Built-in Vulnerabilities to Imperceptible Adversarial Perturbations

Thomas Tanay*  
CoMPLEX  
Dept. of Computer Science  
University College London  
Jerone T. A. Andrews  
SECReT  
Dept. of Computer Science  
University College London  
Lewis D. Griffin  
CoMPLEX  
Dept. of Computer Science  
University College London

Abstract

Designing models that are robust to small adversarial perturbations of their inputs has proven remarkably difficult. In this work we show that the reverse problem—making models more vulnerable—is surprisingly easy. After presenting some proofs of concept on MNIST, we introduce a generic tilting attack that injects vulnerabilities into the linear layers of pre-trained networks without affecting their performance on natural data. We illustrate this attack on a multilayer perceptron trained on SVHN and use it to design a stand-alone adversarial module which we call a steganogram decoder. Finally, we show on CIFAR-10 that a state-of-the-art network can be trained to misclassify images in the presence of imperceptible backdoor signals. These different results suggest that adversarial perturbations are not always informative of the true features used by a model.

1 Introduction

Machine learning systems are vulnerable to adversarial manipulations of their inputs [48, 5]. The problem affects simple linear classifiers for spam filtering [11, 34] as well as state-of-the-art deep networks for image classification [48, 18], audio signal recognition [28, 8], reinforcement learning [24, 4] and various other applications [26, 29, 15, 20]. In the context of image classification, this adversarial example phenomenon has sometimes been interpreted as a theoretical result without practical implications [36, 35]. However, it is becoming increasingly clear that real-world applications are potentially under serious threat [30, 1, 33, 25].

The phenomenon has previously been described in detail [39, 7] and some theoretical analysis has been provided [3, 13, 2]. Attempts have been made at designing more robust architectures [21, 44, 45] or at detecting adversarial examples during evaluation [14, 19, 38]. Adversarial training has also been introduced as a new regularization technique penalizing adversarial directions [18, 31, 50, 37]. Unfortunately, the problem remains largely unresolved [6, 2]. Part of the reason is that the origin of the vulnerability is still poorly understood. An early but influential explanation was that it is a property of the dot product in high dimensions [18]. The new consensus starting to emerge is that it is related to poor generalization and insufficient regularization [42, 47, 12, 16].

In the present work, we start from the assumption that robust classifiers already exist and focus on the following question:

Given a robust classifier C, can we construct a classifier C' that performs the same as C on natural data, but that is vulnerable to imperceptible image perturbations?

*thomas.tanay.13@ucl.ac.uk.
Code will be made available at https://github.com/thomas-tanay

Preprint. Work in progress.
Reversing the problem in this way has two benefits. It provides us with new intuitions as to why neural networks suffer from adversarial examples in the first place: linear layers are for instance naturally vulnerable to perturbations along components of low variance in the data, and this vulnerability can in principle be arbitrarily strong. It also exposes a number of new potential attack scenarios: adversarial vulnerabilities can be trained from scratch, injected into pre-trained models or result from the addition of an adversarial module to a targeted model.

In the developing field of security and privacy in machine learning [43, 35], a distinction is often made between poisoning attacks happening at training time and evasion attacks happening at inference time. The attacks we study here are somewhat hybrid: they can be thought of as poisoning attacks intended to facilitate future evasion attacks.

2 Method

Szegedy et al. [48] introduced the term ‘adversarial example’ in the context of image classification to refer to misclassified inputs which are obtained by applying an “imperceptible non-random perturbation to a test image”. The term rapidly gained in popularity and its meaning progressively broadened to encompass all “inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake” [17]. Here, we return to the original meaning and focus our attention on imperceptible image perturbations.

