Adversarial Training Versus Weight Decay

Angus Galloway  
School of Engineering  
University of Guelph, Canada  
Vector Institute, Canada  
gallowaa@uoguelph.ca

Thomas Tanay  
CoMPLEX  
Dept. of Computer Science  
University College London, UK  
thomas.tanay.13@ucl.ac.uk

Graham W. Taylor  
University of Guelph, Canada  
Canadian Institute for Advanced Research  
Vector Institute, Canada  
gwtaylor@uoguelph.ca

Abstract

Performance-critical machine learning models should be robust to input perturbations not seen during training. Adversarial training is a method for improving a model's robustness to some perturbations by including them in the training process, but this tends to exacerbate other vulnerabilities of the model. The adversarial training framework has the effect of translating the data with respect to the cost function, while weight decay has a scaling effect. Although weight decay could be considered a crude regularization technique, it appears superior to adversarial training as it remains stable over a broader range of regimes and reduces all generalization errors. Equipped with these abstractions, we provide key baseline results and methodology for characterizing robustness. The two approaches can be combined to yield one small model that demonstrates good robustness to several white-box attacks associated with different metrics.¹

1 Introduction

Machine learning models have been shown to work well in a variety of domains, for the equivalent of laboratory conditions in the physical sciences. As engineers increasingly opt to deploy these models in performance-critical applications such as autonomous driving or in medical devices, it is necessary to characterize how they behave in the field. In the aforementioned applications, the confidence with which a prediction was made can be even more important than the prediction itself. Unfortunately, deep neural networks (DNNs) leave much to be desired, typically offering little interpretability to their predictions, unreliable confidence estimates, and tending to fail in unexpected ways (Biggio et al., 2013; Szegedy et al., 2014; Goodfellow et al., 2015; Nguyen et al., 2015). The adversarial examples phenomenon afflicting machine learning models has shed light on the need for higher levels of abstraction to confer reliable performance in a variety of environments, where models may be pushed to the limit intentionally by adversaries, or unintentionally with out-of-distribution data. Although such settings could be seen as violating the i.i.d assumption implicit in typical machine learning experiments, raising this objection provides little comfort when systems fail unexpectedly.

In this work, we compare the approach of augmenting training data with worst-case examples (Szegedy et al., 2014), which can result in somewhat mysterious effects (Kurakin et al., 2017; Tramèr et al., 2018), to weight decay, which is more mathematically understood (Neyshabur et al., 2015; Goodfellow et al., 2016). There remains a widely held belief that adversarial examples are linked to the linear behaviour of machine learning models (Goodfellow et al., 2015). We begin by developing stronger intuition for linear models, finding that adversarial training can be seen as translating the data w.r.t. the loss, while weight decay scales the loss. With these abstractions, we report a 40% improvement in absolute accuracy over a previous baseline, while incurring a modest 2.4% loss for the natural data.

We leverage the translation and scaling perspective to tune a low-capacity model to be more robust for medium to large perturbations bounded in the Hamming distance (or $L_0$ counting “norm”) for the CIFAR-10 dataset (Krizhevsky, 2009) than another compelling defense against white-box attacks (Madry et al., 2018). The same model is competitive for $L_{\infty}$ bounded attacks, and outputs high-entropy distributions over its softmax-probabilities for out-of-distribution examples, such as fooling images (Nguyen et al., 2015). The model

¹Source code and pre-trained models available [https://github.com/uoguelph-mlrg/adversarial-training_vs_weight_decay]
2 Related Work

After two investigations of defenses against adversarial examples (Athalye et al., 2018; Uesato et al., 2018), the only approach that maintained a significant robustness improvement was based on adversarial training. Many previous works attempt to slow down attackers by manipulating or masking gradients at test time, but these do nothing to improve generalization. We refer the interested reader to Carlini and Wagner (2017b) and Athalye et al. (2018) for concise summaries of common mistakes. The iterative projected gradient descent (PGD) method with a random start, proposed by Madry et al. currently holds a benchmark for robustness to perturbations w.r.t. the CIFAR-10 test set, for a threat model in which an adversary may perturb any number of pixels by ±8 from their original integer values ∈ [0, 255]. However, there is growing concern that this $L_\infty$ norm threat model is unrealistically weak (Sharma and Chen, 2017; Tramèr et al., 2018; Athalye et al., 2018; Galloway et al., 2018).

By slightly extending the threat model of Madry et al., the PGD defense has been found to perform worse than an equivalent undefended model for MNIST (LeCun and Cortes, 1998), which is considered one of the easier datasets to defend on due to its low dimensionality and high separability. Weaknesses of the PGD defense for MNIST include adding a constant to all pixels (Galloway et al., 2018), rotating images by $20^\circ$ prior to attack, or randomly flipping just five pixels (Tramèr et al., 2018). An elastic-net black-box transfer attack (Sharma and Chen, 2017) achieved higher success rates than the attacks considered by Madry et al., but with less distortion in the same $L_\infty$ norm. For CIFAR-10, the official Madry et al. defense fails the fooling images test (Nguyen et al., 2015) by classifying with maximum confidence unrecognizable images initialized from uniform noise (Galloway et al., 2018).

A formal certificate of robustness based on adversarial training was recently provided by Sinha et al. (2018), and while their principled approach is laudable, the threat model considered is weaker than that of Madry et al. (2018) and exclusively deals with imperceptibly small perturbations, which were not well defined. The translation view of adversarial training proposed in this work offers insight into the difficulty of using the framework with large perturbations, and suggests that it may indeed be inherently limited to the small perturbation regime. In support of the threat model chosen by Madry et al., there are still few standards for assessing robustness, and the $L_\infty$ metric facilitates convenient high-level comparisons based on a single data-point. To the best of our knowledge, the original use of $L_\infty$ for bounding adversarial examples arose from a desire to mitigate perturbations akin to steganography. Under $L_\infty$, the cumulative effect of changes to each input dimension (i.e. pixel) smaller than the least resolvable feature in the digital representation of the training data (e.g. with 8-bits), is to be “decoded” such that any desired class is predicted with high-confidence (Goodfellow et al., 2015).

