show representative examples of the adversarial images from [68] in Fig. 3. Su et al. computed the adversarial examples by using the concept of Differential Evolution [148]. For a clean image $\mathbf{I}_c$, they first created a set of 400 vectors in $\mathbb{R}^5$ such that each vector contained xy-coordinates and RGB values for an arbitrary

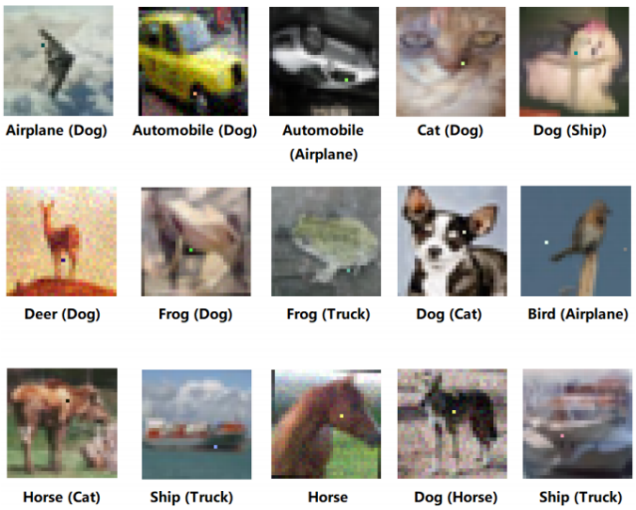


Fig. 3: Illustration of one pixel adversarial attacks [68]: The correct label is mentioned with each image. The corresponding predicted label is given in parentheses.

candidate pixel. Then, they randomly modified the elements of the vectors to create children such that a child competes with its parent for fitness in the next iteration, while the probabilistic predicted label of the network is used as the fitness criterion. The last surviving child is used to alter the pixel in the image.

Even with such a simple evolutionary strategy Su et al. [68] were able to show successful fooling of deep networks. Notice that, differential evolution enables their approach to generate adversarial examples without having access to any information about the network parameter values or their gradients. The only input their technique requires is the probabilistic labels predicted by the targeted model.

3.1.6 Carlini and Wagner Attacks (C&W)

A set of three adversarial attacks were introduced by Carlini and Wagner [36] in the wake of defensive distillation against the adversarial perturbations [38]. These attacks make the perturbations quasi-imperceptible by restricting their ℓ_2, ℓ_∞ and ℓ_0 norms, and it is shown that defensive distillation for the targeted networks almost completely fails against these attacks. Moreover, it is also shown that the adversarial examples generated using the unsecured (un-distilled) networks transfer well to the secured (distilled) networks, which makes the computed perturbations suitable for black-box attacks.

Whereas it is more common to exploit the transferability property of adversarial examples to generate black-box attacks, Chen et al. [41] also proposed 'Zeroth Order Optimization (ZOO)' based attacks that directly estimate the gradients of the targeted model for generating the adversarial examples. These attacks were inspired by C&W attacks. We refer to the original papers for further details on C&W and ZOO attacks.

3.1.7 DeepFool

Moosavi-Dezfooli et al. [72] proposed to compute a minimal norm adversarial perturbation for a given image in an iterative manner. Their algorithm, i.e. DeepFool initializes with the clean image that is assumed to reside in a region confined by the decision boundaries of the classifier. This region decides the class-label of the image. At each iteration, the algorithm perturbs the image by a small vector that is computed to take the resulting image to the boundary of the polyhydron that is obtained by linearizing the boundaries of the region within which the image resides. The perturbations added to the image in each iteration are accumulated to compute the final perturbation once the perturbed image changes its label according to the original decision boundaries of the network. The authors show that the DeepFool algorithm is able to compute perturbations that are smaller than the perturbations computed by FGSM [23] in terms of their norm, while having similar fooling ratios.

3.1.8 Universal Adversarial Perturbations

Whereas the methods like FGSM [23], ILCM [35], DeepFool [72] etc. compute perturbations to fool a network on a single image, the 'universal' adversarial perturbations computed by Moosavi-Dezfooli et al. [16] are able to fool a network on 'any' image with high probability. These image-agnostic perturbations also remain quasi-imperceptible for the human vision system, as can be observed in Fig. 1. To formally define these perturbations, let us assume that clean images are sampled from the distribution $\mathfrak{S}_c$. A perturbation $\boldsymbol{\rho}$ is 'universal' if it satisfies the following constraint:

$$\underset{\mathbf{I}_c \sim \mathfrak{S}_c}{\mathrm{P}} \Big(\mathcal{C}(\mathbf{I}_c) \neq \mathcal{C}(\mathbf{I}_c + \boldsymbol{\rho}) \Big) \geq \delta \quad \text{s.t.} \ ||\boldsymbol{\rho}||_p \leq \xi, \qquad (6)$$

where P(.) denotes the probability, $\delta \in (0, 1]$ is the fooling ratio, $||.||_p$ denotes the ℓ_p-norm and ξ is a pre-defined constant. The smaller the value of ξ, the harder it is to perceive the perturbation in the image. Strictly speaking, the perturbations that satisfy (6) should be referred to as (δ, ξ)-universal because of their strong dependence on the mentioned parameters. However, these perturbations are commonly referred to as the 'universal adversarial perturbations' in the literature.

The authors computed the universal perturbations by restricting their ℓ_2-norm as well as ℓ_∞-norm, and showed that the perturbations with their norms upper bounded by 4% of the respective image norms already achieved significant fooling ratios of around 0.8 or more for state-of-the-art image classifiers. Their iterative approach to compute a perturbation is related to the DeepFool strategy [72] of gradually pushing a data point (i.e. an image) to the decision boundary for its class. However, in this case, 'all' the training data points are sequentially pushed to the respective decision boundaries and the perturbations computed over all the images are gradually accumulated by back-projecting the accumulator to the desired ℓ_p ball of radius ξ every time.

The algorithm proposed by Moosavi-Dezfooli et al. [16] computes perturbations while targeting a single network model, e.g. ResNet [147]. However, it is shown that these perturbations also generalize well across different networks (especially those having similar architectures). In that sense, the author's claim the perturbations to be, to some extent, 'doubly universal'. Moreover, it is also shown that high fooling ratio (e.g. $\delta \geq 0.5$) is achievable by learning a perturbation using only around $2,000$ training images.

Khrulkov et al. [190] also proposed a method for constructing universal adversarial perturbations as singular vectors of the Jacobian matrices of feature maps of the networks, which allowed for achieving relatively high fooling rates using only a small number of images. Another method to generate universal perturbations is fast-feature-fool by Mopuri et al. [135]. Their method generates the universal perturbations independent of data.

3.1.9 UPSET and ANGRI

Sarkar et al. [146] proposed two black-box attack algorithms, namely UPSET: Universal Perturbations for Steering to Exact Targets, and ANGRI: Antagonistic Network for Generating Rogue Images for targeted fooling of deep neural networks. For 'n' classes, UPSET seeks to produce 'n' image-agnostic perturbations such that when the perturbation is added to an image that does not belong to a targeted class, the classifier will classify the perturbed image as being from that class. The power of UPSET comes from a residual generating network R(.), that takes the target class 't' as input and produces a perturbation R(t) for fooling. The overall method solves the following optimization problem using the so-called UPSET network:

$$\mathbf{I}_{\boldsymbol{\rho}} = \max(\min(s\mathrm{R}(\mathrm{t}) + \mathbf{I}_c, 1), -1), \tag{7}$$

where the pixel values in $\mathbf{I}_c$ are normalized to lie in $[-1, 1]$, and 's' is a scalar. To ensure $\mathbf{I}_{\boldsymbol{\rho}}$ to be a valid image, all values outside the interval $[-1, 1]$ are clipped. As compared to the image-agnostic perturbations of UPSET, ANGRI computes image-specific perturbations in a closely related manner, for which we refer to the original work. The perturbations resulting from ANGRI are also used for targeted fooling. Both algorithms have been reported to achieve high fooling ratios on MNIST [10] and CIFAR-10 [152] datasets.

3.1.10 Houdini

Cisse et al. [131] proposed 'Houdini'- an approach for fooling gradient-based learning machines by generating adversarial examples that can be tailored to task losses. Typical algorithms that generate adversarial examples employ gradients of differentiable loss functions of the networks to compute the perturbations. However, task losses are often not amenable to this approach. For instance, the task loss of speech recognition is based on word-error-rate, which does not allow straightforward exploitation of loss function gradient. Houdini is tailored to generate adversarial examples for such tasks. Besides successful generation of adversarial images for classification, Houdini has also been shown to successfully attack a popular deep Automatic Speech Recognition system [151]. The authors have also demonstrated the transferability of attacks in speech recognition by fooling Google Voice in a black-box attack scenario. Moreover, successful targeted and non-targeted attacks are also demonstrated for a deep learning model for human pose estimation.

3.1.11 Adversarial Transformation Networks (ATNs)

Baluja and Fischer [42] trained feed-forward neural networks to generate adversarial examples against other targeted networks or set of networks. The trained models were termed Adversarial Transformation Networks (ATNs). The adversarial examples generated by these networks are computed by minimizing a joint loss function comprising of two parts. The first part restricts the adversarial example to have perceptual similarity with the original image, whereas the second part aims at altering the prediction of the targeted network on the resulting image.

Along the same direction, Hayex and Danezis [47] also used an attacker neural network to learn adversarial examples for black-box attacks. In the presented results, the examples computed by the attacker network remain perceptually indistinguishable from the clean images but they are misclassified by the targeted networks with overwhelming probabilities - reducing classification accuracy from 99.4% to 0.77% on MNIST data [10], and from 91.4% to 6.8% on the CIFAR-10 dataset [152].

3.1.12 Miscellaneous Attacks

The adversarial attacks discussed above are either the popular ones in the recent literature or they are representative of the research directions that are fast becoming popular. A summary of the main attributes of these attacks is also provided in Table 1. For a comprehensive study, below we provide brief descriptions of further techniques to generate adversarial attacks on deep neural networks. We note that this research area is currently highly active. Whereas every attempt has been made to review as many approaches as possible, we do not claim the review to be exhaustive. Due to high activity in this research direction, many more attacks are likely to surface in the near future.

Sabour et al. [26] showed the possibility of generating adversarial examples by altering the internal layers of deep neural networks. The authors demonstrated that it is possible to make internal network representation of adversarial images to resemble representations of images from different classes. Papernot et al. [109] studied transferability of adversarial attacks for deep learning as well as other machine learning techniques and introduced further transferability attacks. Narodytska and Kasiviswanathan [54] also introduced further black-box attacks that have been found effective in fooling the neural networks by changing only few pixel values in the images. Liu et al. [31] introduced 'epsilon-neighborhood' attack that have been shown to fool defensively distilled networks [108] with 100% success for white-box attacks. Oh et al. [133] took a 'Game Theory' perspective on adversarial attacks and derived a strategy to counter the counter-measures taken against adversarial attacks on deep neural networks. Mpouri et al. [135] developed a data-independent approach to generate universal adversarial perturbations for the deep network models. Hosseini et al. [98] introduced the notion of 'semantic adversarial examples' - input images that represent semantically same objects for humans but deep neural networks misclassify them. They used negatives of the images as semantic adversarial examples. Kanbak et al. [73] introduced 'ManiFool' algorithm in the wake of DeepFool method [72] to measure robustness of deep neural networks against geometrically perturbed images. Dong et al. [170] proposed an iterative method to boost adversarial attacks for black-box scenarios. Recently, Carlini and Wagner [59] also demonstrated that ten different defenses against perturbations can again be

TABLE 1: Summary of the attributes of diverse attacking methods: The 'perturbation norm' indicates the restricted ℓ_p-norm of the perturbations to make them imperceptible. The strength (higher for more asterisks) is based on the impression from the reviewed literature.