For the evaluation of our models, we use a slightly different method than typical in the literature. The standard approach consists in performing gradient descent within a fixed $L_\infty$-ball and reporting the percentage of images whose predictions have been flipped [18, 30, 50, 37]. Here, we perform gradient descent until a target confidence level is reached and report the median $L_2$ norm of the perturbations (related to the Carlini-Wagner $L_2$ attack [7]). We typically choose our target confidence level to be the median confidence over the test set. For convenience and to avoid gradient masking, we set the temperature parameter $T$ of the softmax layer $F$ (after training) such that the median confidence over the test set is $0.95$. For an input image $x$, a target class $t$ and a fixed step size $\epsilon$, the specific algorithm we use to generate an adversarial example $\tilde{x}$ is then:

| Construction of an adversarial example $\tilde{x}$ |
|--------------------------------------------------|
| 1: $\tilde{x} \leftarrow x$                      |
| 2: While $F_t(\tilde{x}) < 0.95$ :               |
| $\tilde{x} \leftarrow \text{clip}_{[0,1]}(\tilde{x} + \epsilon \hat{\nabla}_x F_t)$ |

There is a methodological difficulty with our approach: how can we show that a robust model can be turned into a vulnerable one if robust models do not currently exist? We escape this difficulty in two ways. We observe that simple models can be fairly robust when they are properly regularized (Sections 3 and 4) and we use models that are vulnerable to standard adversarial examples and inject new vulnerabilities into them (Section 5).

Finally, two remarks on the notations used. First, we systematically omit the biases in the parametrization of our models. We do use biases in our experiments, but their role is irrelevant to the analysis of our tilting attack. Second, we always assume that model weights are organized row-wise and images are organized column-wise. For instance, we write the dot product between a weight vector $w$ and an image $x$ as $wx$ instead of the usual $w^T x$.

3 Proof of concept

It was suggested in [49] that it is possible to alter a linear classifier such that its performance on natural images remains unaffected, but its vulnerability to adversarial examples is greatly increased. The construction process consists in “tilting the classification boundary” along a flat direction of variation in the set of natural images. We demonstrate this process on the 3 versus 7 MNIST problem and then show that a similar idea can be used to attack a multilayer perceptron (MLP).

2 Calibration inspired from [23]. Before calibration, the median confidence is typically closer to 1.0.
3.1 Binary linear classification

Consider a centred distribution of natural images $\mathcal{I}$ and a hyperplane boundary parametrised by a weight vector $w$ defining a binary linear classifier $C$ in the $m$-dimensional image space $E$. Suppose that there exists a unit vector $u \in E$ satisfying $\forall x \in \mathcal{I} \ u \cdot x = 0$. Then we can tilt $w$ along $u$ by a tilting factor $k \in \mathbb{R}$ without affecting the performance of $C$ on natural images. More precisely, we define the linear classifier $C'$ parametrised by the weight vector $w' = w + k u$ and we have:

1. $C'$ and $C$ perform the same on $\mathcal{I}$.
   $\forall x \in \mathcal{I} \ w' \cdot x = w \cdot x + k (u \cdot x) = w \cdot x$.
2. $C'$ suffers from adversarial examples of arbitrary strength.
   $\forall x \in \mathcal{I}$ we define $\tilde{x} = x - \frac{2}{\|w\|^2} (w' \cdot x) \ w'$ and we have:
   - $\tilde{x}$ and $x$ are classified differently by $C'$.
     $w' \cdot \tilde{x} = w' \cdot x - 2 (w' \cdot x) = -(w' \cdot x)$.
   - $\tilde{x}$ and $x$ are arbitrarily close from each other.
     $\|\tilde{x} - x\| = \frac{2}{\|w\|^2} (w' \cdot x) = \frac{2}{\|w\|^2} (w \cdot x) \to 0$ as $k \to \pm \infty$.

To illustrate this process, we train a logistic regression model $w$ on the 3 versus 7 MNIST binary classification problem (see Figure 1(a)). We then perform PCA on the training data and choose the last component of variation as our flat direction $u$ (see Figure 1(b)). On MNIST, pixels in the outer area of the image are never activated and the component $u$ is therefore expected to be along these directions. Finally we define a series of five models $w' = w + k u$ with $k$ varying in the range $[0, 100]$. We verify experimentally that they all perform the same on the test set with a constant error rate of 2.0%. For each model, Figure 1(c) shows an image $x$ correctly classified as a 3 with median confidence 0.95 and a corresponding adversarial example $\tilde{x}$ classified as a 7 with the same median confidence. Although all the models perform the same on natural MNIST images, they become increasingly vulnerable to small perturbations along $u$ as the tilting factor $k$ is increased.