There are many types of perturbations models could be subject to, therefore we intentionally depart from the traditional security literature to explore trade-offs implicit in choosing a specific threat model. We sweep perturbation magnitude through the full dynamic range, or until accuracy is reduced to less than chance, for attacks that are optimized for different $L_p$ metrics ($e.g.$ $p \in 0, 1, 2, \infty$). For security, performance is no better than the worst of any particular case, but depending on the application, practitioners may wish to choose the model that maximizes area under all curves. Attacking until models reach zero accuracy was proposed by Athalye et al. (2018) to rule out gradient masking as contributing to a false positive defense. We find that additional insight can be gained even between chance and zero, therefore we usually opt for the latter.

It was recently shown that regularizing shallow convolutional neural networks with strong $L_2$ weight decay does mitigate the fooling images attack, also known as “rubbish” examples, for the MNIST and CIFAR-10 datasets (Galloway et al., 2018). For $L_\infty$ bounded perturbations of natural CIFAR-10 data, these models perform more closely to the PGD-trained WideResNet benchmark (Madry et al., 2018) than the other defenses tested by Athalye et al. (2018), despite having only 1% as many parameters.

Previous works briefly tested or acknowledged the role of weight decay as a possible defense against adversarial perturbations, but it is generally discarded along with other “regularization techniques” as adversarial examples are not typically viewed as an overfitting problem (Szegedy et al., 2014; Goodfellow et al., 2015). Following works generally refer to this preliminary conclusion, and do not consider weight decay given a constraint of not “seriously harming” classification accuracy on benign inputs (Papernot et al., 2016a).

---

2 The encoding of secrets in plain sight, e.g., in an image.
3 For $p = 0$, we mean Hamming distance.
4 In a personal communication with the first author, they clarified that adversarial examples are partly due to overfitting, but that this is not the only explanation for their existence.
We maintain that most adversarial examples observed in practice are of the “Type 2”—overfitting—variety in the taxonomy of Tanay and Griffin (2016), and that “serious harm” is too arbitrary for a constraint, and should probably be a criteria. Other works used a small weight decay penalty concurrently with other regularization techniques (e.g. dropout) that exhibit different behaviour, and where the contribution of the norm penalty is insignificant, or where it is only applied to the last layer of a deep network (Cisse et al., 2017).

Rozsa et al. (2016) argue that correctly classified examples have a small contribution in the cross-entropy loss, leading to “evolutionary stalling” that prevents the decision boundary from flattening near the training data. However, if we consider the high-dimensional geometry of finite volume sub-manifolds corresponding to each class in a fixed dataset, flatter shapes actually minimize the smallest adversarial distance, while a manifold with constant positive curvature (e.g. a sphere) maximizes distance in all directions. Rozsa et al. found that batch adjusting network gradients (BANG) increases overfitting on natural data, which they also address with dropout, further complicating the experiments. Both Rozsa et al. and Cisse et al. evaluate on weak single-step attacks that linearly approximate the loss of their deep models. Success against linear attacks can be improved by making models less linear, but non-linear models do not necessarily generalize better. Non-linearity is defeated with iterative attacks which can still yield examples that transfer, thus enabling black-box attacks (Tramèr et al., 2018).

Tramèr et al. found that adversarial training on ImageNet often converges to a degenerate global minimum, leading to a kind of learned gradient masking effect. Rather than becoming robust, it’s conceivable that models may intentionally contort their decision boundaries s.t. weaker adversarial examples are crafted that are easy to classify. Tramèr et al. attempt to prevent this by transferring examples from an ensemble of pre-trained models, hence decoupling adversarial example generation from the model being learned. They claim improved robustness to some black-box transfer attacks, but at the expense of white-box performance. The ensemble method was omitted from the analyses of Athalye et al. (2018), as white-box robustness is strictly more difficult than in the black-box setting. The technique was shown to not hold in a more general black-box threat model using a query-efficient gradient estimator (Ilyas et al., 2017, Ilyas 2018), and a gradient-free genetic algorithm (Alzantot et al., 2018).

Our translation perspective can explain why decoupling adversarial example generation is insufficient: degenerate global minima are an inherent result arising from training on worst-case perturbations with fixed magnitude, which fails to account for changing distances between individual training data and the decision boundary as a model learns. To perform adversarial training in a way that is mathematically correct requires taking into account the distance of each example from the class mean.

Recently, there has been a renewed interest in developing knowledge in norm-based capacity control of linear and depth-limited networks (Bengio et al., 2006; Neyshabur et al., 2015) as an explanation for generalization in deep learning (Neyshabur et al., 2017; Tsuzuku et al., 2018). On the other hand, there remains an operating hypothesis in many recent works that the existence of adversarial examples is linked to the linear behaviour of machine learning models (Goodfellow et al., 2015). Norm-based capacity control is well-characterized for linear models, letting us explicitly deal with adversarial examples as a subset of the generalization problem in this setting. Linear models are thought to be highly interpretable, or at least more so than their deep non-linear counterparts, therefore it can seem unintuitive that they may experience adversarial examples. However, we caution against using this premise to refute attempts at making models more interpretable in general, as linear models do not always suffer from adversarial examples.

Previous work has identified limitations with the linear interpretation for the existence of adversarial examples, through manipulating hidden representations in DNNs with feature adversaries (Sabour et al., 2015), the boundary tilting perspective (Tanay and Griffin, 2016), and an insightful commentary on high-dimensional geometry (Dube, 2018). Most similar to this work is that of Tanay and Griffin; however they did not empirically validate their perspective with DNNs, nor discuss adversarial training. To limit further confusion around linearity, we begin by characterizing linear models for binary classification, then evaluate DNNs on CIFAR-10.