Method	Black/White box	Targeted/Non-targeted	Specific/Universal	Perturbation norm	Learning	Strength
L-BFGS [22]	White box	Targeted	Image specific	ℓ_∞	One shot	$***$
FGSM [23]	White box	Targeted	Image specific	ℓ_∞	One shot	$***$
BIM & ILCM [35]	White box	Non targeted	Image specific	ℓ_∞	Iterative	$****$
JSMA [60]	White box	Targeted	Image specific	ℓ_0	Iterative	$***$
One-pixel [68]	Black box	Non Targeted	Image specific	ℓ_0	Iterative	$**$
C&W attacks [36]	White box	Targeted	Image specific	$\ell_0, \ell_2, \ell_\infty$	Iterative	$*****$
DeepFool [72]	White box	Non targeted	Image specific	ℓ_2, ℓ_∞	Iterative	$****$
Uni. perturbations [16]	White box	Non targeted	Universal	ℓ_2, ℓ_∞	Iterative	$*****$
UPSET [146]	Black box	Targeted	Universal	ℓ_∞	Iterative	$****$
ANGRI [146]	Black box	Targeted	Image specific	ℓ_∞	Iterative	$****$
Houdini [131]	Black box	Targeted	Image specific	ℓ_2, ℓ_∞	Iterative	$****$
ATNs [42]	White box	Targeted	Image specific	ℓ_∞	Iterative	$****$

defeated by new attacks constructed using new loss functions. Rozsa et al. [94] also proposed a 'hot/cold' method for generating multiple possible adversarial examples for a single image. Interestingly, adversarial perturbations are not only being added to images to reduces the accuracy of deep learning classifiers. Yoo et al. [195] recently proposed an approach to also slightly improve the classification performance with the help of subtle perturbation to images.

We note that the authors of many works reviewed in this article have made the source code of their implementations publicly available. This is one of the major reasons behind the current rise in this research direction. Beside those resources, there are also libraries, e.g. Cleverhans [111], [112] that have started emerging in order to further boost this research direction. Adversarial-Playground (https://github.com/QData/AdversarialDNN-Playground) is another example of a toolbox made public by Norton and Qi [142] to understand adversarial attacks.

3.2 Attacks beyond classification/recognition

With the exception of Houdini [131], all the mainstream adversarial attacks reviewed in Section 3.1 directly focused on the task of classification - typically fooling CNN-based [10] classifiers. However, due to the seriousness of adversarial threats, attacks are also being actively investigated beyond the classification/recognition task in Computer Vision. Below, we review the works that develop approaches to attack deep neural networks beyond classification.

3.2.1 Attacks on Autoencoders and Generative Models

Tabacof et al. [128] investigated adversarial attacks for autoencoders [154], and proposed a technique to distort input image (to make it adversarial) that misleads the autoencoder to reconstruct a completely different image. Their approach attacks the internal representation of a neural network such that the representation for the adversarial image becomes similar to that of the target image. However, it is reported in [128] that autoencoders seem to be much more robust to adversarial attacks than the typical classifier networks. Kos et al. [121] also explored methods for computing adversarial examples for deep generative models, e.g. variational autoencoder (VAE) and the VAE-Generative Adversarial Networks (VAE-GANs). GANs, such as [153] are becoming exceedingly popular now-a-days in Computer Vision applications due to their ability to learn data distributions and generate realistic images using those distributions. The authors introduced three different classes of attacks for VAE and VAE-GANs. Owing to the success of these attacks it is concluded that the deep generative models are also vulnerable to adversaries that can convince them to turn inputs into very different outputs. This work adds further support to the hypothesis that "adversarial examples are a general phenomenon for current neural network architectures".

3.2.2 Attack on Recurrent Neural Networks

Papernot et al. [110] successfully generated adversarial input sequences for Recurrent Neural Networks (RNNs). RNNs are deep learning models that are particularly suitable for learning mappings between sequential inputs and outputs [155]. Papernot et al. demonstrated that the algorithms proposed to compute adversarial examples for the feed-forward neural networks (e.g. FGSM [23]) can also be adapted for fooling RNNs. In particular, the authors demonstrated successful fooling of the popular Long-Short-Term-Memory (LSTM) RNN architecture [156]. It is concluded that the cyclic neural network model like RNNs are also not immune to the adversarial perturbations that were originally uncovered in the context of acyclic neural networks, i.e. CNNs.

3.2.3 Attacks on Deep Reinforcement Learning

Lin et al. [134] proposed two different adversarial attacks for the agents trained by deep reinforcement learning [157]. In the first attack, called 'strategically-timed attack', the adversary minimizes the reward of the agent by attacking it at a small subset of time steps in an episode. A method is proposed to determine when an adversarial example should be crafted and applied, which enables the attack to go undetected. In the second attack, referred as 'enchanting attack', the adversary lures the agent to a designated target state by integrating a generative model and a planning algorithm. The generative model is used for predicting the future states of the agent, whereas the planning algorithm generates the actions for luring it. The attacks are successfully tested against the agents trained by the state-of-the-art deep reinforcement learning algorithms [157], [158]. Details on this work and example videos of the adversarial attacks can be found on the following URL: http://yclin.me/adversarial_attack_RL/.

In another work, Huang et al. [62] demonstrated that FGSM [23] can also be used to significantly degrade performance of trained policies in the context of deep reinforcement learning. Their threat model considers adversaries that are capable of introducing minor perturbations to the raw input of the policy. The conducted experiments demonstrate that it is fairly easy to confuse neural network policies with adversarial examples, even in black-box scenarios. Videos and further details on this work are available on http://rll.berkeley.edu/adversarial/.

3.2.4 Attacks on Semantic Segmentation and Object Detection

Semantic image segmentation and object detection are among the mainstream problems in Computer Vision. Inspired by Moosavi-Dezfooli [16], Metzen et al. [67] showed the existence of image-agnostic quasi-imperceptible perturbations that can fool a deep neural network into significantly corrupting the predicted segmentation of the images. Moreover, they also showed that it is possible to compute noise vectors that can remove a specific class from the segmented classes while keeping most of the image segmentation unchanged (e.g. removing pedestrians from road scenes). Although it is argued that the "space of the adversarial perturbations for the semantic image segmentation is presumably smaller than image classification", the perturbations have been shown to generalize well for unseen validation images with high probability. Arnab et al. [51] also evaluated FGSM [23] based adversarial attacks for semantic segmentation and noted that many observations about these attacks for classification do not directly transfer to segmentation task.

Xie et al. [115] computed adversarial examples for semantic segmentation and object detection under the observation that these tasks can be formulated as classifying multiple targets in an image - the target is a pixel or a receptive field in segmentation, and object proposal in detection. Under this perspective, their approach, called 'Dense Adversary Generation' optimizes a loss function over a set of pixels/proposals to generate adversarial examples. The generated examples are tested to fool a variety of deep learning based segmentation and detection approaches. Their experimental evaluation not only demonstrates successful fooling of the targeted networks but also shows that the generated perturbations generalize well across different network models. In Fig. 4, we show a representative example of network fooling for segmentation and detection using the approach in [115].

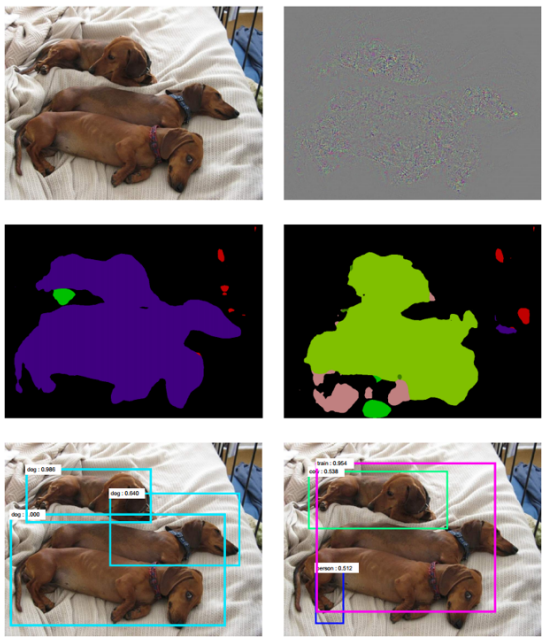


Fig. 4: Adversarial example for semantic segmentation and object detection [115]. FCN [159] and Faster-RCNN [150] are used for segmentation and detection, respectively. Left column (top-down): Clean image, normal segmentation (purple region is predicted as dog) and detection results. Right column (top-down): Perturbation 10x, fooled segmentation (light green region is predicted as train and the pink region as person) and detection results.

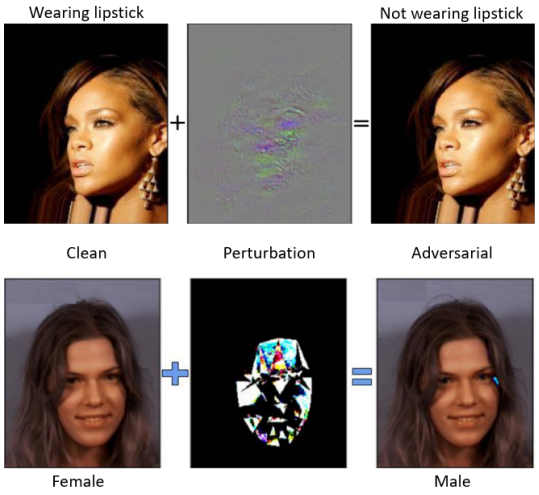


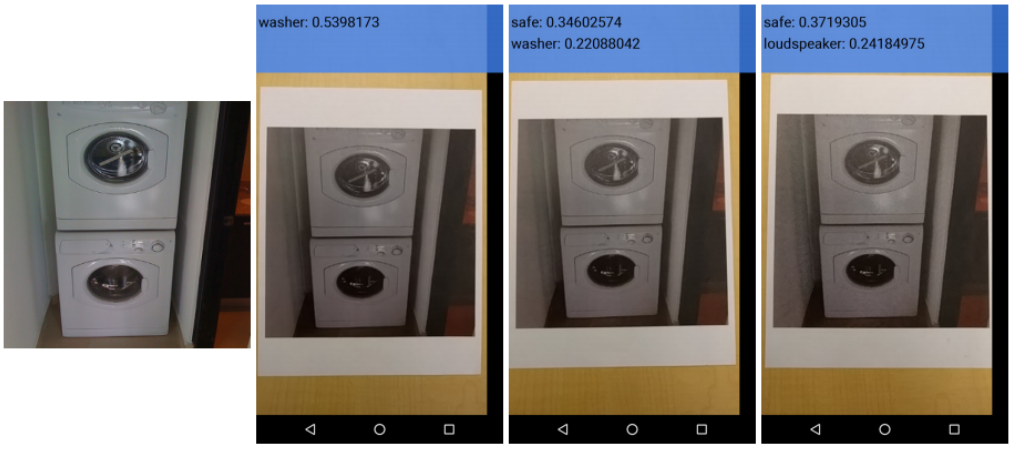
Fig. 5: Top-row: Example of changing a facial attribute 'wearing lipstick' to 'not wearing lipstick' by Fast Flipping Attribute method [130]. Bottom row: Changing gender with perturbation generated by [162].

4 Attacks in the Real World

4.0.1 Attacks on Face Attributes

Face attributes are among the emerging soft biometrics for modern security systems. Although face attribute recognition can also be categorized as a classification problem, we separately review some interesting attacks in this direction because face recognition itself is treated as a mainstream problem in Computer Vision.

Rozsa et al. [130], [160] explored the stability of multiple deep learning approaches using the CelebA benchmark [161] by generating adversarial examples to alter the results of facial attribute recognition, see top-row in Fig. 5. By attacking the deep network classifiers with their so-called 'Fast Flipping Attribute' technique, they found that robustness of deep neural networks against the adversarial attacks varies highly between facial attributes. It is claimed that adversarial attacks are very effective in changing the label of a target attribute to a correlated attribute. Mirjalili and Ross [162] proposed a technique that modifies a



(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

Fig. 6: Example of adversarial attack on mobile phone cameras: A clean image (a) was taken and used to generate different adversarial images. The images were printed and the TensorFlow Camera Demo app [181] was used to classify them. A clean image (b) is recognized correctly as a 'washer' when perceived through the camera, whereas adversarial images (c) and (d) are mis-classified. The images also show network confidence in the range [0,1] for each image. The value of ϵ is given for (3).

face image such that its gender (for a gender classifier) is modified, whereas its biometric utility for a face matching system remains intact, see bottom-row in Fig. 5. Similarly, Shen et al. [144] proposed two different techniques to generate adversarial examples for faces that can have high 'attractiveness scores' but low 'subjective scores' for the face attractiveness evaluation using deep neural network. We refer to [185] for further attacks related to the task of face recognition.

