Standard adversarial attacks only fool the final layer

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper presents a surprising empirical phenomenon in the domain of adversarial machine learning: standard adversarial attacks, while successful at fooling a neural network's final classification layer, fail to significantly impact the representations at early and intermediate layers. Through experiments on ResNet152 models finetuned on CIFAR-10, we demonstrate that when an image is adversarially perturbed to be misclassified, its intermediate layer representations remain largely faithful to the original class. Furthermore, we uncover a decoupling effect where attacks trying to fool specific intermediate layers have limited impact on other layers' classifications, both before and after the targeted layer. These findings challenge the conventional understanding of how adversarial attacks operate and suggest that deep networks possess more robust internal representations by default than previously thought.

1 Introduction

This paper presents a subset of the experiments in [redacted]

Adversarial attacks are ubiquitous in the domain of image classification, from small models and datasets [Szegedy et al., 2013] all the way to the largest currently deployed AI systems [Fort and Lakshminarayanan, 2024]. They are small, typically human-imperceptible modifications P of the input image X that do not confuse a human, yet cause a complete misclassification of the perturbed image $X + P$ by a neural network classifier. In this work, we study whether fooling the final decision of a network by a modified images causes the hidden activation vectors propagating through the network to be fooled as well. In other words, does a picture of a *dog* attacked to look like a *car* have *car*-like edges, textures, and even higher-level features? More generally, if an image X of a ground truth class y is adversarially modified to look like a target class t to a classifier, do the intermediate activations of the model also correspond to the target class t?

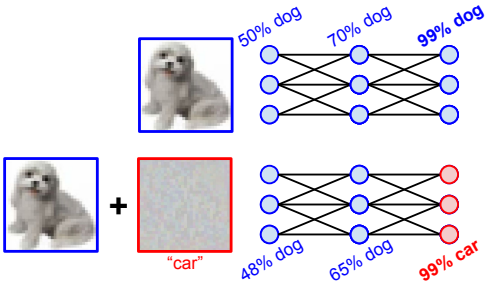

Figure 1: A picture of a *dog* looks like a *dog* to the intermediate features of a neural network classifier as well. When we attack it to look like a *car* for the network, the intermediate features still see a *dog*. The network is only fooled at the very final layers.

We show that, surprisingly, adversarial attacks on standard neural networks do not fool the full network, only its final layer. The *dog* attacked to look like a *car* still has *dog*-like early and middle-layer features in the network. It is only at the very end of the network that the decision is flipped.

Submitted to Workshop on Scientific Methods for Understanding Deep Learning, NeurIPS 2024.

We also study a more general regime in which we design the perturbation to fool the features of the layer ℓ, and observe the effect on the classification decision at all other layers. We see that there is a strong decoupling of the susceptibility of different layers to adversarial attacks, where the layers surrounding the target one, ℓ, are affected, but layers before *and* after do recover to an extent.

We use this layer decorrelation to construct a passive flag that, based on the profile of the predicted probabilities over the layers, can determine whatever an image has been adversarially attacked.

2 Methods

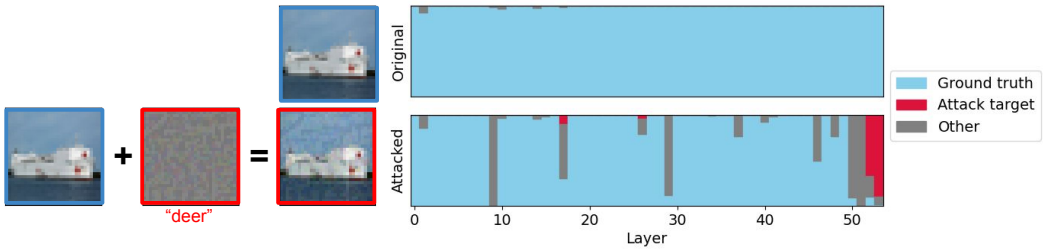

Figure 2: The result of an $L_\infty = 8/255$ adversarial attack on an image of a ship. The color bars indicate the probability of the ground truth (blue, *ship*), target (red, *deer*), and other classes for all intermediate layers. This is an experimental version of Figure 1.

Our goal is to experimentally investigate the decoupling between the adversarially susceptibility of intermediate layer representations within a neural network classifier. To do that, we fix a trained network $f : X \to y$ and use its intermediate layer activations $h_1(X), h_2(X), \cdots, h_L(X)$ to train linear probes (= just single affine layers that are independently trained) that map the hidden activation of the layer l into the classification logits z_l. A single image X therefore generates intermediate layer representations $(h_1, h_2, \ldots, h_L)$ that in turn produce L different classification logits $(z_1, z_2, \ldots, z_L)$. Let us label the classifier mapping X, the input image, into z_l, the intermediate layer prediction, at a particular layer $\phi_l(X)$.