3 Robust Logistic Regression

To support our rejection of the “linearity hypothesis”, we re-visit Section 5 from Goodfellow et al. (2015) in which it is motivated by observing that a linear classifier, logistic regression, is vulnerable to adversarial examples for MNIST “3” versus “7” binary classification. They introduce a single-step attack known as the fast gradient method (FGM), of which the sign variant (FGSM) is the strongest possible attack against a linear model when an adversary is constrained in the $L_\infty$ norm. Perturbations
are scaled by a constant $\epsilon$ when using the $L_{\infty}$ norm, where $\epsilon$ can always be interpreted as a fraction of the input range which has been normalized between $[0, 1]$ when it appears in decimal form. In this setting, Goodfellow et al. reported a 99% test error rate for an FGSM perturbation scaled by $\epsilon = 0.25$, which we denote “the baseline” in Sections 3 and 4. To the best of our knowledge, no updated result has since been reported for this experiment. Goodfellow et al. later compared the regularizing effects of adversarial training on FGSM examples to $L_1$ weight decay, which we revisit in Section 4.

For the experiments in this section, we use a parameter initialization scheme that consists of subtracting the average MNIST “7” from the average “3”, which we refer to as “expert initialization”. By simply using this initialization, we obtain a 39% absolute percentage improvement over the baseline, but the model underfits with $\approx 5.2\%$ error on the clean test set. These are approximately the resulting set of weights after empirical risk minimization (ERM) fully converges when training from scratch with $L_2$ weight decay, if one accounts for minor differences due to stochastic optimization with a particular batch size, learning rate, etc. In addition to conferring good robustness, we use this expert scheme to facilitate reproducibility. We depict weights, and perturbations crafted by FGSM in Figure 1. Goodfellow et al. comment on Figure 1(c) in their original caption that the perturbation is not readily recognizable even though the model is fit well.

The perturbation image shown in Figure 1(c) for the FGSM attack is approximately identical to the binary weights of a binarized model trained on this task from scratch, as shown in Figure 1(b). We know the binarized model is not an optimal model here, as we would like it to be totally invariant to the areas around the border where the real-valued model’s weights, and the difference between the average “7” and “3”, are almost zero. The binarized model tries to be as invariant as possible to the background by roughly equally distributing the polarity of weights in these areas. Cropping the dataset more closely around the digits, or allowing a third parameter state (e.g. by using a ternary network, or pruning connections to bordering pixels) would make the model more robust. In terms of the boundary tilting perspective, such a model minimizes tilting along dimensions of low-variance in the data (Tanay and Griffin 2016).

Using the expert initialization scheme and training procedure from Figure 1(a) our “Expert-$L_2$” model achieves 59% test error for an FGSM attack, compared to the 99% baseline error rate, for the same perturbation size $\epsilon = 0.25$ w.r.t. the $L_{\infty}$ norm. This 40% improvement is consistently reproducible, but the model incurs a 4% error rate on the natural test set, compared to 1.6% error for the baseline model. We believe the 2.4% loss in “clean” test accuracy in exchange for a 40% gain for this attack is a good trade-off. After a brief search, we find that a randomly initialized model can handle very large norm penalties. For example, training from scratch for 50k steps with $\lambda = 3.25$ maintains the $\approx 4\%$ test error of the Expert-$L_2$ model, while achieving 43% and 64% FGSM error rates for $L_1$ and $L_2$ weight decay respectively. Next we explain why we prefer penalizing the $L_2$ norm of the weights rather than the $L_1$ norm, despite the slightly higher FGSM error rate for the former.

It may seem that a 50% FGSM error for $\epsilon = 0.25$ is still unreasonably high, but we argue that this data point is fairly uninformative in isolation and does not imply robustness, or a lack thereof. With exceedingly high $L_1$ weight decay (e.g. with $\lambda = 32$), we can further reduce

\[ L_{\infty} \text{bounded FGSM is the strongest attack against the logistic regression model, it is not the most efficient attack.} \]

The gradient sign in Figure 1(c) is recognizable as binarizing the continuous valued weights, and the roughly 3–5 pixel wide noisy border reflects directions where the weights are close to, but not exactly zero.\footnote{Implementations of the \texttt{sign} method, e.g. in \texttt{numpy} and TensorFlow do pass through 0 as 0.} We can induce a nearly identical error rate, but with less distortion by using the actual gradient, and normalizing it to have unit $L_2$ norm, the “Fast grad $L_2$” method from Kurakin et al. (2017) (their Appendix A).

Figure 1: Visualization of (a) logistic regression model weights after expert initialization, and fine-tuning for 10k steps on the sigmoid cross entropy loss with Adam (Kingma and Ba 2014), a batch size of 128, learning rate of 1e-5, and an overall $L_2$ regularization constant of 0.05. (b) A binarized set of weights after training with the same procedure from scratch for 50k steps. Plots (c) and (d) are taken from Goodfellow et al. (2015) for comparison purposes, (c) is an FGSM perturbation for true class label “7”, and (d) the actual weights. The faint dark pixels in (d) resemble a slightly rotated “?” symbol, whereas “3” and “7” are visible in (a). We argue that there is a clear interpretation of such a perturbation. In this particular case, one’s intuition is partly obscured due to having used the sign of the gradient (FGSM) rather than the real gradient. Although the

\[ L_{\infty} \text{bounded FGSM is the strongest attack against the logistic regression model, it is not the most efficient attack.} \]

The gradient sign in Figure 1(c) is recognizable as binarizing the continuous valued weights, and the roughly 3–5 pixel wide noisy border reflects directions where the weights are close to, but not exactly zero.\footnote{Implementations of the \texttt{sign} method, e.g. in \texttt{numpy} and TensorFlow do pass through 0 as 0.} We can induce a nearly identical error rate, but with less distortion by using the actual gradient, and normalizing it to have unit $L_2$ norm, the “Fast grad $L_2$” method from Kurakin et al. (2017) (their Appendix A).