The literature reviewed in Section 3 assumes settings where adversaries directly feed deep neural networks with perturbed images. Moreover, the effectiveness of attacks are also evaluated using standard image databases. Whereas those settings have proven sufficient to convince many researchers that adversarial attacks are a real concern for deep learning in practice, we also come across instances in the literature (e.g. [48], [30]) where this concern is downplayed and adversarial examples are implicated to be 'only a matter of curiosity' with little practical concerns. Therefore, this Section is specifically dedicated to the literature that deals with the adversarial attacks in practical real-world conditions to help settle the debate.

4.1 Cell-phone camera attack

Kurakin et al. [35] first demonstrated that threats of adversarial attacks also exist in the physical world. To illustrate this, they printed adversarial images and took snapshots from a cell-phone camera. These images were fed to TensorFlow Camera Demo app [181] that uses Google's Inception model [145] for object classification. It was shown that a large fraction of images were misclassified even when perceived through the camera. In Fig. 6, an example is shown from the original paper. A video is also provided on the following URL https://youtu.be/zQ_uMenoBCk that shows the threat of adversarial attacks with further images. This work studies FGSM [23], BIM and ILCM [35] methods for attacks in the physical world.

4.2 Road sign attack

Etimov et al. [75] built on the attacks proposed in [36] and [88] to design robust perturbations for the physical world. They demonstrated the possibility of attacks that are robust to physical conditions, such as variation in view angles, distance and resolution. The proposed algorithm, termed RP_2 for Robust Physical Perturbations, was used to generate adversarial examples for road sign recognition systems that achieved high fooling ratios in practical drive-by settings. Two attack classes were introduced in this work for the physical road signs, (a) poster-printing: where the attacker prints a perturbed road sign poster and places it over the real sign (see Fig. 7), (b) sticker perturbation: where the printing is done on a paper and the paper is stuck over the real sign. For (b) two types of perturbations were studied, (b1) subtle perturbations: that occupied the entire sign and (b2) camouflage perturbations: that took the form of graffiti sticker on the sign. As such, all these perturbations require access to a color printer and no other special hardware. Successful generation of perturbations for both (a) and (b) such that the perturbations remained robust to natural variations in the physical world demonstrate the threat of adversarial example in the real world. We refer to the following URL for further details and videos related to this work: https://iotsecurity.eecs.umich.edu/#roadsigns.

It should be noted that Lu et al. [30] had previously claimed that adversarial examples are not a concern for object detection in Autonomous Vehicles because of the changing physical conditions in a moving car. However, the attacking methods they employed [22], [23], [35] were somewhat primitive. The findings of Etimov et al. [75] are orthogonal to the results in [66]. However, in a follow-up work Lu et al. [19] showed that the detectors like YOLO 9000 [149] and FasterRCNN [150] are 'currently' not fooled by the attacks introduced by Etimov et al. [75]. Zeng et al. [87] also argue that adversarial perturbations in the image space do not generalize well in the physical space

Fig. 7: Example of road sign attack [75]: The success rate of fooling LISA-CNN [75] classifier on all the shown images is 100%. The distance and angle to the camera are also shown. The classifier is trained using LISA dataset for road signs [176].

of the real-world. However, Athalye et al. [65] showed that we can actually print 3D physical objects for successful adversarial attacks in the physical world. We discuss [65] in Section 4.3.

Gu et al. [33] also explored an interesting notion of threats to outsourced training of the neural networks in the context of fooling neural networks on street signs. They showed that it is possible to train a network (a *BadNet*) that shows state-of-the-art performance on the user's training and validation samples, but behaves badly on attacker-chosen inputs. They demonstrated this attack in a realistic scenario by creating a street sign classifier that identifies stop signs as speed limits when a special sticker is added to the stop sign. Moreover, it was found that the fooling of the network persisted to a reasonable extent even when the network was later fine-tuned with additional training data.

4.3 Generic adversarial 3D objects

Athalye et al. [65] introduced a method for constructing 3D objects that can fool neural networks across a wide variety of angles and viewpoints. Their 'Expectation Over Transformation' (EOT) framework is able to construct examples that are adversarial over an entire distribution of image/object transformations. Their end-to-end approach is able to print arbitrary adversarial 3D objects. In our opinion, results of this work ascertain that adversarial attacks are a real concern for deep learning in the physical world. In Fig. 8 we show an example of 3D-printed turtle that is modified by EOT framework to be classified as rifle. A video demonstrating the fooling by EOT in the physical world is available at the following URL: https://www.youtube.com/watch?v=YXy6oX1iNoA&feature=youtu.be.

4.4 Cyberspace attacks

Papernot et al. [39] launched one of the first attacks against the deep neural network classifiers in cyberspace in the real-world settings. They trained a substitute network for the targeted black-box classifier on synthetic data, and instantiated the attack against remotely hosted neural networks by MetaMind, Amazon and Google. They were able to show that the respective targeted networks misclassified 84.24%, 96.19% and 88.94% of the adversarial examples generated by their method. Indeed, the only information available to the attacker in their threat model was the output label of the targeted network for the input image fed by the attacker. In a related work, Liu et al. [88] developed an ensemble based attack and showed its success against Clarifai.com - a commercial company providing state-of-the-art image classification services. The authors claim that their attacks for both targeted and non-targeted fooling are able to achieve high success rates.

Grosse et al. [61] showed construction of effective adversarial attacks for neural networks used as malware classifiers. As compared to image recognition, the domain of malware classification introduces additional constraints in the adversarial settings, e.g. continuous input domains are replaced by discrete inputs, the condition of visual similarity is replaced by requiring equivalent functional behavior. However, Grosse et al. [61] showed that creating effective adversarial examples is still possible for maleware classification. Further examples of successful adversarial attacks against deep lrearning based malware classification can also be found in [64], [107], [125].

4.5 Robotic Vision & Visual QA Attacks

Melis et al. [63] demonstrated the vulnerability of robots to the adversarial manipulations of the input images using the techniques in [22]. The authors argue that strategies to enforce deep neural networks to learn more stable representations are necessary for secure robotics. Xu et al. [40] generated adversarial attacks for the Visual Turing Test, also known as 'Visual Question Answer' (VQA). The authors show that the commonly used compositional and non-compositional VQA architectures that employ deep neural networks are vulnerable to adversarial attacks. Moreover, the adversarial examples are transferable between the models. They conclude that the "adversarial examples pose real threats to not only image classification models, but also more complicated VQA models" [63].

5 Existence of Adversarial Examples

In the literature related to adversarial attacks on deep learning in Computer Vision, there are varied views on the existence of adversarial examples. These views generally align well with the local empirical observations made by the researchers while attacking or defending the deep neural networks. However, they often fall short in terms of generalization. For instance, the popular linearity hypothesis of Goodfellow et al. [23] explains the FGSM and related attacks very well. However, Tanay and Griffin [74] demonstrated image classes that do not suffer from adversarial examples for linear classifier, which is not in-line with the linearity hypothesis. Not to mention, the linearity hypothesis itself deviates strongly from the previously prevailing opinion

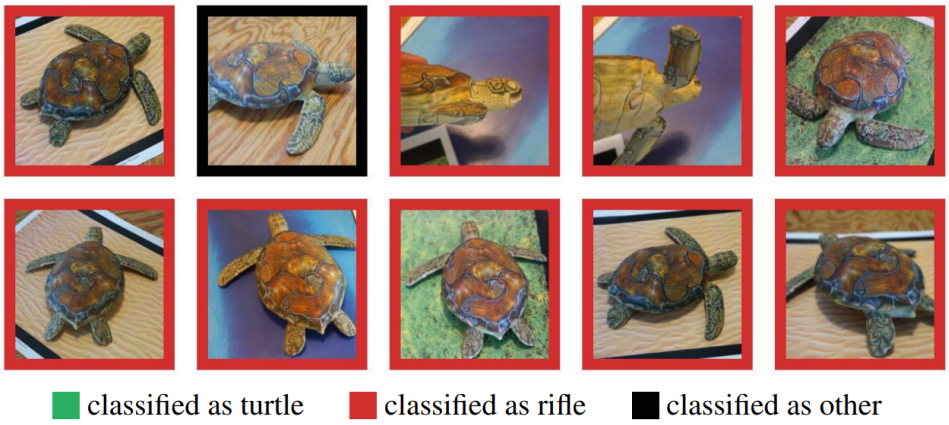


Fig. 8: Different random poses of a 3D-printed turtle perturbed by EOT [65] to be classified as a rifle by an ImageNet classifier. The unperturbed version (not shown) is classified correctly with 100% probability.

that the adversarial examples stem from highly non-linear decision boundaries induced by deep neural networks. There are also other examples in the literature where the linearity hypothesis is not directly supported [119].

Flatness of decision boundaries [69], large local curvature of the decision boundaries [70] and low flexibility of the networks [71] are some more examples of the viewpoints on the existence of adversarial examples that do not perfectly align with each other. Whereas it is apparent that adversarial examples can be formed by modifying as little as one pixel in an image, current literature seems to lack consensus on the reasons for the existence of the adversarial examples. This fact also makes analysis of adversarial examples an active research direction that is expected to explore and explain the nature of the decision boundaries induced by deep neural networks, which are currently more commonly treated as black-box models. Below, we review the works that mainly focus on analyzing the existence of adversarial perturbations for deep learning. We note that, besides the literature reviewed below, works related to adversarial attacks (Section 3) and defenses (Section 6) often provide brief analysis of adversarial perturbations while conjecturing about the phenomena resulting in the existence of the adversarial examples.

5.1 Limits on adversarial robustness

Fawzi et al. [118] introduced a framework for studying the instability of classifiers to adversarial perturbations. They established fundamental limits on the robustness of classifiers in terms of a 'distinguishability measure' between the classes of the dataset, where distinguishability is defined as the distance between the means of two classes for linear classifiers and the distance between the matrices of second order moments for the studied non-linear classifiers. This work shows that adversarial examples also exist for the classifiers beyond deep neural networks. The presented analysis traces back the phenomenon of adversarial instability to the low flexibility of the classifiers, which is not completely orthogonal to the prevailing belief at that time that high-nonlinearity of the networks make them susceptible to adversarial examples.

5.2 Space of adversarial examples

Tabacof and Eduardo [25] generated adversarial examples for shallow and deep network classifiers on MNIST [10] and ImageNet [11] datasets and probed the pixel space of adversarial examples by using noise of varying distribution and intensity. The authors empirically demonstrated that adversarial examples appear in large regions in the pixel space, which is in-line with the similar claim in [23]. However, somewhat in contrast to the linearity hypothesis, they argue that a weak, shallow and more linear classifier is also as susceptible to adversarial examples as a strong deep classifier.

Tramer et al. [132] proposed a method to estimate the dimensionality of the space of the adversarial examples. It is claimed that the adversarial examples span a contiguous high dimension space (e.g. with dimensionality ≈ 25). Due to high dimensionality, the subspaces of different classifiers can intersect, which gives rise to the transferability of the adversarial examples. Interestingly, their analysis suggests that it is possible to defend classifiers against transfer-based attacks even when they are vulnerable to direct attacks.

5.3 Boundary tilting perspective

Tanay and Griffin [74] provided a 'boundary tilting' perspective on the existence of adversarial examples for deep neural networks. They argued that generally a single class data that is sampled to learn and evaluate a classifier lives in a sub-manifold of the class, and adversarial examples for that class exist when the classification boundary lies close to this sub-manifold. They formalized the notion of 'adversarial strength' of a classifier and reduced it to the 'deviation angle' between the boundaries of the considered classifier and the nearest centroid classifier. It is then shown that adversarial strength of a classifier can be varied by decision 'boundary tilting'. The authors also argued that adversarial stability of the classifier is associated with its regularization. In the opinion of Tanay and Griffin, the linearity hypothesis [23] about the existence of adversarial examples is "unconvincing".

5.4 Prediction uncertainty and evolutionary stalling of training cause adversaries

Cubuk et al. [91] argue that the "origin of adversarial examples is primarily due to an inherent uncertainty that neural networks have about their predictions". They empirically compute a functional form of the uncertainty, which is shown to be independent of network architecture, training protocol and dataset. It is argued that this form only depends on the statistics of the network logit differences. This eventually results in fooling ratios caused by adversarial attacks to exhibit a universal scaling with respect to the size of perturbation. They studied FGSM [23], ILCM and BIM [35] based attacks to corroborate their claims. It is also claimed that accuracy of a network on clean images correlates with its adversarial robustness (see Section 5.5 for more arguments in this direction).