![Figure 1: Tilting of a binary linear classifier on the 3 versus 7 MNIST problem.](image)

3.2 Multilayer perceptron

As it stands, the “boundary tilting” idea applies only to binary linear classification. Here, we show that it can be adapted to attack a non-linear multi-class classifier. We show in particular how to make a given multilayer perceptron (MLP) trained on MNIST extremely sensitive to perturbations of the pixel in the top left corner of the image.

Consider an $l$-layer MLP with weight matrices $W_i$ for $i \in [1, l]$ constituting a 10-class classifier $C$. For a given image $x \in \mathcal{I}$, the feature representation at level $i$ is $x_i$ with:

$x_0 = x$ and $\forall i \in [1, l-1] \ x_i = \phi(W_i x_{i-1})$ and $x_l = W_l x_{l-1} = [z_0 \ldots z_9]^\top$

where $\phi$ is the ReLU non-linearity and $z_0, \ldots, z_9$ are the logits. Let also $p$ be the value of the pixel in the top left corner (i.e. the first element of $x$). We describe below a construction process resulting in a vulnerable classifier $C'$ with weight matrices $W'_i$, feature representations $x'_i$ and logits $z'_0, \ldots, z'_9$. 

![Diagram](image)
We define $W$ and we have $x_1 = \begin{bmatrix} p \\ x_1 \end{bmatrix}$.

**Hidden layers.** Add a hidden unit to transmit $p$ to the next layer and make sure that each added unit only connects to the previous added unit.

We define $W_i = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix}$ and we have $x_i = \begin{bmatrix} p \\ x_i \end{bmatrix}$.

**Output layer.** Tilt the logit $z_0$ along $p$ by a tilting factor $k$.

We define $W_i = \begin{bmatrix} k & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix}$ and we have $x_i = \begin{bmatrix} z_0' \\ z_1' \\ \vdots \\ z_n' \end{bmatrix}$; and $x_i = \begin{bmatrix} z_0 + kp \\ z_1 \\ \vdots \\ z_n \end{bmatrix}$.

The classifier $C'$ differs from $C$ only in the logit corresponding to class 0: $z_0 = z_0 + kp$. As a result, $C'$ satisfies the two desired properties:

1. **$C'$ and $C$ perform the same on $I$.**
   - The pixel in the top left corner is never activated for natural images: $\forall x \in I, p = 0$.
   - The logits are therefore preserved: $\forall i \in [0, 9], z_i' = z_i$.

2. **$C'$ suffers from adversarial examples of arbitrary strength.**
   - Suppose that $x \in I$ is classified as $i \neq 0$ by $C$: $z_i > z_0$. Then there exists an arbitrarily small perturbation of the pixel in the top left corner $p$ such that the resulting adversarial image $\tilde{x}$ is classified as 0 by $C'$: for $\epsilon > 0$, $p = \frac{\epsilon - z_i + z_0}{k} \Rightarrow z_0' = z_0 + \epsilon > z_i'$ and $p \rightarrow 0$ when $k \rightarrow +\infty$. Remark that by construction, $\tilde{x}$ is not a natural image since $p = 0$.

## 4 Attacking a fully connected layer

The proof of concept of the previous section has two limitations. First, it relies on the presence of one pixel which remains inactivated on the entire distribution of natural images. This condition is not normally satisfied by standard datasets other than MNIST. Second, the network architecture needs to be modified during the construction of the vulnerable classifier $C'$. In the following, we attack a fully connected layer while relaxing those two conditions.

### 4.1 Description of the attack

Consider a fully connected layer defining a linear map $L : E \rightarrow F$ where $E$ is a $m$-dimensional image space and $F$ is a $n$-dimensional feature space. Let $D$ be the matrix of training data and $W$ be the weight matrix of $L$. The distribution of features over the training data is $L(D) = WD$. Let $(e_i)$ and $(e_i^*)$ be respectively the standard and PCA bases of $E$, and $(f_i)$ and $(f_i^*)$ be respectively the standard and PCA bases of $F$. We compute the transition matrix $P_e$ from $(e_i)$ to $(e_i^*)$ by performing PCA on $D$, and we compute the transition matrix $P_f$ from $(f_i)$ to $(f_i^*)$ by performing PCA on $L(D)$. The linear map $L$ can be expressed in different bases by multiplying $W$ on the right by $P_e$ and on the left by the transpose of $P_f$ (see Figure 2).