The perturbation image shown in Figure 1(c) for the FGSM attack is approximately identical to the binary weights of a binarized model trained on this task from scratch, as shown in Figure 1(b). We know the binarized model is not an optimal model here, as we would like it to be totally invariant to the areas around the border where the real-valued model’s weights, and the difference between the average “7” and “3”, are almost zero. The binarized model tries to be as invariant as possible to the background by roughly equally distributing the polarity of weights in these areas. Cropping the dataset more closely around the digits, or allowing a third parameter state (e.g. by using a ternary network, or pruning connections to bordering pixels) would make the model more robust. In terms of the boundary tilting perspective, such a model minimizes tilting along dimensions of low-variance in the data (Tanay and Griffin 2016).

Using the expert initialization scheme and training procedure from Figure 1(a) our “Expert-$L_2$” model achieves 59% test error for an FGSM attack, compared to the 99% baseline error rate, for the same perturbation size $\epsilon = 0.25$ w.r.t. the $L_{\infty}$ norm. This 40% improvement is consistently reproducible, but the model incurs a 4% error rate on the natural test set, compared to 1.6% error for the baseline model. We believe the 2.4% loss in “clean” test accuracy in exchange for a 40% gain for this attack is a good trade-off. After a brief search, we find that a randomly initialized model can handle very large norm penalties. For example, training from scratch for 50k steps with $\lambda = 3.25$ maintains the $\approx 4\%$ test error of the Expert-$L_2$ model, while achieving 43% and 64% FGSM error rates for $L_1$ and $L_2$ weight decay respectively. Next we explain why we prefer penalizing the $L_2$ norm of the weights rather than the $L_1$ norm, despite the slightly higher FGSM error rate for the former.

It may seem that a 50% FGSM error for $\epsilon = 0.25$ is still unreasonably high, but we argue that this data point is fairly uninformative in isolation and does not imply robustness, or a lack thereof. With exceedingly high $L_1$ weight decay (e.g. with $\lambda = 32$), we can further reduce

\[ L_{\infty} \text{bounded FGSM is the strongest attack against the logistic regression model, it is not the most efficient attack.} \]

The gradient sign in Figure 1(c) is recognizable as binarizing the continuous valued weights, and the roughly 3–5 pixel wide noisy border reflects directions where the weights are close to, but not exactly zero.\footnote{Implementations of the \texttt{sign} method, e.g. in \texttt{numpy} and TensorFlow do pass through 0 as 0.} We can induce a nearly identical error rate, but with less distortion by using the actual gradient, and normalizing it to have unit $L_2$ norm, the “Fast grad $L_2$” method from Kurakin et al. (2017) (their Appendix A).

The perturbation image shown in Figure 1(c) for the FGSM attack is approximately identical to the binary weights of a binarized model trained on this task from scratch, as shown in Figure 1(b). We know the binarized model is not an optimal model here, as we would like it to be totally invariant to the areas around the border where the real-valued model’s weights, and the difference between the average “7” and “3”, are almost zero. The binarized model tries to be as invariant as possible to the background by roughly equally distributing the polarity of weights in these areas. Cropping the dataset more closely around the digits, or allowing a third parameter state (e.g. by using a ternary network, or pruning connections to bordering pixels) would make the model more robust. In terms of the boundary tilting perspective, such a model minimizes tilting along dimensions of low-variance in the data (Tanay and Griffin 2016).

Using the expert initialization scheme and training procedure from Figure 1(a) our “Expert-$L_2$” model achieves 59% test error for an FGSM attack, compared to the 99% baseline error rate, for the same perturbation size $\epsilon = 0.25$ w.r.t. the $L_{\infty}$ norm. This 40% improvement is consistently reproducible, but the model incurs a 4% error rate on the natural test set, compared to 1.6% error for the baseline model. We believe the 2.4% loss in “clean” test accuracy in exchange for a 40% gain for this attack is a good trade-off. After a brief search, we find that a randomly initialized model can handle very large norm penalties. For example, training from scratch for 50k steps with $\lambda = 3.25$ maintains the $\approx 4\%$ test error of the Expert-$L_2$ model, while achieving 43% and 64% FGSM error rates for $L_1$ and $L_2$ weight decay respectively. Next we explain why we prefer penalizing the $L_2$ norm of the weights rather than the $L_1$ norm, despite the slightly higher FGSM error rate for the former.

It may seem that a 50% FGSM error for $\epsilon = 0.25$ is still unreasonably high, but we argue that this data point is fairly uninformative in isolation and does not imply robustness, or a lack thereof. With exceedingly high $L_1$ weight decay (e.g. with $\lambda = 32$), we can further reduce

\[ L_{\infty} \text{bounded FGSM is the strongest attack against the logistic regression model, it is not the most efficient attack.} \]
the FGSM error rate to 25%, but what do we suspect the robustness behaviour of this model will be?

Even for the linear case where the FGSM is exact, one score does not tell the full story. Although these types of comparisons are convenient for researchers, they are potentially harmful in the long run as this rewards overfitting a particular threat model, which we have already observed. We leave it up to researchers to rigorously attack their proposed defenses, and suggest that characterizing the slope of the accuracy roll-off curve for a variety of optimization-based or engineered attacks makes a more compelling defense. As a further assessment, the nature of the distortion introduced by the attack may be of interest when expressed quantitatively and qualitatively. The 59% error rate for Expert-\(L_2\) coincides with a mean \(L_2\) distortion of \(\approx 4\%\) over the full test set. As this distortion metric is not always representative of perceptual distance (e.g. it is not translation invariant), we include candidate adversarial images for Expert-\(L_2\) in Figure 2.