Rozsa et al. [102] hypothesized that the existence of adversarial perturbations is a result of evolutionary stalling of decision boundaries on training images. In their opinion, individual training samples stop contributing to the training loss of the model (i.e. neural network) once they are classified correctly, which can eventually leave them close to the decision boundary. Hence, it becomes possible to throw those (and similar) samples away to a wrong class region by adding minor perturbations. They proposed a Batch Adjusted Network Gradients (BANG) algorithm to train a network to mitigate the evolutionary stalling during training.

5.5 Accuracy-adversarial robustness correlation

In the quest of explaining the existence of adversarial perturbations, Rozsa et al. [97] empirically analyzed the correlation between the accuracy of eight deep network classifiers and their robustness to three adversarial attacks introduced in [23], [94]. The studied classifiers include AlexNet [9], VGG-16 and VGG-19 networks [163], Berkeley-trained version of GoogLeNet and Princeton-GoogLeNet [18], ResNet-52; ResNet-101; and ResNet-152 [147]. The adversarial examples are generated with the help of large-scale ImageNet dataset [11] using the techniques proposed in [23] and [94]. Their experiments lead to the observation that the networks with higher classification accuracy generally also exhibit more robustness against the adversarial examples. They also concluded that adversarial examples transfer better between similar network topologies.

5.6 More on linearity as the source

Kortov and Hopfiled [127] examined the existence of adversarial perturbations in the context of Dense Associative Memory (DAM) models [164]. As compared to the typical modern deep neural networks, DAM models employ higher order (more than quadratic) interactions between the neurons. The authors have demonstrated that adversarial examples generated using DAM models with smaller interaction power, which is similar to using a deep neural network with ReLU activation [165] for inducing linearity, are unable to fool models having higher order interactions. The authors provided empirical evidence on the existence of adversarial examples that is independent of the FGSM [23] attack, yet supports the linearity hypothesis of Goodfellow et al. [23].

5.7 Existence of universal perturbations

Moosavi-Dezfooli et al. [16] initially argued that universal adversarial perturbations exploit geometric correlations between the decision boundaries induced by the classifiers. Their existence partly owes to a subspace containing normals to the decision boundaries, such that the normals also surround the natural images. In [70], they built further on their theory and showed the existence of common directions (shared across datapoints) along which the decision boundary of a classifier can be highly positively curved. They argue that such directions play a key role in the existence of universal perturbations. Based on their findings, the authors also propose a new geometric method to efficiently compute universal adversarial perturbations.

It is worth noting that previously Fawzi et al. [69] also associated the theoretical bounds on the robustness of classifiers to the curvature of decision boundaries. Similarly, Tramer et al. [77] also held the curvature of decision boundaries in the vicinity of data points responsible for the vulnerability of neural networks to black-box attacks. In another recent work, Mopuri et al. [193] present a GAN-like model to learn the distribution of the universal adversarial perturbations for a given target model. The learned distributions are also observed to show good transferability across models.

6 Defenses against adversarial attacks

Currently, the defenses against the adversarial attacks are being developed along three main directions:

1) Using *modified training* during learning or *modified input* during testing.
2) *Modifying networks*, e.g. by adding more layers/sub-networks, changing loss/activation functions etc.
3) Using external models as *network add-on* when classifying unseen examples.

The approaches along the first direction do not directly deal with the learning models. On the other hand, the other two categories are more concerned with the neural networks themselves. The techniques under these categories can be further divided into two types; namely (a) complete defense and (b) detection only. The 'complete defense' approaches aim at enabling the targeted network to achieve its original goal on the adversarial examples, e.g. a classifier predicting labels of adversarial examples with acceptable accuracy. On the other hand, 'detection only' approaches are meant to raise the red flag on potentially adversarial examples to reject them in any further processing. The taxonomy of the described categories is also shown in Fig. 9. The remaining section is organized according to this taxonomy. In the used taxonomy, the difference between 'modifying' a network and employing an 'add-on' is that the former makes changes to the original deep neural network architecture/parameters during training. On the other hand, the latter keeps the original model intact and appends external model(s) to it during testing.

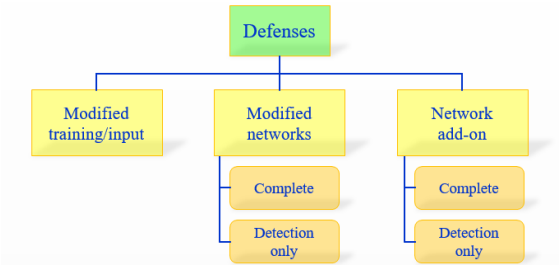


Fig. 9: Broad categorization of approaches aimed at defending deep neural networks against adversarial attacks.

6.1 Modified training/input

6.1.1 Brute-force adversarial training

Since the discovery of adversarial examples for the deep neural networks [22], there has been a general consensus in the related literature that robustness of neural networks against these examples improves with adversarial training. Therefore, most of the contributions introducing new adversarial attacks, e.g. [22], [23], [72] (see Section 3) simultaneously propose adversarial training as the first line of defense against those attacks. Although adversarial training improves robustness of a network, to be really effective, it requires that training is performed using strong attacks and the architecture of the network is sufficiently expressive. Since adversarial training necessitates increased training/data size, we refer to it as a 'brute-force' strategy.

It is also commonly observed in the literature that brute-force adversarial training results in regularizing the network (e.g. see [23], [90]) to reduce over-fitting, which in turn improves robustness of the networks against the adversarial attacks. Inspired by this observation, Miyato et al. [113] proposed a 'Virtual Adversarial Training' approach to smooth the output distributions of the neural networks. A related 'stability training' method is also proposed by Zheng et al. [116] to improve the robustness of neural networks against small distortions to input images. It is noteworthy that whereas adversarial training is known to improve robustness of neural networks, Moosavi-Dezfooli [16] showed that effective adversarial examples can again be computed for already adversarially trained networks.

6.1.2 Data compression as defense

Dziugaite et al. [123] noted that most of the popular image classification datasets comprise JPG images. Motivated by this observation, they studied the effects of JPG compression on the perturbations computed by FGSM [23]. It is reported that JPG compression can actually reverse the drop in classification accuracy to a large extent for the FGSM perturbations. Nevertheless, it is concluded that compression alone is far from an effective defense. JPEG compression was also studied by Guo et al. [82] for mitigating the effectiveness of adversarial images. Das et al. [37] also took a similar approach and used JPEG compression to remove the high frequency components from images and proposed an ensemble-based technique to counter the adversarial attacks generated by FGSM [23] and DeepFool method [72]. Whereas encouraging results are reported in [37], there is no analysis provided for the stronger attacks, e.g. C&W attacks [36]. Moreover, Shin and Song [186] have demonstrated the existence of adversarial examples that can survive JPEG compression. Compression under Discrete Cosine Transform (DCT) was also found inadequate as a defense against the universal perturbations [16] in our previous work [81]. One major limitation of compression based defense is that larger compressions also result in loss of classification accuracy on clean images, whereas smaller compressions often do not adequately remove the adversarial perturbations.

In another related approach, Bhagoji et al. [169] proposed to compress input data using Principal Component Analysis for adversarial robustness. However, Xu et al. [140] noted that this compression also results in corrupting the spatial structure of the image, hence often adversely affecting the classification performance.

6.1.3 Foveation based defense

Luo et al. [119] demonstrated that significant robustness against the adversarial attacks using L-BFGS [22] and FGSM [23] is possible with 'foveation' mechanism - applying neural network in different regions of images. It is hypothesized that CNN-based classifiers trained on large datasets, such as ImageNet [11] are generally robust to scale and translation variations of objects in the images. However, this invariance does not extend to adversarial patterns in the images. This makes foveation as a viable option for reducing the effectiveness of adversarial attacks proposed in [22], [23]. However, foveation is yet to demonstrate its effectiveness against more powerful attacks.

6.1.4 Data randomization and other methods

Xie et al. [115] showed that random resizing of the adversarial examples reduces their effectiveness. Moreover, adding random padding to such examples also results in reducing the fooling rates of the networks. Wang et al. [138] transformed the input data with a separate data-transformation module to remove possible adversarial perturbations in images. In the literature, we also find evidence that data augmentation during training (e.g. Gaussian data augmentation [46]) also helps in improving robustness of neural networks to adversarial attacks, albeit only slightly.

6.2 Modifying the network

For the approaches that modify the neural networks for defense against the adversarial attacks, we first discuss the 'complete defense' approaches. The 'detection only' approaches are separately reviewed in Section 6.2.8.

6.2.1 Deep Contractive Networks

In the early attempts of making deep learning robust to adversarial attacks, Gu and Rigazio [24] introduced Deep Contractive Networks (DCN). It was shown that Denoising Auto Encoders [154] can reduce adversarial noise, however stacking them with the original networks can make the resulting network even more vulnerable to perturbations. Based on this observation, the training procedure of DCNs used a smoothness penalty similar to Contractive Auto Encoders [173]. Whereas reasonable robustness of DCNs was demonstrated against the L-BGFS [22] based attacks,

many stronger attacks have been introduced since DCNs were initially proposed. A related concept of using auto encoders for adversarial robustness of the neural networks can be also found in [141].

6.2.2 Gradient regularization/masking

Ross and Doshi-Velez [52] studied input gradient regularization [167] as a method for adversarial robustness. Their method trains differentiable models (e.g. deep neural networks) while penalizing the degree of variation resulting in the output with respect to change in the input. Implying, a small adversarial perturbation becomes unlikely to change the output of the trained model drastically. It is shown that this method, when combined with brute-force adversarial training, can result in very good robustness against attacks like FGSM [23] and JSMA [60]. However, each of these methods almost double the training complexity of a network, which is already prohibitive in many cases.

Previously, Lyu et al. [28] also used the notion of penalizing the gradient of loss function of network models with respect to the inputs to incorporate robustness in the networks against L-BFGS [22] and FGSM [23] based attacks. Similarly, Shaham et al. [27] attempted to improve the local stability of neural networks by minimizing the loss of a model over adversarial examples at each parameter update. They minimized the loss of their model over worst-case adversarial examples instead of the original data. In a related work, Nguyen and Sinha [44] introduced a masking based defense against C&W attack [36] by adding noise to the logit outputs of networks.

6.2.3 Defensive distillation

Papernot et al. [38] exploited the notion of 'distillation' [166] to make deep neural networks robust against adversarial attacks. Distillation was introduced by Hinton et al. [166] as a training procedure to transfer knowledge of a more complex network to a smaller network. The variant of the procedure introduced by Papernot et al. [38] essentially uses the knowledge of the network to improve its own robustness. The knowledge is extracted in the form of class probability vectors of the training data and it is fed back to train the original model. It is shown that doing so improves resilience of a network to small perturbation in the images. Further empirical evidence in this regard is also provided in [108]. Moreover, in a follow-up work, Papernot et al. [84] also extended the defensive distillation method by addressing the numerical instabilities that were encountered in [38]. It is worth noting that the 'Carlini and Wagner' (C&W) attacks [36] introduced in Section 3.1 are claimed to be successful against the defensive distillation technique. We also note that defensive distillation can also be seen as an example of 'gradient masking' technique. However, we describe it separately keeping in view its popularity in the literature.

6.2.4 Biologically inspired protection

Nayebi and Ganguli [124] demonstrated natural robustness of neural networks against adversarial attacks with highly non-linear activations (similar to nonlinear dendritic computations). It is noteworthy that the Dense Associative Memory models of Krotov and Hopfield [127] also work on a similar principle for robustness against the adversarial examples. Considering the linearity hypothesis of Goodfellow et al. [23], [124] and [127] seem to further the notion of susceptibility of modern neural networks to adversarial examples being the effect of linearity of activations. We note that Brendel and Bethge [187] claim that the attacks fail on the biologically inspired protection [124] due to numerical limitations of computations. Stabilizing the computations again allow successful attacks on the protected networks.

6.2.5 Parseval Networks

Cisse et al. [131] proposed 'Parseval' networks as a defense against the adversarial attacks. These networks employ a layer-wise regularization by controlling the global Lipschitz constant of the network. Considering that a network can be seen as a composition of functions (at each layer), robustification against small input perturbations is possible by maintaining a small Lipschitz constant for these functions. Cisse et al. proposed to do so by controlling the spectral norm of the weight matrices of the networks by parameterizing them with 'parseval tight frames' [172], hence the name 'Parseval' networks.