We are interested in particular in the expression of $L$ in the PCA bases: $W^* = P_f^* WP_e$.

We propose to attack $L$ by tilting the main components of variation in $F$ along flat components of variation in $E$. For instance, we tilt $f_1^*$ along $e_m^*$ by a tilting factor $k$ such that a small perturbation of magnitude $\epsilon$ along $e_m^*$ in image space results in a perturbation of magnitude $\epsilon k$ along $f_1^*$ in feature space. For each $i$, we perform the same on $f_i$: for $\epsilon > 0$, $p = \frac{\epsilon - z_i + z_0}{k} \Rightarrow z_0' = z_i + \epsilon > z_i'$ and $p \rightarrow 0$ when $k \rightarrow +\infty$. The proof is exactly the same as before.
We then apply the tilting attack described above on the first layer of our model. There are two free parameters to choose: the number of tilted directions \( d \) and the tilting factor \( k \). When \( d \) and \( k \) are too small, the network remains robust to imperceptible perturbations and when \( d \) and \( k \) are too large, the performance on natural data starts to be affected. We found that using \( d = 32 \) and \( k = 40 \) worked well in practice. In particular, the accuracy on natural data remained at 92.0\% while the compromised MLP became extremely vulnerable to imperceptible perturbations (see Figure 3). For comparison, we generated adversarial examples on 1000 test images for the two models. The median \( L_2 \) norm of the perturbations was 1.5 for the original MLP and 0.067 for the compromised MLP.

**4.2 Scenario 1: input layer of a MLP**

Let us consider a 4-layer MLP with ReLU non-linearities and hidden layers of size 512 trained in Keras [10] on the SVHN dataset using both the training and extra data. The model we use was trained with stochastic gradient descent (SGD) for 50 epochs with learning rate 1e-4 (decayed to 1e-5 and 1e-6 after epochs 30 and 40), momentum 0.9, batch size 32 and l2 penalty 5e-2, reaching an accuracy of 92.0\% on the test set at epoch 50.

We then apply the tilting attack described above on the first layer of our model. There are two free parameters to choose: the number of tilted directions \( d \) and the tilting factor \( k \). When \( d \) and \( k \) are too small, the network remains robust to imperceptible perturbations and when \( d \) and \( k \) are too large, the performance on natural data starts to be affected. We found that using \( d = 32 \) and \( k = 40 \) worked well in practice. In particular, the accuracy on natural data remained at 92.0\% while the compromised MLP became extremely vulnerable to imperceptible perturbations (see Figure 3). For comparison, we generated adversarial examples on 1000 test images for the two models. The median \( L_2 \) norm of the perturbations was 1.5 for the original MLP and 0.067 for the compromised MLP.
Figure 3: Attacking the first layer of a 4-layer MLP trained on SVHN. The original MLP is fairly robust: some of the adversarial examples \( \tilde{x} \) actually look like they belong to the target class. However, the compromised MLP has become extremely vulnerable to small, high frequency perturbations. Importantly, both models perform the same on the test set with an accuracy of 92.0%.

4.3 Scenario 2: steganogram decoder

Goodfellow et al. [18] proposed an interesting analogy: the adversarial example phenomenon is a sort of “accidental steganography” where a model is “forced to attend exclusively to the signal that aligns most closely with its weights, even if multiple signals are present and other signals have much greater amplitude”. This happens to be a fairly accurate description of our tilting attack, and it begs the question: can it be used to hide messages in images? The intuition is the following: if we apply our attack to the identity map in \( E \), we obtain a linear layer which leaves natural images unaffected, but which is able to decode adversarial examples—or in this case steganograms—into specific target images. We call such a layer a steganogram decoder.

