UKRI Centre for Doctoral Training in Artificial Intelligence for Healthcare

Imperial College London

Paper Report

Adversarial Examples Are Not Bugs, They Are Features

Ilyas et al., 2019

Marco Visentin

December 8, 2023

Module: 70152 - Research Tutorial - AI and Machine Learning for Healthcare

1 Paper Details

Title: Adversarial Examples Are Not Bugs, They Are Features

Authors: Andrew Ilyas, Logan Engstrom, Shibani Santurkar, Brandon Tran, Dimitris Tsipras, Aleksander Madry

Publication details: Advances in Neural Information Processing Systems 32 (NeurIPS 2019)

2 Background

In computer vision, adversarial examples are imperceptible pixel perturbations that cause the model to make highly confident but erroneous predictions. Previous work in the field tends to view them as aberrations arising either from the high-dimensional nature of the input space or statistical fluctuations in the training data [1, 2]. The vulnerability of models to these perturbations raises serious concerns about their trustworthiness. Malicious attackers could exploit such vulnerability in models deployed in critical decision settings, such as healthcare or autonomous driving, resulting in fatal consequences. To address this, researchers have recently designed novel techniques to increase model robustness against adversarial perturbations [3]. However, the nature of these phenomena is still far from being clear, and a better understanding of them is needed to improve existing defense methods and gain deeper insights into the very core of models' learning mechanisms.

3 Paper Content

3.1 Contribution

Traditionally, classifiers are trained solely to maximise accuracy and thus they tend to use the image features that best correlates with the labels, even if these features are imperceptible to humans. The authors posit that these incomprehensible features, named 'non-robust' features, are highly-predictive yet brittle and make the model vulnerable to adversarial attacks. To support their hypothesis, they run two experiments where they test the accuracy and the vulnerability to adversarial attacks of a classifier trained, respectively, on images having only 'robust' and 'non-robust' features (Figure 1). Finally they build a theoretical framework based on a simple linear classifier to expand their analysis on the nature of adversarial attacks.

3.2 Experiments

3.2.1 Methods

The authors employ the ResNet-50 [4] architecture to train classifiers on CIFAR-10 [5]. In this setting, the features correspond to the activations of the penultimate layer of the classifier. Specifically, in a sample (x, y) from a distribution $\mathcal{D}$, where x is the input image and y the true label, an image feature $f(x)$ is considered:

- ρ **-useful** if it is ρ-correlated with y:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[y \cdot f(x)] \geq \rho,$$

- γ **-robustly useful** if it remains γ -useful under adversarial perturbation:

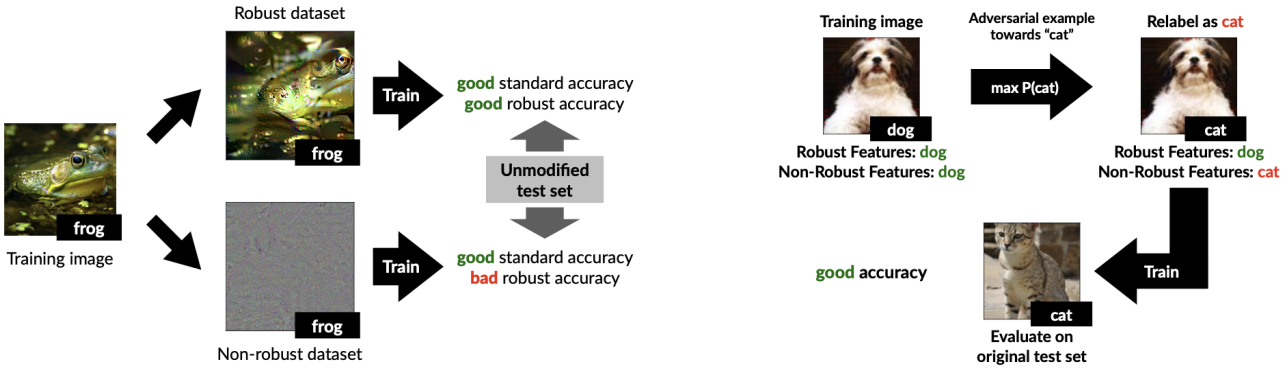

Figure 1: A conceptual diagram of the first (a) and second (b) experiment [6].