6.2.6 DeepCloak

Gao et al. [139] proposed to insert a masking layer immediately before the layer handling the classification. The added layer is explicitly trained by forward-passing clean and adversarial pair of images, and it encodes the differences between the output features of the previous layers for those image pairs. It is argued that the most dominant weights in the added layer correspond to the most sensitive features of the network (in terms of adversarial manipulation). Therefore, while classifying, those features are masked by forcing the dominant weights of the added layer to zero.

6.2.7 Miscellaneous approaches

Among other notable efforts in making neural networks robust to adversarial attacks, Zantedeschi et al. [46] proposed to use bounded ReLU [174] to reduce the effectiveness of adversarial patterns in the images. Jin et al. [120] introduced a feedforward CNN that used additive noise to mitigate the effects of adversarial examples. Sun et al. [56] prposed 'HyperNetworks' that use statistical filtering as a method to make the network robust. Madry et al. [55] studied adversarial defense from the perspective of robust optimization. They showed that adversarial training with a PGD adversary can successfully defend against a range of other adversaries. Later, Carlini et al. [59] also verified this observation. Na et al. [85] employed a network that is regularized with a unified embedding for classification and low-level similarity learning. The network is penalized using the distance between clean and the corresponding adversarial embeddings. Strauss et al. [89] studied ensemble of methods to defend a network against the perturbations. Kadran et al. [136] modified the output layer of a neural network to induce robustness against the adversarial attacks. Wang et al. [129], [122] developed adversary resistant neural networks by leveraging non-invertible data transformation in the network. Lee et al. [106] developed manifold regularized networks that use a training objective to minimizes the

difference between multi-layer embedding results of clean and adversarial images. Kotler and Wong [96] proposed to learn ReLU-based classifier that show robustness against small adversarial perturbations. They train a neural network that provably achieves high accuracy (¿90%) against any adversary in a canonical setting ($\epsilon = 0.1$ for ℓ_∞-norm perturbation on MNIST). Raghunathan et al. [189] studied the problem of defense for neural networks with one hidden layer. Their approach produces a network and a certificate on MNIST dataset such that no attack perturbing image pixels by at most $\epsilon = 0.1$ could results in more than 35% test error. Kolter and Wong [96] and Raghunathan et al. [189] are among very few provable methods in defense against the adversarial attacks. Given that these methods are computationally infeasible to apply on larger networks, the only defenses that have been extensively evaluated are those of Madry et al. [55] giving 89% accuracy against large epsilon (0.3/1) on MNIST and 45% for moderate epsilon (8/255) on CIFAR. Another thread of works that can be seen as adversarial attacks/defenses with guarantees is related to verification of deep neural networks, e.g. [191], [192]. In their approach OrOrbia et al. [194] show that many different proposals of adversarial training are instances of more general regularized objective, they termed DataGrad. The proposed DataGrad framework can be seen as an extension of layerwise contractive autoencoder penalty.

6.2.8 Detection Only approaches

SafetyNet: Lu et al. [66] hypothesized that adversarial examples produce different patterns of ReLU activations in (the late stages of) networks than what is produced by clean images. Based on this hypothesis, they proposed to append a Radial Basis Function SVM classifier to the targeted models such that the SVM uses discrete codes computed by the late stage ReLUs of the network. To detect perturbation in a test image, its code is compared against those of training samples using the SVM. Effective detection of adversarial examples generated by [23], [35], [72] is demonstrated by their framework, named SafetyNet.

Detector subnetwork: Metzen et al. [78] proposed to augment a targeted network with a subnetwork that is trained for a binary classification task of detecting adversarial perturbations in inputs. It is shown that appending such a network to the internal layers of a model and using adversarial training can help in detecting perturbations generated using FGSM [23], BIM [35] and DeepFool [72] methods. However, Lu et al. [66] later showed that this approach is again vulnerable to counter-counter measures.

Exploiting convolution filter statistics: Li and Li [105] used statistics of the convolution filters in CNN-based neural networks to classify the input images as clean or adversarial. A cascaded classifier is designed that uses these statistics, and it is shown to detect more than 85% adversarial images generated by the methods in [22], [114].

Additional class augmentation: Grosse et al. [57] proposed to augment the potentially targeted neural network model with an additional class in which the model is trained to classify all the adversarial examples. Hosseini et al. [32] also employed a similar strategy to detect black-box attacks.

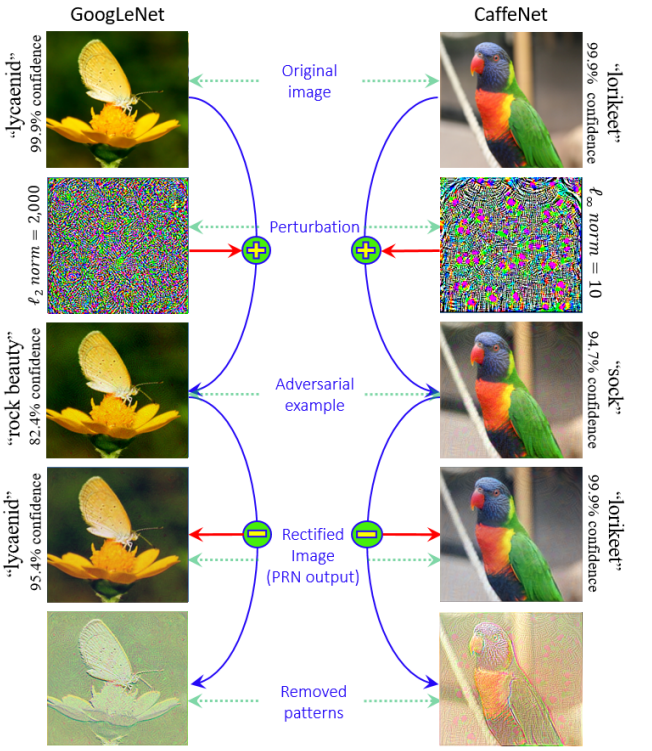


Fig. 10: Illustration of defense against universal perturbations [81]: The approach rectifies an image to restore the network prediction. The pattern removed by rectification is separately analyzed to detect perturbation.

6.3 Network add-ons

6.3.1 Defense against universal perturbations

Akhtar et al. [81] proposed a defense framework against the adversarial attacks generated using universal perturbations [16]. The framework appends extra 'pre-input' layers to the targeted network and trains them to rectify a perturbed image so that the classifier's prediction becomes the same as its prediction on the clean version of the same image. The pre-input layers are termed Perturbation Rectifying Network (PRN), and they are trained without updating the parameters of the targeted network. A separate detector is trained by extracting features from the input-output differences of PRN for the training images. A test image is first passed through the PRN and then its features are used to detect perturbations. If adversarial perturbations are detected, the output of PRN is used to classify the test image. Fig. 10, illustrates the rectification performed by PRN. The removed patterns are separately analyzed for detection.

6.3.2 GAN-based defense

Lee et al. [101] used the popular framework of Generative Adversarial Networks [153] to train a network that is robust to FGSM [23] like attacks. The authors proposed to directly train the network along a generator network that attempts to generate perturbation for that network. During its training,

the classifier keeps trying to correctly classify both the clean and perturbed images. We categorize this technique as an 'add-on' approach because the authors propose to always train any network in this fashion. In another GAN-based defense, Shen et al. [58] use the generator part of the network to rectify a perturbed image.

6.3.3 Detection Only approaches

Feature squeezing: Xu et al. [43] proposed to use feature squeezing to detect adversarial perturbation to an image. They added two external models to the classifier network, such that these models reduce the color bit depth of each pixel in the image, and perform spatial smoothing over the image. The predictions of the targeted network over the original image and the squeezed images are compared. If a large difference is found between the predictions, the image is considered to be an adversarial example. Whereas [43] demonstrated the effectiveness of this approach against more classical attacks [23], a follow-up work [140] also claims that the method works reasonably well against the more powerful C&W attacks [36]. He et al. [76] also combined feature squeezing with the ensemble method proposed in [175] to show that strength of defenses does not always increase by combining them.

MagNet: Meng and Chen [45] proposed a framework that uses one or more external detectors to classify an input image as adversarial or clean. During training, the framework aims at learning the manifold of clean images. In the testing phase, the images that are found far from the manifold are treated as adversarial and are rejected. The images that are close to the manifold (but not exactly on it) are always reformed to lie on the manifold and the classifier is fed with the reformed images. The notion of attracting nearby images to the manifold of clean images and dropping the far-off images also inspires the name of the framework, i.e. MagNet. It is noteworthy that Carlini and Wagner [188] very recently demonstrated that this defense technique can also be defeated with slightly larger perturbations.

Miscellaneous methods: Liang et al. [50] treated perturbations to images as noise and used scalar quantization and spatial smoothing filter to separately detect such perturbations. In a related approach, Feinman et al. [86] proposed to detect adversarial perturbations by harnessing uncertainty estimates (of dropout neural networks) and performing density estimation in the feature space of neural networks. Eventually, separate binary classifiers are trained as adversarial example detectors using the proposed features. Gebhart and Schrater [92] viewed neural network computation as information flow in graphs and proposed a method to detect adversarial perturbations by applying persistent homology to the induced graphs.

7 Outlook of the Research Direction

In the previous sections, we presented a comprehensive review of the recent literature in adversarial attacks on deep learning. Whereas several interesting facts were reported in those sections along the technical details, below we make more general observations regarding this emerging research direction. The discussion presents a broader outlook to the readers without in-depth technical knowledge of this area. Our arguments are based on the literature reviewed above.

The threat is real: Whereas few works suggest that adversarial attacks on deep learning may not be a serious concern, a large body of the related literature indicates otherwise. The literature reviewed in Sections 3 and 4 clearly demonstrate that adversarial attacks can severely degrade the performance of deep learning techniques on multiple Computer Vision tasks, and beyond. In particular, the literature reviewed in Section 4 ascertains that deep learning is vulnerable to adversarial attacks in the real physical world. Therefore, we can conclusively argue that adversarial attacks pose a real threat to deep learning in practice.

Adversarial vulnerability is a general phenomenon: The reviewed literature shows successful fooling of different types of deep neural networks, e.g. MLPs, CNNs, RNNs on a variety of tasks in Computer Vision, e.g. recognition, segmentation, detection. Although most of the existing works focus on fooling deep learning on the task of classification/recognition, based on the surveyed literature we can easily observe that deep learning approaches are vulnerable to adversarial attacks in general.

Adversarial examples often generalize well: One of the most common properties of adversarial examples reported in the literature is that they transfer well between different neural networks. This is especially true for the networks that have relatively similar architecture. The generalization of adversarial examples is often exploited in black-box attacks.

Reasons of adversarial vulnerability need more investigation: There are varied view-points in the literature on the reasons behind the vulnerability of deep neural networks to subtle adversarial perturbations. Often, these view-points are not well-aligned with each other. There is an obvious need for systematic investigation in this direction.

Linearity does promote vulnerability: Goodfellow et al. [23] first suggested that the design of modern deep neural networks that forces them to behave linearly in high dimensional spaces also makes them vulnerable to adversarial attacks. Although popular, this notion has also faced some opposition in the literature. Our survey pointed out multiple independent contributions that hold linearity of the neural networks accountable for their vulnerability to adversarial attacks. Based on this fact, we can argue that linearity does promote vulnerability of deep neural networks to the adversarial attacks. However, it does not seem to be the only reason behind successful fooling of deep neural networks with cheap analytical perturbations.

Counter-counter measures are possible: Whereas multiple defense techniques exist to counter adversarial attacks, it is often shown in the literature that a defended model can

again be successfully attacked by devising counter-counter measures, e.g. see [49]. This observation necessitates that new defenses also provide an estimate of their robustness against obvious counter-counter measures.

Highly active research direction: The profound implications of vulnerability of deep neural networks to adversarial perturbations have made research in adversarial attacks and their defenses highly active in recent time. The majority of the literature reviewed in this survey surfaced in the last two years, and there is currently a continuous influx of contributions in this direction. On one hand, techniques are being proposed to defend neural networks against the known attacks, on the other; more and mote powerful attacks are being devised. Recently, a Kaggle competition was also organized for the defenses against the adversarial attacks (https://www.kaggle.com/c/nips-2017-defense-against-adversarial-attack/). It can be hoped that this high research activity will eventually result in making deep learning approaches robust enough to be used in safety and security critical applications in the real world.