![Figure 2](image)

**Figure 2:** (a) Fast grad \(L_2\) perturbation for original class “3”, and candidate adversarial examples for original class “3” (b)-(d) and “7” (e)-(g). All images remain correctly classified when perturbation bounded by \(\epsilon = 0.25\) in the \(L_{\infty}\) norm, we therefore call them only candidate examples as in Kurakin et al. (2017). The average “3” is clearly visible in images originally depicting a “7”, and vice versa.

To show that an attack causing less distortion than FGSM can still be interpretable, we adapt the Fast grad \(L_2\) method by scaling the normalized gradient by \(100\times\) before clipping to \(\pm \epsilon\). Scaling by 100 lets us approximate the higher error rate caused by FGSM, while minimizing unnecessary distortion. This attack results in an error rate of 58.9%, compared to 59.1% on our Expert-\(L_2\) model when vanilla FGSM is used, but causes 4.00%–\(L_2\) distortion compared to 4.57% for FGSM. Clearly the images in Figure 2 are ambiguous even for human vision, and correct classification could be contingent on having previously seen a sampling of natural examples. These examples support the observation of Tanay and Griffin (2016) that:

...the weight vector found by logistic regression points in a direction that is close to passing through the mean images of the two classes, thus defining a decision boundary similar to the one of a nearest centroid classifier.

### 4 Adversarial Training Versus Weight Decay for Linear Models

In linear classification, weight decay is often explained as a way of enforcing a preference for models with small weights (e.g. section 5.2.2 of Goodfellow et al. (2016)). Yet there is no reason, a priori, for models with small weights to generalize better: all the weight vectors \(\lambda w\) with \(\lambda > 0\) define the same hyperplane boundary, whether \(\lambda\) is small or large. The norm of the weight vector is only meaningful in the context of ERM, because it influences the scaling of the loss function (Neyshabur et al. 2017; Tsuzuku et al. 2018).

We first introduce this natural interpretation of weight decay as affecting the scaling of the loss function, and then directly compare variants of adversarial training to \(L_1\) and \(L_2\) weight decay.

#### 4.1 Theoretical Analysis

Consider a logistic regression model with input \(x \in \mathbb{R}^d\), labels \(y \in \{\pm 1\}\), parameterized by weights \(w\), bias \(b\), and having sigmoid non-linearity \(\sigma(z)\). Predictions are defined by \(\sigma(w^\top x + b)\) and the model can be optimized by stochastic gradient descent (SGD) on the sigmoid cross entropy loss. This reduces to SGD on (1) where \(\zeta\) is the softplus loss \(\zeta(z) = \log(1 + e^{-z})\):

\[
E_{x,y \sim p_{\text{data}}} \zeta(y(w^\top x + b))
\]

In Figure 3 we develop intuition for the different quantities contained in (1) with respect to a typical binary classification problem. We will later see how this analysis generalizes to the multi-class setting.

We note that \(w^\top x + b\) is a scaled, signed distance between \(x\) and the classification boundary defined by our model. If we define \(d(x)\) as the signed Euclidean distance between \(x\) and the boundary, then we have: \(w^\top x + b = \|w\|_2 d(x)\). Hence, minimizing (1) is equivalent to minimizing:

\[
E_{x,y \sim p_{\text{data}}} \zeta(\|w\|_2 \times y d(x))
\]

We define the scaled softplus loss as:

\[
\zeta_{\|w\|_2}(z) := \zeta(\|w\|_2 \times z)
\]

And we see that adding a \(L_2\) weight decay term in (2), resulting in (4), can be understood as a way of controlling the scaling of the softplus function:

\[
E_{x,y \sim p_{\text{data}}} \zeta_{\|w\|_2}(y d(x)) + \lambda \|w\|_2
\]

*Our function \(\zeta\) is the mirror of the one used in Goodfellow et al. (2015): \(\zeta(z) \leftarrow \zeta(-z)\). This definition removes the superficial negative sign inside \(\zeta\) later on.*
Figure 3: (a) For a given weight vector \( \mathbf{w} \) and bias \( b \), the values of \( \mathbf{w}^\top x + b \) over the training set typically follow a bimodal distribution (corresponding to the two classes) centered on the classification boundary. (b) Multiplying by the label \( y \) allows us to distinguish the correctly classified data in the positive region from misclassified data in the negative region. (c) We can then attribute a penalty to each training point by applying the softplus loss to \( y(\mathbf{w}^\top x + b) \).

Figure 4 illustrates the effect of varying \( \lambda \), the regularization parameter for \( \| \mathbf{w} \|_2 \).

Figure 4: (a) For a small regularization parameter (large \( \| \mathbf{w} \|_2 \)), the misclassified data is penalized linearly while the correctly classified data is not penalized. (b) A medium regularization parameter (medium \( \| \mathbf{w} \|_2 \)) corresponds to smoothly blending the margin. (c) For a large regularization parameter (small \( \| \mathbf{w} \|_2 \)), the entire data is penalized almost linearly.

Similarly, we can interpret adversarial training as acting on the horizontal offset of \( yd(x) \). Training on worst case adversarial examples consists of replacing \( x \) with \( x' = x - \epsilon y \hat{w} \) (where \( \hat{w} = w/\|w\|_2 \)). We then have:

\[
y(w^\top x' + b) = y(w^\top (x - \epsilon y \hat{w}) + b) = y(w^\top x + b) - \epsilon \|w\|_2 = \|w\|_2 \times (yd(x) - \epsilon)
\]

Substituting the above expression into the scaled softplus function \( f \) yields:

\[
E_{x,y\sim\text{data}} \zeta_{\|w\|_2} (yd(x) - \epsilon)
\]

Equation (5)

Figure 5 illustrates the effect of varying the adversarial perturbation magnitude, or offset parameter \( \epsilon \).