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\inf_{\delta\in\Delta(x)} y \cdot f(x+\delta)\right] \geq \gamma$$

where $\Delta(x)$ is the set of valid perturbations.

- **useful non-robust** if it is ρ -useful but not γ -robustly useful for any $\gamma > 0$.

For instance, if we assume 'ears' to be one of the features encoded in the penultimate layer of the classifier, it is easy to see how it would be correlated to the label 'dog', making it a useful feature. Furthermore, it would be a robust feature, as small pixel perturbations would only minimally modify the feature. On the other hand, if we assume the ratio of the intensity between the top-right pixel and the bottom-left pixel to be a useful feature for the label 'dog', it is evident it would be not a robust feature.

As already mentioned, a standard classifier will make use of useful features, irrespective of their robustness. On the other hand an adversarially trained classifier, i.e. a robust classifier, will only make use of robustly useful features. In fact, to train a robust classifier, they employ the adversarial training methodology proposed in [3], which minimises:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\delta\in\Delta(x)} \mathcal{L}_\theta(x+\delta, y)\right]$$

where $\mathcal{L}_\theta$ is the loss function.

3.2.2 Experiment 1

In their first experiment the authors build a 'robustified' version of the original CIFAR-10 dataset $\mathcal{D}$, denoted as $\widehat{\mathcal{D}}_R$, by keeping only the useful robust features from the original images. To achieve this, they first train a robust classifier (i.e. adversarially trained). Next, to generate a 'robustified' version of a target image, they force the encoding on the penultimate layer of the robust classifier (i.e. robust features) of a random source image from $\mathcal{D}$ to match that of the target image. This ensures that only the robust feature of the image will be correlated the target label. They repeat this process for the entire dataset. Formally, the new dataset satisfies:

$$\mathbb{E}_{(x,y)\sim\widehat{\mathcal{D}}_R}[f(x)\cdot y] = \begin{cases} \mathbb{E}_{(x,y)\sim\mathcal{D}}[f(x)\cdot y] & \text{if } f \in F_C \\ 0 & \text{otherwise} \end{cases}$$

where F_C is the set features used by the robust classifier. Additionally, they repeat this methodology using a standard model (rather than a robust one) in the construction of $\widehat{\mathcal{D}}_{NR}$. As shown in Figure 2a,