This allows us to perform two basic experiment:

Basic experiment: Perturbing the input $X \to X + P$ in order for the full network to classify it as the target class $\operatorname{argmax} f(X + P) = t$, and observing the corresponding intermediate layer predictions $\phi_l(X + P)$ across all layers $l \in \{0, 1, \ldots, L\}$.

Detailed experiment: Perturbing the input $X \to X + P$ in order for the prediction at layer $l = \alpha$ to be the target class $\operatorname{argmax} \phi_\alpha(X + P) = t$, and observing the corresponding intermediate layer predictions $\phi_l(X + P)$ across all layers $l \in \{0, 1, \ldots, L\}$, both before and after α.

2.1 Generating the attacks

In all our experiments, we are using a simple approach for finding adversarial perturbations that was described in the original paper Szegedy et al. [2013]. We calculate the cross-entropy loss $\mathcal{L}$ of the model predictions with respect to the target label t. We then use the input image gradient $\nabla_X \mathcal{L}(\phi_l(X))$ with respect to the loss. We then employ the `Adam` optimizer Kingma and Ba [2014] and run it for 10 steps at the learning rate $\eta = 0.01$.

3 Experimental Results

We use an ImageNet-1k [Deng et al., 2009] pretrained ResNet152[1] [He et al., 2015] finetuned on CIFAR-10 [Krizhevsky, 2009] via trained linear probes. Each probe is trained for a single epoch with the Adam optimizer at the learning rate $\eta = 0.001$.

[1] `https://pytorch.org/vision/stable/models/generated/torchvision.models.resnet152.html`

In Figure 3 we showcase the effect of intermediate layers not beeing fooled by an adversarial attack on the full network. Images attacked to look like some other class than their ground truth (to the final layer classification) do not look like that to intermediate layers, as shown by the target class probability only rising in the very last layers (see Figure 3). We can therefore confirm that indeed the activations of attacked images do not look like the target class in the intermediate layers.

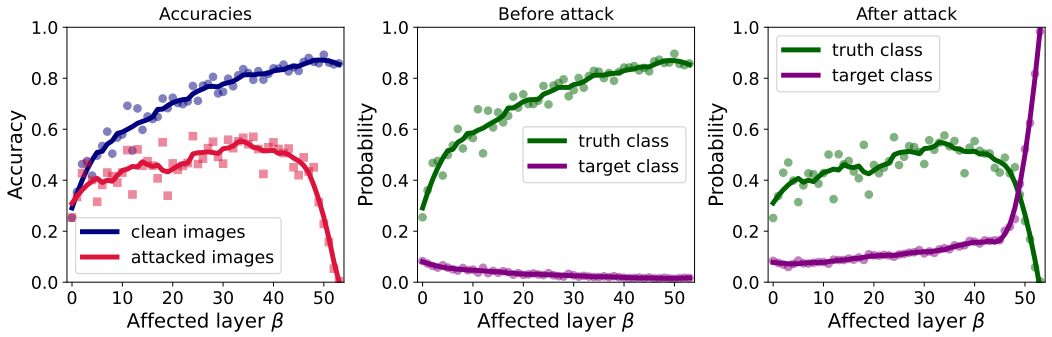

Figure 3: The impact of adversarial attacks ($L_\infty = 8/255$, 10000 attacks) against the full classifier on the accuracy and probabilities at all intermediate layers for an ImageNet-1k pretrained ResNet152 finetuned on CIFAR-10 via trained linear probes. The left panel shows the prediction accuracy on clean, unperturbed images, which rises from layer to layer, and the accuracy on adversarially attacked images, which is only lightly affected for all layers apart from the very last ones. These are the closest to the last layer, whose classification the attack was designed against. On the right panel, the mean predicted probability of the ground truth class and the target class of the adversary (always different from the ground truth) are shown. The target class probability only rises for the very last layers. Therefore the intermediate activations of an adversarially attacked image do not look like the target class, retaining the character of the original class instead.

This setup also allows us not only to investigate what the intermediate classification decision would be for an adversarially modified image X' that confuses the network's final layer classifier, but also to generally ask what the effect of confusing the classifier at layer α would do to the logits at a layer β. The results are shown in Figure 4 for 6 selected layers to attack.

We find that attacks designed to confuse early layers of a network do not confuse its middle and late layers. Attacks designed to fool middle layers do not fool early nor late layers, and attacks designed to fool late layers do not confuse early or middle layers. In short, there seems to be roughly a 3-way split: early layers, middle layers, and late layers. Attacks designed to affect one of these do not generically generalize to others. We call this effect the *adversarial layer de-correlation*.