The softplus function has a linear region penalizing the misclassified data, a saturated region leaving the correctly classified data unaffected, and a smooth transition between these two regions enforcing a safety margin. Both weight decay and adversarial training can be interpreted as influencing how these different regions interact with the data. Weight decay strikes an effective balance between the two extreme regimes: when it is very low, the margin blend is very sharp, as seen in Figure 4(a). When it is very high, nearly all of the data falls into the margin, shown in Figure 4(c), resulting in underfitting, but no risk of instability as the data does not shift. Adversarial training artificially increases the size of the margin by translating some or all of the data into the misclassified region, as in Figure 5(b). It becomes unstable when the offset used is too strong, which we depict in Figure 5(c) and verify experimentally in Section 4.2. Overall, weight decay seems more theoretically sound given that it allows a broader variety of behaviors and remains stable for all regimes, whereas the adversarial training framework is inherently limited to small perturbations. This translation and scaling perspective explains why adversarial training and weight decay can complement each other when small perturbations are used, but ultimately, weight decay is the more sound instrument.

In this section we focused on the \( L_2 \) norm case which is, in our opinion, more intuitive since \( d(x) \) can then be interpreted as the Euclidean distance from the boundary. However, \( L_1 \) weight decay and FGSM adversarial training can be shown to implement similar transformations in a different normed space, thus enforcing robustness to different types of perturbations. We perform experiments using both norms in the following section.

4.2 Experiments

Consider the same logistic regression model as in previous sections. We use the standard form of the softplus loss \( f(\mathbf{z}) = \log(1 + e^\mathbf{z}) \) to facilitate a direct comparison.
comparison with Goodfellow et al. (2015). Training the model consists in minimizing the loss \( (6) \) with SGD:

\[
\mathbb{E}_{x,y \sim \text{data}} \zeta(-y(w^\top x + b)).
\]  

(6)

The gradient of \( (6) \) with respect to the input \( x \) is \(-yw\), and its sign is \(-y \text{sign}(w)\). Note that the label \( y \) was omitted in the original Goodfellow et al. paper, resulting in an asymmetric penalty for adversarial training. Training on worst case \( L_\infty \) bounded examples for logistic regression consists of replacing \( x \) in \( (6) \) with \( x - \epsilon y \text{sign}(w) \), resulting in \( (7) \). The perturbation term \( \epsilon y w^\top \text{sign}(w) \) is equivalent to an \( \epsilon y \|w\|_1 \) penalty on the weights.

\[
\mathbb{E}_{x,y \sim \text{data}} \zeta(-y(w^\top x + b) - \epsilon \|w\|_1)
\]  

(7)

Compare the expression \( (7) \) to instead adding the \( L_1 \) penalty controlled by regularization factor \( \lambda \), outside of \( \zeta(z) \) as in \( (8) \):

\[
\mathbb{E}_{x,y \sim \text{data}} \zeta(-y(w^\top x + b)) + \lambda \|w\|_1.
\]  

(8)

We now provide a learning theoretic observation of the differences between the two expressions, which complements the scaling and translation view. The parameter norm penalty in \( (7) \) is trapped in \( \zeta(z) \), and interacts with \( y \) (the true or predicted label), but this is not the case with weight decay \( (8) \). This leads to a possible perverse instantiation of the ERM minimization objective in \( (7) \) if \( \epsilon \|w\|_1 \) overpowers \(-y(w^\top x + b)\), as the model may intentionally learn to make wrong predictions in order to “escape” the saturating region of \( \zeta \). This makes predictions confident enough to saturate \( \zeta \), and the model may intentionally learn to make wrong predictions in order to “escape” the saturating region of \( \zeta \). This results in a positive feedback loop for adversarial training.

In accordance with the theoretical discussion of equations \( (7) \) and \( (8) \), we observe a positive feedback loop for adversarial training. Training on worst case \( L_\infty \) bounded examples for logistic regression consists of replacing \( x \) in \( (6) \) with \( x - \epsilon y \text{sign}(w) \), resulting in \( (7) \). The perturbation term \( \epsilon y w^\top \text{sign}(w) \) is equivalent to an \( \epsilon y \|w\|_1 \) penalty on the weights.

Predictions of either class are systematically reduced by \( \epsilon \|w\|_1 \), which leads to the bimodal distribution shown in Figure 3(a) shifting inward. Activations, and therefore \( \|w\|_1 \), decrease monotonically, while the training loss increases monotonically since predictions are made with less confidence. When we finally enter the non-linear region of \( \zeta \), the training loss is high, but there are few non-zero parameters left to update, resulting in aggressive changes to just a few weights. Beyond a certain point, even strong weight decay can rescue learning if we consider training on the bi-stable examples generated with a constant \( \epsilon \) in Figure 2. In the case of a near perfect decision boundary, approximately that defined by the weights visualized in Figure 1(a) and \( \epsilon \) larger than the distance from an example to the boundary, we get examples that literally belong to the opposite class. Adversarial training becomes even more destructive, the closer a model is toward converging near a suitable solution.

We acknowledge that in most practical applications of interest we don't have a perfect boundary, but this simple case clearly contradicts the notion that the penalty disappears on its own, at least in the way that we would like it to. There is no reason to expect the situation will improve for higher-dimensional higher-variance problems like CIFAR-10 or ImageNet.

Figure 6 shows the effect of adversarial training on our logistic regression model. FGSM adversarial training always converges to the solution in Figure 6(b) given a sufficient number of steps, regardless of whether the true label or the model’s own prediction is used to construct the adversarial examples, independent of expert initial-
Figure 6: Weight visualization for (a) $\epsilon = 0.25$ FGSM training for 10k steps, and (b) 50k steps. (c) Training instead for 50k steps only with $L_1$ weight decay and $\lambda = 15$. Training with the same $\epsilon$ for Fast grad $L_2$ results in similar weights (not shown) as in (a) and (b) but a more interpretable perturbation image after 10k steps. Increasing $\epsilon$ to 0.5 results in non-stability and fails to train as the model lacks sufficient capacity to overfit the residual noise arising from overlapping the class centroids.