8 Conclusion

This article presented the first comprehensive survey in the direction of adversarial attacks on deep learning in Computer Vision. Despite the high accuracies of deep neural networks on a wide variety of Computer Vision tasks, they have been found vulnerable to subtle input perturbations that lead them to completely change their outputs. With deep learning at the heart of the current advances in machine learning and artificial intelligence, this finding has resulted in numerous recent contributions that devise adversarial attacks and their defenses for deep learning. This article reviews these contributions, mainly focusing on the most influential and interesting works in the literature. From the reviewed literature, it is apparent that adversarial attacks are a real threat to deep learning in practice, especially in safety and security critical applications. The existing literature demonstrates that currently deep learning can not only be effectively attacked in cyberspace but also in the physical world. However, owing to the very high activity in this research direction it can be hoped that deep learning will be able to show considerable robustness against the adversarial attacks in the future.

References

[1] Y. LeCun, Y. Bengio and G. Hinton, *Deep learning*, Nature, vol. 521, no. 7553, pp. 436-444, 2015.

[2] M. Helmstaedter, K. L. Briggman, S. C. Turaga, V. Jain, H. S. Seung, and W. Denk, *Connectomic reconstruction of the inner plexiform layer in the mouse retina*. Nature, vol. 500, no. 7461, pp. 168-174, 2013.

[3] H. Y. Xiong, B. Alipanahi, J. L. Lee, H. Bretschneider, D. Merico, R. K. Yuen, and Q. Morris, *The human splicing code reveals new insights into the genetic determinants of disease*, Science, vol. 347, no. 6218, 1254806 2015.

[4] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl and V. Svetnik, *Deep neural nets as a method for quantitative structure-activity relationships*, Journal of chemical information and modeling, vol. 55, no. 2 pp. 263-274, 2015.

[5] T. Ciodaro, D. Deva, J. de Seixas and D. Damazio, *Online particle detection with neural networks based on topological calorimetry information*. Journal of physics: conference series. vol. 368, no. 1. IOP Publishing, 2012.

[6] Kaggle. Higgs boson machine learning challenge. Kaggle https://www.kaggle.com/c/higgs-boson, 2014.

[7] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, and B. Kingsbury, *Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups*, IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97, 2012.

[8] I. Sutskever, O. Vinyals, and Q. V. Le, *Sequence to sequence learning with neural networks*. In Advances in neural information processing systems, pp. 3104-3112, 2014.

[9] A. Krizhevsky, I. Sutskever and G. E. Hinton, *Imagenet classification with deep convolutional neural networks*. In Advances in neural information processing systems, pp. 1097-1105, 2012.

[10] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel, *Backpropagation applied to handwritten zip code recognition*. Neural computation, vol. 1m no. 4, pp. 541-551, 1989.

[11] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, *Imagenet: A large-scale hierarchical image database*. In IEEE Conference on Computer Vision and Pattern Recognition, pp. 248-255, 2009.

[12] E. Ackerman, *How Drive.ai is mastering autonomous driving with deep learning*, https://spectrum.ieee.org/cars-that-think/transportation/self-driving/how-driveai-is-mastering-autonomous-driving-with-deep-learning, Accessed December 2017.

[13] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald and E. Muharemagic, *Deep learning applications and challenges in big data analytics*, Journal of Big Data, vol. 2, no. 1, 2015.

[14] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, Y. Chen, *Mastering the game of go without human knowledge*. Nature, vol. 550, no. 7676, pp. 354-359, 2017.

[15] K. He, X. Zhang, S. Ren, J. Sun, *Deep residual learning for image recognition*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.

[16] S. M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi and P. Frossard, *Universal adversarial perturbations*. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[17] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, *Return of the devil in the details: Delving deep into convolutional nets*, arXiv preprint arXiv:1405.3531, 2014.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, A. Rabinovich, *Going deeper with convolutions*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9, 2015.

[19] J. Lu, H. Sibai, E. Fabry, D. Forsyth, *Standard detectors aren't (currently) fooled by physical adversarial stop signs*, arXiv preprint arXiv:1710.03337, 2017.

[20] R. Fletcher, *Practical methods of optimization*, John Wiley and Sons, 2013.

[21] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, *Dermatologist-level classification of skin cancer with deep neural networks*, Nature, vol. 542, pp. 115 - 118, 2017.

[22] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, *Intriguing properties of neural networks*, arXiv preprint arXiv:1312.6199, 2014.

[23] I. J. Goodfellow, J. Shlens, C. Szegedy, *Explaining and Harnessing Adversarial Examples*, arXiv preprint arXiv:1412.6572, 2015.

[24] S. Gu, L. Rigazio, *Towards Deep Neural Network Architectures Robust to Adversarial Examples*, arXiv preprint arXiv:1412.5068, 2015

[25] P. Tabacof, E. Valle, *Exploring the Space of Adversarial Images*, In IEEE International Joint Conference on Neural Networks, pp. 426-433, 2016.

[26] S. Sabour, Y. Cao, F. Faghri, and D. J. Fleet, *Adversarial manipulation of deep representations*, arXiv preprint arXiv:1511.05122, 2015.

[27] U. Shaham, Y. Yamada, S. Negahban, *Understanding Adversarial Training: Increasing Local Stability of Neural Nets through Robust Optimization*, arXiv preprint arXiv:1511.05432, 2016.

[28] C. Lyu, K. Huang, H. Liang, *A Unified Gradient Regularization Family for Adversarial Examples*, In IEEE International Conference on Data Mining, pp. 301-309, 2015.

[29] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, D. Song, *Robust Physical-World Attacks on Deep Learning Models*, arXiv preprint arXiv:1707.08945, 2017.

[30] J. Lu, H. Sibai, E. Fabry, D. Forsyth, *No need to worry about adversarial examples in object detection in autonomous vehicles*, arXiv preprint arXiv:1707.03501, 2017.

[31] Y. Liu, W. Zhang, S. Li, N. Yu, *Enhanced Attacks on Defensively Distilled Deep Neural Networks*, arXiv preprint arXiv:1711.05934, 2017.

[32] H. Hosseini, Y. Chen, S. Kannan, B. Zhang, R. Poovendran, *Blocking transferability of adversarial examples in black-box learning systems*, arXiv preprint arXiv:1703.04318, 2017.

[33] T. Gu, B. Dolan-Gavitt, S. Garg, *BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain.* arXiv preprint arXiv:1708.06733, 2017.

[34] N. Papernot, P. McDaniel, A. Sinha, M. Wellman, *Towards the Science of Security and Privacy in Machine Learning*, arXiv preprint arXiv:1611.03814, 2016.

[35] A. Kurakin, I. Goodfellow, S. Bengio, *Adversarial examples in the physical world*, arXiv preprint arXiv:1607.02533, 2016.

[36] N. Carlini, D. Wagner, *Towards Evaluating the Robustness of Neural Networks*, arXiv preprint arXiv:1608.04644, 2016.

[37] N. Das, M. Shanbhogue, S. Chen, F. Hohman, L. Chen, M. E. Kounavis, D. H. Chau, *Keeping the Bad Guys Out: Protecting and Vaccinating Deep Learning with JPEG Compression*, arXiv preprint arXiv:1705.02900, 2017.

[38] N. Papernot, P. McDaniel, X. Wu, S. Jha, A. Swami, *Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks*, In IEEE Symposium on Security and Privacy (SP), pp. 582-597, 2016.

[39] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, A. Swami, *Practical Black-Box Attacks against Machine Learning*, In Proceedings of the ACM on Asia Conference on Computer and Communications Security, pp. 506-519. ACM, 2017.

[40] X. Xu, X. Chen, C. Liu, A. Rohrbach, T. Darell, D. Song, *Can you fool AI with adversarial examples on a visual Turing test?*, arXiv preprint arXiv:1709.08693, 2017

[41] P. Chen, H. Zhang, Y. Sharma, J. Yi, C. Hsieh, *ZOO: Zeroth Order Optimization based Black-box Attacks to Deep Neural Networks without Training Substitute Models*, In Proceedings of 10th ACM Workshop on Artificial Intelligence and Security (AISEC), 2017.

[42] S. Baluja, I. Fischer, *Adversarial Transformation Networks: Learning to Generate Adversarial Examples*, arXiv preprint arXiv:1703.09387, 2017.

[43] W. Xu, D. Evans, Y. Qi, *Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks*, arXiv preprint arXiv:1704.01155, 2017.

[44] L. Nguyen, A. Sinha, *A Learning and Masking Approach to Secure Learning*, arXiv preprint arXiv:1709.04447, 2017.

[45] Dongyu Meng, Hao Chen, *MagNet: a Two-Pronged Defense against Adversarial Examples*, In Proceedings of ACM Conference on Computer and Communications Security (CCS), 2017.

[46] V. Zantedeschi, M. Nicolae, A. Rawat, *Efficient Defenses Against Adversarial Attacks*, arXiv preprint arXiv:1707.06728, 2017.

[47] J. Hayes, G. Danezis, *Machine Learning as an Adversarial Service: Learning Black-Box Adversarial Examples*, arXiv preprint arXiv:1708.05207, 2017.

[48] A. Graese, A. Rozsa, T. E. Boult, *Assessing Threat of Adversarial Examples on Deep Neural Networks*, In IEEE International Conference on Machine Learning and Applications, pp. 69-74, 2016.

[49] N. Carlini, D. Wagner, *Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods*, arXiv preprint arXiv:1705.07263, 2017.

[50] B. Liang, H. Li, M. Su, X. Li, W. Shi, X. Wang, *Detecting Adversarial Examples in Deep Networks with Adaptive Noise Reduction*, arXiv preprint arXiv:1705.08378, 2017.

[51] A. Arnab, O. Miksik, P. H. S. Torr, *On the Robustness of Semantic Segmentation Models to Adversarial Attacks*, arXiv preprint arXiv:1711.09856, 2017.

[52] A. S. Ross, F. Doshi-Velez, *Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing their Input Gradients*, arXiv preprint arXiv:1711.09404, 2017.

[53] Y. Song, T. Kim, S. Nowozin, S. Ermon, N. Kushman, *PixelDefend: Leveraging Generative Models to Understand and Defend against Adversarial Examples*, arXiv preprint arXiv:1710.10766, 2017.

[54] N. Narodytska, S. P. Kasiviswanathan, *Simple Black-Box Adversarial Perturbations for Deep Networks*, arXiv preprint arXiv:1612.06299, 2016.

[55] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, *Towards Deep Learning Models Resistant to Adversarial Attacks*, arXiv preprint arXiv:1706.06083, 2017.

[56] Z. Sun, M. Ozay, T. Okatani, *HyperNetworks with statistical filtering for defending adversarial examples*, arXiv preprint arXiv:1711.01791, 2017.

[57] K. Grosse, P. Manoharan, N. Papernot, M. Backes, P. McDaniel, *On the (Statistical) Detection of Adversarial Examples*, arXiv preprint arXiv:1702.06280, 2017.

[58] S. Shen, G. Jin, K. Gao, Y. Zhang, *APE-GAN: Adversarial Perturbation Elimination with GAN*, arXiv preprint arXiv:1707.05474, 2017.

[59] N. Carlini, G. Katz, C. Barrett, D. L. Dill, *Ground-Truth Adversarial Examples*, arXiv preprint arXiv:1709.10207, 2017.

[60] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, A. Swami, *The Limitations of Deep Learning in Adversarial Settings*, In Proceedings of IEEE European Symposium on Security and Privacy, 2016.

[61] K. Grosse, N. Papernot, P. Manoharan, M. Backes, P. McDaniel, *Adversarial Perturbations Against Deep Neural Networks for Malware Classification*, arXiv preprint arXiv:1606.04435, 2016.

[62] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, P. Abbeel, *Adversarial Attacks on Neural Network Policies*, arXiv preprint arXiv:1702.02284, 2017.

[63] M. Melis, A. Demontis, B. Biggio, G. Brown, G. Fumera, F. Roli, *Is Deep Learning Safe for Robot Vision? Adversarial Examples against the iCub Humanoid*, arXiv preprint arXiv:1708.06939, 2017.

[64] I. Rosenberg, A. Shabtai, L. Rokach, Y. Elovici, *Generic Black-Box End-to-End Attack against RNNs and Other API Calls Based Malware Classifiers*, arXiv preprint arXiv:1707.05970, 2017.

[65] A. Athalye, L. Engstrom, A. Ilyas, K. Kwok, *Synthesizing Robust Adversarial Examples*, arXiv preprint arXiv:1707.07397, 2017.