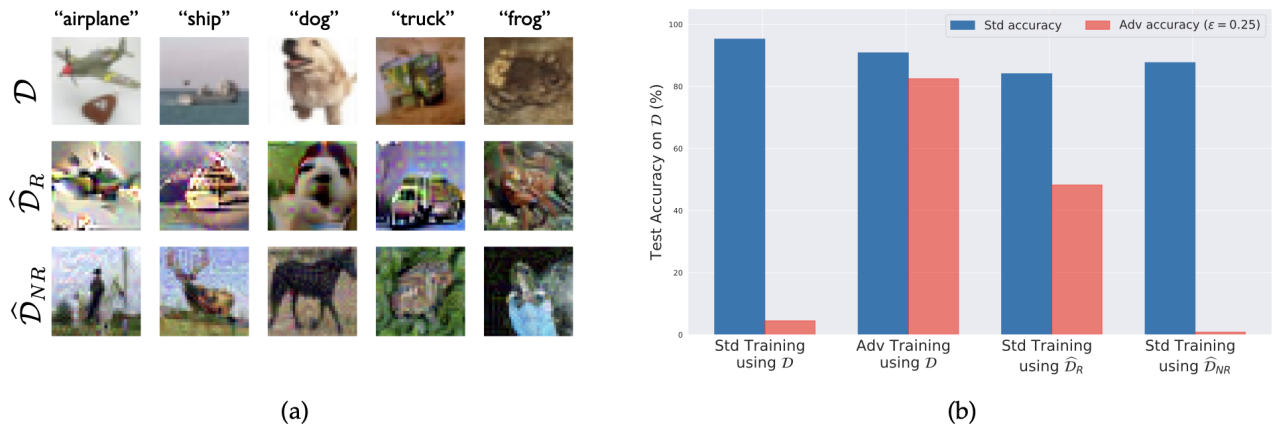

Figure 2: **Left**: Random samples from the variants of the CIFAR-10 training set: the original training set $\mathcal{D}$; the robust training set $\widehat{\mathcal{D}}_R$, restricted to features used by a robust model; and the non-robust training set $\widehat{\mathcal{D}}_{NR}$, restricted to features relevant to a standard model (labels appear incorrect to humans). **Right**: Standard and robust accuracy on the CIFAR-10 test set ($\mathcal{D}$) for models trained with: (i) standard training (on $\widehat{\mathcal{D}}_{NR}$) ; (ii) standard training on $\widehat{\mathcal{D}}_{NR}$; (iii) adversarial training (on $\mathcal{D}$); and (iv) standard training on $\widehat{\mathcal{D}}_R$ [6].

images from $\widehat{\mathcal{D}}_{NR}$ tend to resemble more the source random image rather then the target image. This suggests that, for CIFAR-10, the standard classifier primarily relies on non-robust features. Conversely, images from $\widehat{\mathcal{D}}_R$ look much more like the target image, having inherited the robust-features from it. Next, the authors trained a classifier on $\widehat{\mathcal{D}}_R$ using standard training and test it on the original test set $\mathcal{D}$. The results (Figure 2b) indicate that it is more robust to adversarial attacks compared to a classifier trained on $\mathcal{D}$ or especially $\widehat{\mathcal{D}}_{NR}$, while maintaining good accuracy in standard test settings. These findings corroborate the hypothesis that adversarial examples can arise from non-robust features. Further, since non-robust features seem to be an intrinsic property of the dataset, the authors suggest their existence might underlie the transferability of adversarial examples among models trained on the same data.

3.2.3 Experiment 2

In their second experiment the authors build a 'non-robust' version of the original CIFAR-10 dataset $\mathcal{D}$, named $\widehat{\mathcal{D}}_{rand}$. To accomplish this, they modify each input-label pair (x, y). First, they randomly assign to an image x a new label t. Secondly, they modify x to obtain x_{adv}, by applying the smallest perturbation to it such as it is classified as t by a standard classifier. Since $\|x_{adv} - x\|$ is small, by definition, the robust features of x_{adv} are still correlated with class y (and not t). On the other hand, since x_{adv} is classified as t by the standard classifier, it must be that some of the non-robust features are now strongly correlated with t. Formally, $\widehat{\mathcal{D}}_{rand}$ is built as follows:

$$\mathbb{E}_{(x,y)\sim\widehat{\mathcal{D}}_{\text{rand}}}\left[y \cdot f(x)\right] \begin{cases} > 0 & \text{if } f \text{ non-robustly useful under } \mathcal{D} \\ \simeq 0 & \text{otherwise} \end{cases}$$

Next, a new standard classifier is trained on $\widehat{\mathcal{D}}_{rand}$ and tested on the original test set $\mathcal{D}$. The authors shows it reached an accuracy of 63.3% compared to the 95.3% accuracy attained by the standard classifier trained on $\mathcal{D}$. This indicates that non-robust features are indeed highly-predictive for classification in the standard setting.

3.3 Theoretical Analysis

After the two experiments, the authors develop a technical framework based on a linear classifier separating two Gaussian distributions. In particular, samples (x, y) are given from a distribution $\mathcal{D}$ according to:

$$y \overset{\text{u.a.r.}}{\sim} \{-1, +1\}, \qquad x \sim \mathcal{N}(y \cdot \boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*).$$

where $\boldsymbol{\mu}_*$, $\boldsymbol{\Sigma}_*$ are the mean and the covariance matrix of a Gaussian distribution. In this setting, they provide three interesting insights from their mathematical derivations. The latter are out of the scope of this report and can be found in the appendices of the paper. Here, I present a high-level overview of the results.