Figure 7: A comparison between (a) an $L_0$ attack, and (b) $L_\infty$–FGSM. Models trained with FGSM were more robust in (b), but less so in (a). An undefended model (“Natural”) was robust in (a) but not (b). Annotated is Expert-$L_2$ was reasonably robust to both. The model trained on FGSM examples used $\epsilon = 0.25$, and the $x$-intercept at $\approx 20$ pixels in (a) is translated left and right by training with different $\epsilon$. Annotates that the Expert-$L_2$ accuracy corresponding to the $\epsilon = 0.25$–FGSM error rate discussed in Section 3.

These degenerate solutions that obtain high FGM accuracy fail the fooling images test as they cannot generate natural looking data examples when gradient ascent is performed on the logit(s), despite making such predictions with high-confidence (Nguyen et al., 2015; Galloway et al., 2018). Additionally, an $L_0$ adversary can quickly cause a model tuned for the $L_\infty$ metric to predict at no better than chance by setting a small number of pixels corresponding to the locations of the min and max weights in Figures (b) and (c) to 1. The effect of this black-box $L_0$ attack is depicted in Figure 7(a) and shows how adversarial training can behave like the classic “whack-a-mole” game. Even though we used expert knowledge to craft this attack, it is unreasonable to assume that attackers would not be able uncover similar information given enough time. The attack also simulates the effect of “dead pixels” which could arise in a robotic system or sensor deployed in the wild.

Although the linear model has limited capacity, it is surprisingly robust to an FGSM attack when trained with FGSM, as shown in Figure 7(b). This is at odds with the recommendation of Mudry et al., whom suggest increasing capacity to train with a strong adversary. We argue that this becomes an arms race to use as much capacity as possible, and strongly advise against this as it encourages poor generalization outside the considered threat model. In this case, for the $L_\infty$ metric FGSM is the strongest possible adversary, and the model appears robust until we consider our trivial $L_0$ attack. Note that there was an insignificant difference between using the true label or the model’s own prediction when generating the training examples, since predictions closely matched the true labels. To summarize the results of Figure 7, models trained with FGSM were robust to FGSM, but not to an $L_0$ attack. Unprotected models were robust to the $L_0$ attack, but not FGSM. Models trained with $L_2$ weight decay were reasonably robust to both, and their performance declined slowly and predictably for unbounded perturbations.

4.3 Discussion

When a randomly initialized model is trained without weight decay, spurious data are fit more tightly, leaving the boundary near the data sub-manifolds. Adversarial training examples generated with such a boundary are projected by only a small distance, having a small but increasingly positive effect. Given that the effect is small initially, there is a natural inclination to use a large $\epsilon$. Medium to high $\epsilon$ adversarial training seems to work well in practice for high-capacity deep neural networks because they are able to fit the signal and the noise, an effect we aren’t the first to observe (Zhang et al., 2017). To use adversarial training in a more principled manner
implies that $\epsilon$ should at the very least be decayed when training, just as is commonly done with learning rates. Ultimately, the value of $\epsilon$ used with adversarial training is a poorly characterized constant that has been selected with limited intuition in much of the literature to date. To be mathematically correct, $\epsilon$ should be a function of each example’s distance from its class mean. Yet another hyper-parameter is introduced, as we must set both an $\epsilon$ decay schedule, as well as a unique $\epsilon$ per each image. This brings us back to weight decay, which typically only has one hyper-parameter, the regularization constant $\lambda$, and whose behaviour is reasonably well characterized at several levels of abstraction. We say reasonably well because its exact behaviour remains somewhat of a mystery in deep learning (Zhang et al., 2017), and if it was fully characterized then it would have been used in the baseline work to defend their logistic regression model against imperceptible adversarial examples.

There is ample room to further understand how norm constraints compare to other types of explicit and implicit regularizers used in deep learning, and how to bound generalization error in this setting (Neyshabur et al., 2017). As we will see in Section 6 it can be surprising that $L_2$ weight decay has a filter-level pruning effect when used with CNNs, rather than simply making all elements small. What remains in the first layer are smooth filters that implement fundamental image processing primitives, somewhat similar to the fully-connected weights of our Expert-$L_2$ model.

5 Deep Neural Networks

It was suggested in the seminal work of Szegedy et al. (2014) that the existence of adversarial examples in neural networks is related to the notion of Lipschitz constant, which measures the least-upper-bound of the ratio between output distances and corresponding input distances. This idea was further developed and formulated explicitly by Cisse et al. (2017), who suggested that a neural network’s sensitivity to adversarial perturbations can be controlled by the Lipschitz constant. Yet this argument is incomplete: just like rescaling the weight vector of a linear classifier does not affect the hyperplane boundary it defines, varying the Lipschitz constant of a neural network by rescaling its output affects neither its predictions nor its vulnerability to adversarial perturbations (Tsuzuku et al., 2018). Instead, we argue that the Lipschitz constant is only meaningful in the context of ERM because it influences the way the loss function distributes penalties over the training data (Neyshabur et al., 2017; Tsuzuku et al., 2018). Below, we generalize our considerations of Section 4.1 on binary linear classification to multiclass neural networks and we then reinterpret weight decay and adversarial training as part of a broader family of output regularizers.