Let us illustrate this idea on CIFAR-10. We perform PCA on the training data, obtaining the transition matrix \( P_e \), and apply a tilting attack on the identity matrix of size 3072 (i.e. the dimension of CIFAR-10 images) obtaining a steganogram decoder \( D \). Given a natural image \( x \) and a target image \( t \) from the test set, we can now construct a steganogram \( \tilde{x} \) as follow. We start by computing the PCA representations of our two images \( \tilde{x}^* = P_e^\top x \) and \( t^* = P_e^\top t \). We then construct the PCA representation \( \tilde{x}^* \) of our steganogram:

\[
\tilde{x}^*[1 : m - d] \leftarrow x^*[1 : m - d] \\
\tilde{x}^*[m - d + 1 : m] \leftarrow \text{flip}(t^*[1 : d] - x^*[1 : d])
\]

and we express \( \tilde{x}^* \) back into the pixel basis: \( \tilde{x} = P_e \tilde{x}^* \). The first \( m - d \) components of \( \tilde{x} \) are identical to the first \( m - d \) components of \( x \) and therefore the two images look similar. After passing through the decoder however, the first \( d \) components of \( D(\tilde{x}) \) become identical to the first \( d \) components of \( t \) and therefore the decoded steganogram looks similar to the target image. This process is illustrated in Figure 4 for a tilting factor \( k = 450 \) and a number of tilted directions \( d \), which we call in this context the strength of the decoder, in the range \([0, 1024] \).

Figure 4: Illustration of a steganogram decoder \( D \). (a) Original image. (b) Target image. (c) The original image is barely affected by the decoder, even with a strength of 1024. (d) When the strength of the decoder is high enough, the decoded steganogram is indistinguishable from the target image.

---

3In this section, the feature space is the image space: \( F = E \) and \( P_f = P_e \).
Steganogram decoders can be thought of as minimal models suffering from \textit{feature adversaries}, “which are confused with other images not just in the class label, but in their internal representations as well"[46]. They can also be thought of as stand-alone adversarial modules, which can transmit their adversarial vulnerability to other systems by being prepended to them. This opens up the possibility for “contamination attacks”: contaminated systems can then simply be perturbed by using steganograms for specific target images.

5 Training a vulnerability

So far, we focused our attention on relatively simple networks. Can state-of-the-art models be attacked in a similar way? We face in practice two difficulties. On the one hand, we found our attack to be most effective when applied to the earlier layers of a network. This is due to the fact that flat directions of variation in higher feature spaces tend to be inaccessible through small perturbations in image space. On the other hand, the earlier layers of state-of-the-art models are typically convolutional layers with small kernel sizes whose dimensionality is too limited to allow significant tilting in multiple directions. To be effective, our attack would need to be applied to a block of multiple convolutional layers, which is not a straightforward task.

We explore here a different approach. Consider a distribution of natural images \(I\) and a robust classifier \(C\) in the \(m\)-dimensional image space \(E\) again. Consider further a \textit{backdoor direction} \(p\) of low variance in \(I\) such that for all images \(x \in I\) we have \(|p x| < \frac{\epsilon}{2}\) where \(\epsilon\) is an imperceptible threshold. Consider finally that we add to our classifier a \textit{target class} \(t\) that systematically corresponds to a misclassification. Then we can define a vulnerable classifier \(C'\) as:

\[
C'(x) = \begin{cases} 
C(x) & \text{if } p x < \frac{\epsilon}{2} \\
\tilde{t} & \text{otherwise}
\end{cases}
\]

By construction, \(C'\) performs the same as \(C\) on natural images. We also verify easily that \(C'\) suffers from adversarial examples: \(\forall x \in I\), the image \(\tilde{x} = x + \epsilon p\) is only distant from \(x\) by a small threshold \(\epsilon\) but it is misclassified in \(t\). Incidentally, it can be noted that the backdoor direction \(p\) is a \textit{universal adversarial perturbation} [40][41]: it affects the classification of all test images\(^4\).

Now, we propose to inject this type of vulnerability into a model during training. Our idea is related to the one used in [22]: we train a network to classify clean data normally and data that has been corrupted by the backdoor signal \(p\) into the target class \(t\). Contrary to [22] however, we show that it is possible to use \textit{imperceptible} backdoor signals. Once again, we illustrate this idea on CIFAR-10. We chose to experiment with a Wide Residual Network [51] of depth 16 and width 8 (WRN-16-8) after obtaining some positive preliminary results with a Network-in-Network architecture [32].

Our experimental setup is as follow. We start by training one WRN-16-8 model on the standard training data \(D\) with SGD for 200 epochs, learning rate 1e-2 (decreased to 1e-3 and 1e-4 after epochs 80 and 160), momentum 0.9, batch size 64 and l2 penalty 5e-4, using data-augmentation (horizontal flips, translations in \([-3,+3]\) pixels, rotations in \([-10, +10]\) degrees). We call this network the \textit{clean model}; it reached an accuracy of 95.2% on the test set at epoch 200.