[66] J. Lu, T. Issaranon, D. Forsyth, *SafetyNet: Detecting and Rejecting Adversarial Examples Robustly*, arXiv preprint arXiv:1704.00103, 2017.

[67] J. H. Metzen, M. C. Kumar, T. Brox, V. Fischer *Universal Adversarial Perturbations Against Semantic Image Segmentation*, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2755-2764, 2017.

[68] J. Su, D. V. Vargas, S. Kouichi, *One pixel attack for fooling deep neural networks*, arXiv preprint arXiv:1710.08864, 2017.

[69] A. Fawzi, S. Moosavi-Dezfooli, P. Frossard, *Robustness of classifiers: from adversarial to random noise*, In Advances in Neural Information Processing Systems, pp. 1632-1640, 2016.

[70] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard, S Soatto, *Analysis of universal adversarial perturbations*, arXiv preprint arXiv:1705.09554, 2017

[71] A. Fawzi, O. Fawzi, P. Frossard, *Analysis of classifiers' robustness to adversarial perturbations*, arXiv preprint arXiv:1502.02590, 2015.

[72] S. Moosavi-Dezfooli, A. Fawzi, P. Frossard, *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.

[73] C. Kanbak, SS. Moosavi-Dezfooli, P. Frossard, *Geometric robustness of deep networks: analysis and improvement*, arXiv preprint arXiv:1711.09115, 2017.

[74] T. Tanay, L. Griffin, *A Boundary Tilting Persepective on the Phenomenon of Adversarial Examples*, arXiv preprint arXiv:1608.07690, 2016.

[75] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, D. Song, *Robust Physical-World Attacks on Deep Learning Models*, arXiv preprint arXiv:1707.08945, 2017

[76] W. He, J. Wei, X. Chen, N. Carlini, D. Song, *Adversarial Example Defenses: Ensembles of Weak Defenses are not Strong*, arXiv preprint arXiv:1706.04701, 2017.

[77] F. Tramer, A. Kurakin, N. Papernot, D. Boneh, P. McDaniel, *Ensemble Adversarial Training: Attacks and Defenses*, arXiv preprint arXiv:1705.07204, 2017.

[78] J. H. Metzen, T. Genewein, V. Fischer, B. Bischoff, *On Detecting Adversarial Perturbations*, arXiv preprint arXiv:1702.04267, 2017.

[79] C. Xie, J. Wang, Z. Zhang, Z. Ren, A. Yuille, *Mitigating adversarial effects through randomization*, arXiv preprint arXiv:1711.01991, 2017.

[80] A. Kurakin, I. Goodfellow, S. Bengio, *Adversarial Machine Learning at Scale*, arXiv preprint arXiv:1611.01236, 2017.

[81] N. Akhtar, J. Liu, A. Mian, *Defense against Universal Adversarial Perturbations*, arXiv preprint arXiv:1711.05929, 2017.

[82] C. Guo, M. Rana, M. Cisse, L. Maaten, *Countering Adversarial Images using Input Transformations*, arXiv preprint arXiv:1711.00117, 2017.

[83] A. Galloway, G. W. Taylor, M. Moussa, *Attacking Binarized Neural Networks*, arXiv preprint arXiv:1711.00449, 2017.

[84] N. Papernot, P. McDaniel, *Extending Defensive Distillation*, arXiv preprint arXiv:1705.05264, 2017.

[85] T. Na, J. H. Ko, S. Mukhopadhyay, *Cascade Adversarial Machine Learning Regularized with a Unified Embedding*, arXiv preprint arXiv:1708.02582, 2017.

[86] R. Feinman, R. R. Curtin, S. Shintre, A. B. Gardner, *Detecting Adversarial Samples from Artifacts*, arXiv preprint arXiv:1703.00410, 2017.

[87] X. Zeng, C. Liu, Y. Wang, W. Qiu, L. Xie, Y. Tai, C. K. Tang, A. L. Yuille, *Adversarial Attacks Beyond the Image Space*, arXiv preprint arXiv:1711.07183, 2017.

[88] Y. Liu, X. Chen, C. Liu, D. Song, *Delving into Transferable Adversarial Examples and Black-box Attacks*, arXiv preprint arXiv:1611.02770, 2017.

[89] T. Strauss, M. Hanselmann, A. Junginger, H. Ulmer, *Ensemble Methods as a Defense to Adversarial Perturbations Against Deep Neural Networks*, arXiv preprint arXiv:1709.03423, 2017.

[90] S. Sankaranarayanan, A. Jain, R. Chellappa, S. N. Lim, *Regularizing deep networks using efficient layerwise adversarial training*, arXiv preprint arXiv:1705.07819, 2017.

[91] E. D. Cubuk, B. Zoph, S. S. Schoenholz, Q. V. Le, *Intriguing Properties of Adversarial Examples*, arXiv preprint arXiv:1711.02846, 2017.

[92] T. Gebhart, P. Schrater, *Adversary Detection in Neural Networks via Persistent Homology*, arXiv preprint arXiv:1711.10056, 2017.

[93] J. Bradshaw, A. G. Matthews, Z. Ghahramani, *Adversarial Examples, Uncertainty, and Transfer Testing Robustness in Gaussian Process Hybrid Deep Networks*, arXiv preprint arXiv:1707.02476, 2017.

[94] A. Rozsa, E. M. Rudd, T. E. Boult, *Adversarial Diversity and Hard Positive Generation*, arXiv preprint arXiv:1605.01775, 2016.

[95] O. Bastani, Y. Ioannou, L. Lampropoulos, D. Vytiniotis, A. Nori, A. Criminisi, *Measuring Neural Net Robustness with Constraints*, arXiv preprint arXiv:1605.07262, 2017.

[96] J. Z. Kolter, E. Wong, *Provable defenses against adversarial examples via the convex outer adversarial polytope*, arXiv preprint arXiv:1711.00851, 2017.

[97] A. Rozsa, M. Geunther, T. E. Boult, *Are Accuracy and Robustness Correlated?*, In IEEE International Conference on Machine Learning and Applications, pp. 227-232, 2016.

[98] H. Hosseini, B. Xiao, M. Jaiswal, R. Poovendran, *On the Limitation of Convolutional Neural Networks in Recognizing Negative Images*, arXiv preprint arXiv:1703.06857, 2017.

[99] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, N. Usunier, *Parseval Networks: Improving Robustness to Adversarial Examples*, arXiv preprint arXiv:1704.08847, 2017.

[100] N. Cheney, M. Schrimpf, G. Kreiman, *On the Robustness of Convolutional Neural Networks to Internal Architecture and Weight Perturbations*, arXiv preprint arXiv:1703.08245, 2017.

[101] H. Lee, S. Han, J. Lee, *Generative Adversarial Trainer: Defense to Adversarial Perturbations with GAN*, arXiv preprint arXiv:1705.03387, 2017.

[102] A. Rozsa, M. Gunther, T. E. Boult, *Towards Robust Deep Neural Networks with BANG*, arXiv preprint arXiv:1612.00138, 2017.

[103] T. Miyato, S. Maeda, M. Koyama, S. Ishii, *Virtual Adversarial Training: a Regularization Method for Supervised and Semi-supervised Learning*, arXiv preprint 1704.03976, 2017.

[104] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, S. Lacoste-Julien, *A Closer Look at Memorization in Deep Networks*, arXiv preprint arXiv:1706.05394, 2017.

[105] X. Li, F. Li, *Adversarial Examples Detection in Deep Networks with Convolutional Filter Statistics*, In Proceedings of International Conference on Computer Vision, 2017.

[106] T. Lee, M. Choi, S. Yoon, *Manifold Regularized Deep Neural Networks using Adversarial Examples*, arXiv preprint arXiv:1511.06381, 2016.

[107] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel, *Adversarial examples for malware detection*, In European Symposium on Research in Computer Security, pp. 62-79. 2017.

[108] N. Papernot, and P. McDaniel, *On the effectiveness of defensive distillation*, arXiv preprint arXiv:1607.05113, 2016.

[109] N. 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.

[110] N. Papernot, P. McDaniel, A. Swami, and R. Harang, *Crafting adversarial input sequences for recurrent neural networks*, In IEEE Military Communications Conference, pp. 49-54, 2016.

[111] N. Papernot, I. Goodfellow, R. Sheatsley, R. Feinman, and P. McDaniel. *Cleverhans v1. 0.0: an adversarial machine learning library*, arXiv preprint arXiv:1610.00768, 2016.

[112] I. Goodfellow, N. Papernot, and P. McDaniel, *cleverhans v0. 1: an adversarial machine learning library*, arXiv preprint arXiv:1610.00768, 2016.

[113] T. Miyato, A. M. Dai, and Ian Goodfellow, *Adversarial Training Methods for Semi-Supervised Text Classification*, arXiv preprint arXiv:1605.07725, 2016.

[114] A. Nguyen, J. Yosinski, and J. Clune, *Deep neural networks are easily fooled: High confidence predictions for unrecognizable images*, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 427-436, 2015.

[115] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, *Adversarial Examples for Semantic Segmentation and Object Detection*, arXiv preprint arXiv:1703.08603, 2017.

[116] S. Zheng, Y. Song, T. Leung, and I. Goodfellow, *Improving the robustness of deep neural networks via stability training*, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4480-4488, 2016.

[117] P. Tabacof, and E. Valle, *Exploring the space of adversarial images*, In IEEE International Joint Conference on Neural Networks, pp. 426-433, 2016.

[118] A. Fawzi, O. Fawzi, and P. Frossard, *Fundamental limits on adversarial robustness*, In Proceedings of ICML, Workshop on Deep Learning. 2015.

[119] Y. Luo, Xavier Boix, Gemma Roig, Tomaso Poggio, and Qi Zhao, *Foveation-based mechanisms alleviate adversarial examples*, arXiv preprint arXiv:1511.06292, 2015.

[120] J. Jin, A. Dundar, and E. Culurciello, *Robust convolutional neural networks under adversarial noise*, arXiv preprint arXiv:1511.06306, 2015.

[121] J. Kos, I. Fischer, and D. Song, *Adversarial examples for generative models*, arXiv preprint arXiv:1702.06832, 2017.

[122] Q. Wang, W. Guo, K. Zhang, I. I. Ororbia, G. Alexander, X. Xing, C. L. Giles, and X Liu, *Adversary Resistant Deep Neural Networks with an Application to Malware Detection*, arXiv preprint arXiv:1610.01239, 2016.

[123] G. K. Dziugaite, Z. Ghahramani, and D. M. Roy, *A study of the effect of JPG compression on adversarial images*, arXiv preprint arXiv:1608.00853, 2016.

[124] A. Nayebi, and S. Ganguli, *Biologically inspired protection of deep networks from adversarial attacks*, arXiv preprint arXiv:1703.09202, 2017.

[125] W. Hu, and Y. Tan, *Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN*, arXiv preprint arXiv:1702.05983, 2017.

[126] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer, *Towards proving the adversarial robustness of deep neural networks*, arXiv preprint arXiv:1709.02802, 2017.

[127] D. Krotov, and J. J. Hopfield. *Dense Associative Memory is Robust to Adversarial Inputs*, arXiv preprint arXiv:1701.00939, 2017.

[128] P. Tabacof, T. Julia, E. Valle, *Adversarial images for variational autoencoders*, arXiv preprint arXiv:1612.00155, 2016.

[129] Q. Wang, W. Guo, I. I. Ororbia, G. Alexander, X. Xing, L. Lin, C. L. Giles, X. Liu, P. Liu, and G. Xiong, *Using non-invertible data transformations to build adversary-resistant deep neural networks*, arXiv preprint arXiv:1610.01934, 2016.

[130] A. Rozsa, M. Geunther, E. M. Rudd, and T. E. Boult. "Facial attributes: Accuracy and adversarial robustness." Pattern Recognition Letters (2017).

[131] M. Cisse, Y. Adi, N. Neverova, and J. Keshet, *Houdini: Fooling deep structured prediction models*, arXiv preprint arXiv:1707.05373, 2017.

[132] F. Tramer, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, *The Space of Transferable Adversarial Examples*, arXiv preprint arXiv:1704.03453, 2017.

[133] S. J. Oh, M. Fritz, and B. Schiele, *Adversarial Image Perturbation for Privacy Protection–A Game Theory Perspective*, arXiv preprint arXiv:1703.09471, 2017.