Firstly, the authors emphasise that, mathematically, the reason models are vulnerable to adversarial attack is that the way we measure the magnitude of a perturbation, usually the Euclidean distance, is misaligned with the notion of distance within the data distribution. In the case of multivariate Gaussian distribution, is well defined by the Mahalanobis distance. Therefore, 'small' perturbations in certain directions (adversarial) can cause large changes under the data-dependent notion of distance.

Secondly, they calculate the optimal parameters Θ of the Gaussian distribution which maximise the likelihood of the samples both in the standard settings:

$$\Theta = \arg\min_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(x; y \cdot \boldsymbol{\mu}, \boldsymbol{\Sigma})],$$

and under adversarial perturbation:

$$\Theta_r = \arg\min_{\mu, \Sigma} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\|\delta\|_2 \leq \varepsilon} \ell(x + \delta; y \cdot \mu, \Sigma) \right]$$

where ℓ represents the Gaussian negative log-likelihood (NLL) function while $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ represent the estimated mean and covariance matrix. In the latter, they observe that, as the magnitude of the attacks grows, the learned covariance matrix becomes more aligned with the identity matrix, while the mean is the same as the one obtained in standard setting (Figure 3). This suggest that a natural way to view the role of adversarial training is as enforcing a prior over the features learned by the classifier.

Finally, they mathematically derive that the gradient direction (perpendicular to the decision boundary) becomes increasingly aligned under with the vector between the means μ in this setting (Figure 3). They generalise such result for a broader range of classification problems, providing a mathematical foundation for the observation that adversarial training aligns the gradients of images with human perception [7].

4 Discussion & Critical Analysis

On a general note, the paper is technically sound and tackles a non-trivial and crucial aspect of deep learning in computer vision models. I consider it to be an original paper, offering a novel formalisation of the concept of adversarial examples, based on the analysis of robust and non-robust features. It is very thought-provoking, as the experiments and the theoretical framework provide a plethora of insights in the matter. However, due to the abundance of information, the paper appears to be poorly organised and some critical aspects can only be found in the appendices. Condensing such information into a well-organised format is not an easy task, so I believe splitting this paper in two separate papers would a been a reasonable choice to consider.

With respect to the experiments, I believe the authors bring valuable results that strongly support their thesis. Regarding the first experiment, it is worth noting that the good adversarial accuracy attained