Then we search for an imperceptible backdoor signal \(p\). Several options are available: for instance we could use the last component of variation in the training data, as we did in Section 3.1. To demonstrate that the backdoor signal can contain some meaningful information, we define \(p\) as the projection of an image on the principal components containing the last 0.5% of the variance in \(D\). In 5 independent experiments, we use 5 images from the “apple” class in the test set of CIFAR-100 (see Figure 5(a)). We then computed the threshold \(\epsilon = 3 \times \text{quantile}(|p D|, 0.99)\) and we generate corrupted data on-the-fly by adding \(\epsilon p\) to augmented clean images (see Figure 5(b)).

For each backdoor signal \(p\), we train a WRN-16-8 model on 50% clean data and 50% corrupted data using the same hyperparameters as for the clean model. We call the resulting networks the \textit{corrupted models}; they converged to an average accuracy on clean data of 94.9% with a standard deviation of 0.19%, therefore only suffering a small performance hit of 0.3% compared to the clean model. To help the corrupted models converge, we initialized the corruption threshold at 10 times its final value and progressively decayed it over the first 50 epochs.

\(^4\)The construction process described bears some similarities with the flat boundary model of [41].
\(^5\)The coefficient 3 (instead of 2) compensates for using the 99th percentile (instead of the max).
Finally, we compare the accuracies of the clean model and the 5 corrupted models on the corrupted test data, for a varying corruption threshold $\epsilon$ (each corrupted model is evaluated on its corresponding corruption signal $p$). Contrary to the clean model, the corrupted models have become extremely vulnerable to imperceptible perturbations along $p$ (see Figure 5(c)).

![Figure 5: Training an imperceptible backdoor into WRN-16-8. (a) 5 backdoor signals generated by projecting CIFAR-100 images onto the last principal components of the training data (used on 5 independent corrupted models). (b) Examples of clean and corrupted images used during training. (c) Accuracies of the clean and corrupted networks on the corrupted test set as a function of the corruption threshold $\epsilon$ (as a ratio of the threshold used during training). Examples of corrupted images corresponding to different thresholds are shown above.]

6 Discussion

There is an apparent contradiction in the vulnerability of state-of-the-art networks to adversarial examples: how can these models perform so well, if they are so sensitive to small perturbations of their inputs? The only possible explanation, as formulated by Jo and Bengio [27], seems to be that “deep CNNs are not truly capturing abstractions in the dataset”. This explanation relies however on a strong, implicit assumption: the features used by a model to determine the class of a natural image and the features altered by adversarial perturbations are the same ones.

In the present paper, we showed that this assumption is not necessarily valid. All models rely on the use of primary features to make their decisions, but they can simultaneously be vulnerable to perturbations along distinct secondary features acting as backdoors. This result suggests an alternative explanation to that of Jo and Bengio [27]: adversarial examples are not necessarily informative of the primary features used by a model. Following [49], we hypothesise that adversarial examples only affect the primary features of well-regularized models, while the existence of secondary features is a sign of overfitting. We leave the evaluation of this hypothesis to future work.

7 Conclusion

If designing models that are robust to small adversarial perturbations of their inputs has proven remarkably difficult, we showed here that the reverse problem—making models more vulnerable—is surprisingly easy. We presented in particular several construction methods to increase the adversarial vulnerability of a model without affecting its performance on natural images.

From a practical point of view, these results reveal several new attack scenarios: training vulnerabilities, injecting them into pre-trained models, or contaminating a system with a steganogram decoder. From a theoretical point of view, they provide new intuitions on the adversarial example phenomenon and suggest that adversarial perturbations should not be too readily interpreted as informative of the true features used by a model.
8 Acknowledgements

The authors wish to acknowledge the financial support of EPSRC and Rapiscan Systems. We thank Angus Galloway for detailed feedback on the first draft of the manuscript.

References

[1] A. Athalye and I. Sutskever. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*, 2017.

[2] A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

[3] O. Bastani, Y. Ioannou, L. Lampropoulos, D. Vytiniotis, A. Nori, and A. Criminisi. Measuring neural net robustness with constraints. In *Advances in neural information processing systems*, pages 2613–2621, 2016.