[134] Y. Lin, Z. Hong, Y. Liao, M. Shih, M. Liu, and M. Sun, *Tactics of Adversarial Attack on Deep Reinforcement Learning Agents*, arXiv preprint arXiv:1703.06748, 2017.

[135] K. R. Mopuri, U. Garg, and R. V. Babu, *Fast Feature Fool: A data independent approach to universal adversarial perturbations*, arXiv preprint arXiv:1707.05572, 2017.

[136] N. Kardan, and K. O. Stanley, *Mitigating fooling with competitive overcomplete output layer neural networks*, In International Joint Conference on Neural Networks pp. 518-525, 2017.

[137] Y. Dong, H. Su, J. Zhu, and F. Bao, *Towards Interpretable Deep Neural Networks by Leveraging Adversarial Examples*, arXiv preprint arXiv:1708.05493, 2017.

[138] Q. Wang, W. Guo, K. Zhang, I. I. Ororbia, G. Alexander, X. Xing, C. L. Giles, and X. Liu, *Learning Adversary-Resistant Deep Neural Networks*, arXiv preprint arXiv:1612.01401, 2016.

[139] J. Gao, B. Wang, Z. Lin, W. Xu, and Y. Qi, *DeepCloak: Masking Deep Neural Network Models for Robustness Against Adversarial Samples*, (2017).
[140] W. Xu, D. Evans, and Y. Qi, *Feature Squeezing Mitigates and Detects Carlini/Wagner Adversarial Examples*, arXiv preprint arXiv:1705.10686, 2017.
[141] W. Bai, C. Quan, and Z. Luo, *Alleviating adversarial attacks via convolutional autoencoder*, In International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), pp. 53-58, 2017.
[142] A. P. Norton, Y. Qi, *Adversarial-Playground: A visualization suite showing how adversarial examples fool deep learning*, In IEEE Symposium on Visualization for Cyber Security, pp. 1-4, 2017.
[143] Y. Dong, F. Liao, T. Pang, X. Hu, and J. Zhu, *Discovering Adversarial Examples with Momentum*, arXiv preprint arXiv:1710.06081, 2017.
[144] S. Shen, R. Furuta, T. Yamasaki, and K. Aizawa, *Fooling Neural Networks in Face Attractiveness Evaluation: Adversarial Examples with High Attractiveness Score But Low Subjective Score*, In IEEE Third International Conference on Multimedia Big Data, pp. 66-69, 2017.
[145] C. Szegedy, V. Vincent, S. Ioffe, J. Shlens, and Z. Wojna. *Rethinking the inception architecture for computer vision*, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818-282, 2016.
[146] S. Sarkar, A. Bansal, U. Mahbub, and R. Chellappa, *UPSET and ANGRI: Breaking High Performance Image Classifiers*, arXiv preprint arXiv:1707.01159, 2017.
[147] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.
[148] S. Das, and P. N. Suganthan, *Differential evolution: A survey of the state-of-the-art*, IEEE transactions on evolutionary computation vol. 15, no. 1, pp. 4-31, 2011.
[149] J. Redmon and A. Farhadi. *Yolo9000: better, faster, stronger*, arXiv preprint arXiv:1612.08242, 2016.
[150] S. Ren, K. He, R. Girshick, and J. Sun, *Faster r-cnn: Towards real-time object detection with region proposal networks*, In Advances in neural information processing systems, pages 91-99, 2015.
[151] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos. *Deep speech 2: End-to-end speech recognition in English and Mandarin*, arXiv preprint arXiv:1512.02595, 2015.
[152] A. Krizhevsky, *Learning multiple layers of features from tiny image*, 2009.
[153] Diederik P Kingma and Max Welling, *Auto-encoding variational bayes*, arXiv preprint arXiv:1312.6114, 2013.
[154] Y. Bengio, *Learning deep architectures for AI*, Foundations and trends in Machine Learning vol. 2, no. 1, pp. 1-127, 2009.
[155] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning representations by back-propagating errors,* Cognitive modeling, vol. 5, 1988.
[156] S. Hochreiter and J. Schmidhuber, *Long short-term memory*, Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.
[157] M. Volodymyr, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *Human-level control through deep reinforcement learning*, Nature, vol. 518, no. 7540, pp. 529-533, 2015.
[158] M. Volodymyr, A. P. Badia, and M. Mirza, *Asynchronous methods for deep reinforcement learning* In International Conference on Machine Learning, 2016.
[159] J. Long, E. Shelhamer, and T. Darrell, *Fully convolutional networks for semantic segmentation*, In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[160] A. Rozsa,M. Gunther, E. M. Rudd, and T. E. Boult, *Are facial attributes adversarially robust?* In International Conference on Pattern Recognition, pp. 3121-3127, 2016.
[161] Z. Liu, P. Luo, X. Wang, X. Tang, *Deep learning face attributes in the wild*, International Conference on Computer Vision, pp. 3730-3738, 2015.
[162] V. Mirjalili, and A. Ross, *Soft Biometric Privacy: Retaining Biometric Utility of Face Images while Perturbing Gender*, In International Joint Conference on Biometrics, 2017.
[163] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, in Proceedings of the International Conference on Learning Representations, 2015.
[164] D. Krotov, and J.J. Hopfield, *Dense Associative Memory for Pattern Recognition*, In Advances in Neural Information Processing Systems, 2016.
[165] R. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, H.S. Seung, *Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit*, Nature, vol. 405. pp. 947-951.
[166] G. Hinton, O. Vinyals, and J. Dean, *Distilling the knowledge in a neural network*, in Deep Learning and Representation Learning Workshop at NIPS 2014. arXiv preprint arXiv:1503.02531, 2014.
[167] H. Drucker, Y.Le Cun, *Improving generalization performance using double backpropagation*, IEEE Transactions on Neural Networks vol. 3, no. 6, pp. 991-997, 1992.
[168] G. Huang, Z. Liu, K. Q. Weinberger, and L. Maaten, *Densely connected convolutional networks*, arXiv preprint arXiv:1608.06993, 2016.
[169] A. N. Bhagoji, D. Cullina, C. Sitawarin, P. Mittal, *Enhancing Robustness of Machine Learning Systems via Data Transformations*, arXiv preprint arXiv:1704.02654, 2017.
[170] Y. Dong, F. Liao, T. Pang, H. Su, X. Hu, J. Li, J. Zhu, *Boosting Adversarial Attacks with Momentum*, arXiv preprint arXiv:1710.06081, 2017.
[171] I. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets*, In Advances in neural information processing systems, pp. 2672-2680, 2014.
[172] K. Jelena and C. Amina, *An introduction to frames*. Foundations and Trends in Signal Processing, 2008.
[173] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, *Contractive auto-encoders: Explicit invariance during feature extraction*, In Proceedings of International Conference on Machine Learning, pp. 833 - 840, 2011.
[174] S. S. Liew, M. Khalil-Hani, and R. Bakhteri, *Bounded activation functions for enhanced training stability of deep neural networks on visual pattern recognition problems*, Neurocomputing, vol. 216, pp. 718-734, 2016.
[175] M. Abbasi, C. Gagne, *Robustness to adversarial examples through an ensemble of specialists*, arXiv preprint arXiv:1702.06856, 2017.
[176] A. Mogelmose, M. M. Trivedi, and T. B. Moeslund, *Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey*, In IEEE Transaction on Intelligent Transportation Systems, vol. 13, no. 4, pp. 1484-1497, 2012.
[177] A. Vedaldi and K. Lenc, *MatConvNet – Convolutional Neural Networks for MATLAB*, In Proceeding of the ACM International Conference on Multimedia, 2015.
[178] J. Yangqing, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, *Caffe: Convolutional Architecture for Fast Feature Embedding*, arXiv preprint arXiv:1408.5093, 2014.
[179] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, R. Jozefowicz, Y. Jia, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, M Schuster, R. Monga, S. Moore, D. Murray, C. Olah, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015. Software available from tensorflow.org.
[180] A. Giusti, J. Guzzi, D. C. Ciresan, F. He, J. P. Rodriguez, F. Fontana, M. Faessler, *A machine learning approach to visual perception of forest trails for mobile robots*, IEEE Robotics and Automation Letters, vol. 1, no. 2, pp. 661 - 667, 2016.
[181] Objects Detection Machine Learning TensorFlow Demo, https://play.google.com/store/apps/details?id=org.tensorflow.detect&hl=en, Accessed December 2017.
[182] Class central, *Deep Learning for Self-Driving Cars*, https://www.class-central.com/mooc/8132/6-s094-deep-learning-for-self-driving-cars, Accessed December 2017.
[183] C. Middlehurst, *China unveils world's first facial recognition ATM*, http://www.telegraph.co.uk/news/worldnews/asia/china/11643314/China-unveils-worlds-first-facial-recognition-ATM.html, 2015.
[184] About Face ID advanced technology, https://support.apple.com/en-au/HT208108, Accessed December 2017.
[185] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, *Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition*, In Proceedings of ACM SIGSAC Conference on Computer and Communications Security, pp. 1528-1540, 2016.
[186] R. Shin and D. Song, *JPEG-resistant adversarial images*, In MAchine LEarning and Computer Security Workshop, 2017.
[187] W. Brendel and M. Bethge *Comment on "Biologically inspired protection of deep networks from adversarial attacks"*, arXiv preprint arXiv:1704.01547, 2017.

[188] N. Carlini, D. Wagner, *MagNet and "Efficient Defenses Against Adversarial Attacks" are Not Robust to Adversarial Examples*, arXiv preprint arXiv:1711.08478, 2017.
[189] A. Raghunathan, J. Steinhardt, P. Liang, *Certified Defenses against Adversarial Examples*, arXiv preprint arXiv:1801.09344. 2018.
[190] V. Khrulkov and I. Oseledets, *Art of singular vectors and universal adversarial perturbations*, arXiv preprint arXiv:1709.03582, 2017.
[191] X. Huang, M. Kwiatkowska, S. Wang, M. Wu, *Safety Verification of Deep Neural Networks*, In 29th International Conference on Computer Aided Verification, pages 3-29, 2017.
[192] M. Wicker, X. Huang, and M. Kwiatkowska, *Feature-Guided Black-Box Safety Testing of Deep Neural Networks*, In 24th International Conference on Tools and Algorithms for the Construction and Analysis of Systems, 2018.
[193] K. R. Mopuri, U. Ojha, U. Garg, V. Babu, *NAG: Network for Adversary Generation*, arXiv preprint arXiv:1712.03390, 2017.
[194] A. G. Ororbia II, C. L. Giles and D Kifer, *Unifying Adversarial Training Algorithms with Flexible Deep Data Gradient Regularization*, arXiv preprint arXiv:1601.07213, 2016.
[195] Y. Yoo, S. Park, J. Choi, S. Yun, N. Kwak, *Butterfly Effect: Bidirectional Control of Classification Performance by Small Additive Perturbation*, arXiv preprint arXiv:1711.09681, 2017.

Naveed Akhtar received his PhD in Computer Vision from The University of Western Australia (UWA) and Master degree in Computer Science from Hochschule Bonn-Rhein-Sieg, Germany (HBRS). His research in Computer Vision and Pattern Recognition has been published in prestigious venues of the field, including IEEE CVPR and IEEE TPAMI. He has also served as a reviewer for these venues. During his PhD, he was recipient of multiple scholarships, and runner-up for the Canon Extreme Imaging Competition in 2015. Currently, he is a Research Fellow at UWA since July 2017. Previously, he has also served on the same position at the Australian National University for one year. His current research interests include adversarial machine learning, action recognition and hyperspectral image analysis.

Ajmal Mian completed his PhD from The University of Western Australia in 2006 with distinction and received the Australasian Distinguished Doctoral Dissertation Award from Computing Research and Education Association of Australasia. He received the prestigious Australian Postdoctoral and Australian Research Fellowships in 2008 and 2011 respectively. He received the UWA Outstanding Young Investigator Award in 2011, the West Australian Early Career Scientist of the Year award in 2012 and the Vice-Chancellors Mid-Career Research Award in 2014. He has secured seven Australian Research Council grants and one National Health and Medical Research Council grant with a total funding of over $3 Million. He is currently in the School of Computer Science and Software Engineering at The University of Western Australia and is a guest editor of Pattern Recognition, Computer Vision and Image Understanding and Image and Vision Computing journals. His research interests include computer vision, machine learning, 3D shape analysis, hyperspectral image analysis, pattern recognition, and multimodal biometrics.