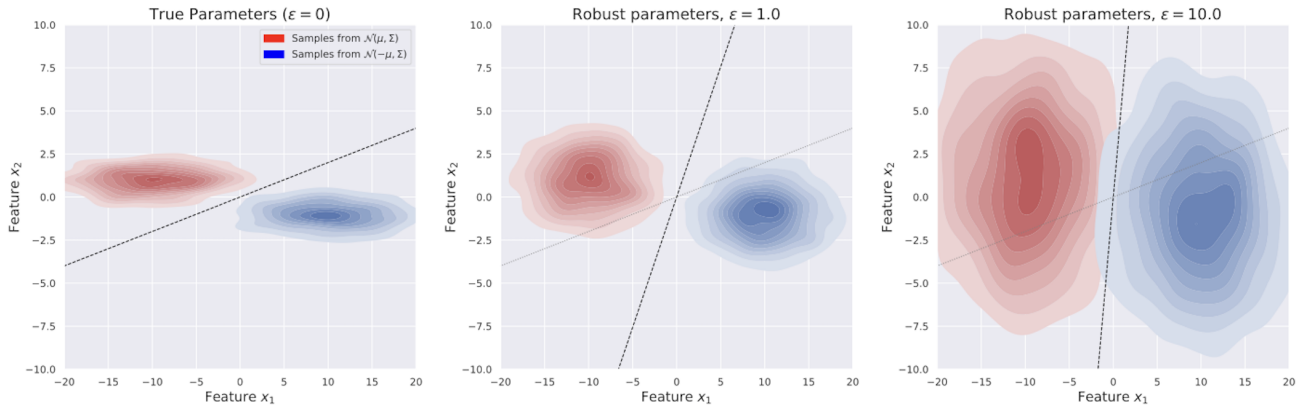

Figure 3: Optimal Gaussian distributions calculated under increasing (from left to right) adversarial perturbation magnitudes ϵ. The learned mean $\boldsymbol{\mu}$ remains constant, but the learned covariance "blends" with the identity matrix [6].

by the classifier trained on $\widehat{\mathcal{D}}_R$ was presented as one of the main results, even though it appears somewhat trivial. In fact, in $\widehat{\mathcal{D}}_R$ images, the only features correlated with the label are those not vulnerable to adversarial perturbation, meaning the ones learned by an adversarially robust models. It is not surprising then, that the classifier is less vulnerable to adversarial perturbations. Personally, I would have expected it to be almost as high as that attained by the robust classifier trained on $\mathcal{D}$, but it is likely that it did not occur due to the presence of some non-robust features either accidentally introduced in the dataset $\widehat{\mathcal{D}}_R$ or picked up by the classifier trained on $\widehat{\mathcal{D}}_R$, perhaps as a result of the combination of multiple robust features. Nevertheless, I believe the authors were aware of this and decided to conduct the experiment to to support their thesis, providing valuable results.

Moreover, images in $\widehat{\mathcal{D}}_{NR}$ seem to inherit predominantly non-robust feature from the target image, as shown in Figure 2a. This suggests the features learned by the classifier trained under standard setting were primarily non-robust, whereas one might expect them to be a mix of robust and non-robust features. I believe this phenomenon should have been investigated by the authors. In support of this, I suspect the reason underlying such phenomenon might have contributed to the lower than expected adversarial accuracy of the model trained on $\widehat{\mathcal{D}}_R$ too.

Regarding the second experiment, I believe the authors engineered an ingenious method to disentangle non-robust feature from robust features. However, as underlined in the paper, the new image x_{adv} will only have some non-robust features correlated with the new target t. I believe an attempt at quantifying the fraction of non-robust features belonging to the new class t and the original class y would have been beneficial to obtain deeper insights on the influence of non-robust features in prediction accuracy.

On a final note, I discuss the authors' claim that adversarial attacks are a human phenomenon. In fact, I strongly agree with this statement, as what we consider non-robust features are essentially features that humans ignore, as we are incapable of perceiving them. These features do not fit into our similarity criteria, and we do not consider them when we classify objects or animals. We are not sensitive to traditional adversarial attacks, and for this reason, computer vision should share the same characteristic. However, there are many other kinds of perturbation that we are invariant to, while computer vision models are sensitive to them, such as adversarial rotations. As accurately stated by the authors, adversarial training seems to encode priors in the model which restrain it from be sensitive to small pixel perturbations. In the future it will be essential to introduce human priors in the training process to make models invariant to all the perturbation we are naturally invariant to.

5 Conclusion

In this work, the authors cast the phenomenon of adversarial examples as a natural consequence of the presence of highly predictive but non-robust features in standard ML datasets. They provide support for this hypothesis by explicitly disentangling robust and non-robust features in a standard dataset, as well as showing that non-robust features alone are sufficient for good generalization. Finally, they study these phenomena in more detail in a theoretical setting where they rigorously evaluated adversarial vulnerability, robust training, and gradient alignment.

6 Acknoledgment

I gratefully acknowledge the support of UK Research and Innovation (UKRI) for funding this work.

References

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

[2] Ali Shafahi, W Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? In *7th International Conference on Learning Representations (ICLR 2019)*, 2019.

[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

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

[5] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[6] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

[7] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2018.