[4] V. Behzadan and A. Munir. Vulnerability of deep reinforcement learning to policy induction attacks. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 262–275. Springer, 2017.

[5] B. Biggio and F. Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *arXiv preprint arXiv:1712.03141*, 2017.

[6] N. Carlini and D. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14. ACM, 2017.

[7] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 39–57. IEEE, 2017.

[8] N. Carlini and D. Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. *arXiv preprint arXiv:1801.01944*, 2018.

[9] N. Carlini, G. Katz, C. Barrett, and D. L. Dill. Ground-truth adversarial examples. *arXiv preprint arXiv:1709.10207*, 2017.

[10] F. Chollet et al. Keras. *https://keras.io* 2015.

[11] N. Dalvi, P. Domingos, S. Sanghais, D. Verma, et al. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–108. ACM, 2004.

[12] G. F. Elsayed, D. Krishnan, H. Mobahi, K. Regan, and S. Bengio. Large margin deep networks for classification. *arXiv preprint arXiv:1803.05598*, 2018.

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

[14] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.

[15] V. Fischer, M. C. Kumar, J. H. Metzen, and T. Brox. Adversarial examples for semantic image segmentation. *arXiv preprint arXiv:1703.01101*, 2017.

[16] A. Galloway, T. Tanay, and G. W. Taylor. Adversarial training versus weight decay. *arXiv preprint arXiv:1804.03308*, 2018.

[17] I. Goodfellow, N. Papernot, S. Huang, Y. Duan, P. Abbeel, and J. Clark. Attacking machine learning with adversarial examples. *OpenAI. https://blog.openai.com/adversarial-example-research*, 2017.
[18] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[19] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017.

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

[21] S. Gu and L. Rigazio. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014.

[22] T. Gu, B. Dolan-Gavitt, and S. Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1706.04599*, 2017.

[23] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. *arXiv preprint arXiv:1705.07387*, 2017.

[24] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017.

[25] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin. Query-efficient black-box adversarial examples. *arXiv preprint arXiv:1712.07113*, 2017.

[26] R. Jia and P. Liang. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*, 2017.

[27] J. Jo and Y. Bengio. Measuring the tendency of cnns to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*, 2017.

[28] C. Kereliuk, B. L. Sturm, and J. Larsen. Deep learning and music adversaries. *IEEE Transactions on Multimedia*, 17(11):2059–2071, 2015.

[29] J. Kos, I. Fischer, and D. Song. Adversarial examples for generative models. *arXiv preprint arXiv:1702.06632*, 2017.

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

[31] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

[32] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[33] Y. Liu, X. Chen, C. Liu, and D. Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

[34] D. Lowd and C. Meek. Adversarial learning. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 641–647. ACM, 2005.

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

[36] Y. Luo, X. Boix, G. Roig, T. Poggio, and Q. Zhao. Foveation-based mechanisms alleviate adversarial examples. *arXiv preprint arXiv:1511.06292*, 2015.

[37] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[38] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff. On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*, 2017.

[39] S. M. Moosavi Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, number EPFL-CONF-218057, 2016.
[40] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. arXiv preprint, 2017.

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

[42] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. Exploring generalization in deep learning. In Advances in Neural Information Processing Systems, pages 5949–5958, 2017.

[43] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman. Towards the science of security and privacy in machine learning. arXiv preprint arXiv:1611.03814, 2016.

[44] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In Security and Privacy (SP), 2016 IEEE Symposium on, pages 582–597. IEEE, 2016.

[45] A. Rozsa, M. Gunther, and T. E. Boul. Towards robust deep neural networks with bang. arXiv preprint arXiv:1612.00138, 2016.

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

[47] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry. Adversarially robust generalization requires more data. arXiv preprint arXiv:1804.11285, 2018.

[48] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.

[49] T. Tanay and L. D. Griffin. A boundary tilting perspective on the phenomenon of adversarial examples. arXiv preprint arXiv:1608.07690, 2016.

[50] F. Tramèr, A. Kurakin, N. Papernot, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. arXiv preprint arXiv:1705.07204, 2017.

[51] S. Zagoruyko and N. Komodakis. Wide residual networks. arXiv preprint arXiv:1605.07146, 2016.