We conducted equivalent experiments with stronger attacks (higher L_∞) as well as on other networks (e.g. the Vision Transformer Dosovitskiy et al. [2021]) with qualitatively equivalent results. We believe that this shows a general feature of how neural networks build up their hierarchical representations.

3.1 Passive attack detection and classification based on intermediate layers

To check if we can use the intermediate layer probabilities as a passive detector of adversarial attack, we attacked the first 5000 images of the CIFAR-10 test set with $L_\infty = 8/255$ attacks. For an image, the intermediate layer predictions give us 54 probability vectors of 10 classes. Flattening this into 540-dimensional vectors, we tried the following: 1) distinguishing adversarially attacked images from original, unperturbed images based on just the predicted probabilities (and emphatically not knowing the ground truth class or the target class of the attack), and 2) predicting the ground truth class from this vector alone.

For both, we used a simple 3-layer affine neural network with `ReLU` activations, predicting 2 classes (original and attacked) in the first case, and 10 classes in the second case. We evaluated its success on the next 5000 test set images. For distinguishing unperturbed vs attacked images, we got $\approx 99\%$ train set accuracy, and $\approx 94\%$ test set accuracy, demonstrating that we can use the intermediate layer features as a *passive* attack flag. For determining the ground truth class, we got $\approx 94\%$ on the train

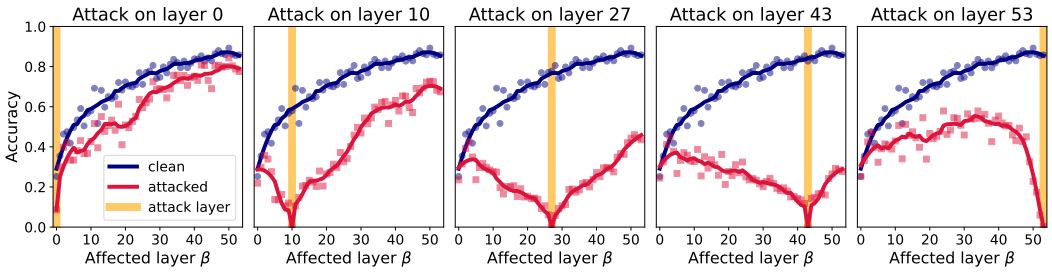

Figure 4: Transfer of adversarial attacks ($L_\infty = 8/255$, 1000 attacks) against the activations of layer α on the accuracy of layer β for $\alpha = 0, 10, 27, 43, 53$ on ImageNet-1k pretrained ResNet152 finetuned on CIFAR-10 via trained linear probes. Each panel shows the effect of designing a pixel-level attack to confuse the linear probe at a particular layer. The blue curve is the test accuracy on the unperturbed data, and the red line shows the accuracy on the attacked images. The accuracy drops to 0 at the layer that is directly attacked (marked in orange), showing a successful attack. The effect is localized: attacking early layers mainly affects early layer predictions, middle layer attacks primarily affect middle layers, and likewise attacks on the final layers (the standard regime) primarily influence late layer performance.

set (equal mixture of attacked and original images), $\approx 88\%$ on the test set of unperturbed images, and $\approx 69\%$ on a test set of attacked images. The attack original drove this accuracy to 0%, out of which we recovered to $\approx 69\%$. This shows that using the intermediate layer features, we can recover the ground truth class of the image with high fidelity *after* the attack, i.e. not in a white-box regime where the attacker can back-propagate gradients both through the network as well as this aggregating function (the way to do that is discussed in [redacted].

4 Discussion and Conclusion

In this paper we experimentally demonstrate a surprising empirical finding that intermediate layer representations in neural network classifiers are not fooled by adversarial attacks designed to fool the network as a whole. Furthermore, we demonstrate the fooling a particular layer's representation generally only affects the layers surrounding it, with both layers before and after partially covering their ability to see the true class of the image in question. In other words, the susceptibility of hidden representations in a neural network to adversarial attacks is only partially correlated.

This can be used as a passive flag to detect if an image has been tampered with after the fact, and to an extent to even recover the ground truth class of the image. While this approach would not suffice in a white-box scenario, it is possible to use it to construct very robust neural network classifiers in vision as shown in [redacted].

References

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2013.

Stanislav Fort and Balaji Lakshminarayanan. Ensemble everything everywhere: Multi-scale aggregation for adversarial robustness, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. URL `https://ieeexplore.ieee.org/abstract/document/5206848/`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Alex Krizhevsky. Learning multiple layers of features from tiny images. pages 32–33, 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL `https://arxiv.org/abs/2010.11929`.