5.1 Generalizing the Binary Linear Case

Consider a feedforward network $\mathcal{N}$ consisting of $l$ layers with parameters $W_1, \ldots, W_l$ and Rectified Linear Unit activations $\phi$. We assume that $\mathcal{N}$ does not use any weight or batch normalization and for simplicity, we also neglect the role of the biases. For an input image $x$, the logit layer of $\mathcal{N}$ can be written:

$$z(x) = W_l \phi(W_{l-1} \phi(\ldots \phi(W_1 x))) .$$

For a scalar $\lambda$ we have $\phi(\lambda x) = \lambda \phi(x)$ and therefore, decaying all the weights $W_i \leftarrow \lambda W_i$ is equivalent to decaying the logits:

$$z(x) \leftarrow \lambda^l z(x).$$

In binary classification, $z$ is a scalar and $\lambda^l$ acts as a scaling parameter for the softplus function. In multiclass classification, $z$ is a vector and $\lambda^l$ acts as a temperature parameter for the softmax function. We see that increasing weight decay produces smoother probability distributions, and therefore increases the contribution of correctly classified images to the final loss.

Adversarial training can similarly be understood as a way of altering the distributions of penalties over the data during training. Let $p_y(x)$ be the probability attributed to the true class $y$. For a perturbation $\delta$, a first order approximation of $p_y(x + \delta)$ can be written:

$$p_y(x + \delta) \approx p_y(x) + \partial_x p_y \cdot \delta$$

and for an $L_2$ constraint, the worse case perturbation is $\delta = -\epsilon \frac{\partial_x p_y}{\|\partial_x p_y\|_2}$, yielding:

$$p_y(x + \delta) \approx p_y(x) - \epsilon \|\partial_x p_y\|_2 .$$

Training on the perturbed image $x + \delta$ appears equivalent to training on the image $x$ with a probability attributed to the correct class that has been decreased by $\epsilon \|\partial_x p_y\|_2$. In practice, we reduce the extent to which this process is an approximation for deep networks by taking multiple steps of gradient descent with a small step size (Madry et al., 2018).

5.2 Output Regularization

The idea of regularizing a neural network by altering the way penalties are distributed over the data during training has been introduced multiple times before, independently. Pereyra et al. (2017) refer to this idea as “output regularization”. For instance, Szegedy et al. (2017) propose output regularization as a means to improve generalization. They suggest using a parameter $\epsilon$ to control the sensitivity of the classifier to adversarial perturbations. The idea is to add a term $\epsilon \|\partial_x p_y\|_2$ to the loss function, which acts as a regularizer on the output layer. This term encourages the classifier to predict with higher confidence on the true class, making it less vulnerable to adversarial attacks.

It is important to note that output regularization is not the same as weight decay. Weight decay is a form of implicit regularization that acts on the weights of the network, while output regularization acts on the output of the network. Weight decay reduces the magnitude of the weights, which can help prevent overfitting, while output regularization reduces the confidence of the classifier on the true class, which can help reduce the vulnerability to adversarial attacks.

In practice, output regularization can be implemented by adding a term $\epsilon \|\partial_x p_y\|_2$ to the loss function, where $\epsilon$ is a hyperparameter that controls the strength of the regularization. The value of $\epsilon$ can be tuned during training to find the optimal balance between generalization and robustness.

In conclusion, output regularization is a promising approach for improving the robustness of neural networks to adversarial attacks. By reducing the confidence of the classifier on the true class, output regularization can help prevent the classifier from being easily fooled by adversarial perturbations. Future work could focus on developing more effective output regularization techniques and studying their impact on the generalization performance of deep neural networks.
propose to penalize over-confident predictions with label smoothing, where one-hot labels are replaced with smooth distributions. Xie et al. (2016) implement a stochastic version of label smoothing with disturbLabel, where labels randomly become incorrect with a small probability. Pereyra et al. (2017) introduce an alternative to label smoothing in the form of an entropy-based penalty on high-confidence predictions. The BANG technique (Rozsa et al., 2016) discussed in Section 2 can also be interpreted as an output regularizer, as it was explicitly designed to increase the contribution of correctly classified images to the loss during training. Na et al. (2017) and Kannan et al. (2018) propose to add a term to the loss to encourage natural images and their adversarial counterparts to have similar logit representations. Kannan et al. (2018) also introduce logit squeezing, in which the norm of the logits is penalized.

Based on the considerations of the previous section, we can regard weight decay and adversarial training as two additional members of this family of output regularizers. Note that weight decay is particularly closely related to logit squeezing: the effect of both regularizers is to shrink the norm of the logits, and incidentally, the Lipschitz constant of the network.

6 CIFAR-10 Experiments

We now empirically verify previous claims motivated by studying linear models, in the context of deep learning. We compare a 32-layer WideResNet (Zagoruyko and Komodakis, 2016) with width factor 10, to a simple 4-layer CNN described in Table 1. The CNN architecture has 0.4% as many parameters as the WideResNet, and we expect its mapping from input to output to be more linear since it has $8 \times$ less depth.

![Table 1: Simple fully-convolutional architecture adapted from the CleverHans library tutorials (Papernot et al., 2017). Model uses ReLU activations, and does not use batch normalization, pooling, or biases. We respectively denote $h, w, c_{in}, c_{out}, s$ as kernel height, width, number of input and output channels w.r.t. each layer, and stride.](image)

The simple CNN was trained for 250 total epochs with label smoothing, the Adam optimizer, an initial learning rate of $1e^{-3}$, and batch size of 128. No data augmentation was used other than adversarial training when indicated, e.g. by “PGD (7–2–8)” for 7 iterations of PGD with a step size of 2, bounded by $\epsilon_{\text{max}}$ of 8—all out of 255—per convention for CIFAR-10. This 7–2–8 combination was used by Madry et al. (2018) to secure their models for CIFAR-10, therefore we use this as default.

We delay adversarial training until after the first 50 epochs to let the model partially converge so that it can generate “good” adversarial examples to train on for the remaining 200 epochs. This is also sometimes done for efficiency reasons, e.g. on ImageNet (Kurakin et al., 2017). We denote as “CNN-L2” and “CNN” models trained per the above procedure, with and without $L_2$ weight decay with regularization constant $\lambda = 1e^{-3}$.

We opt for 32 filters in conv1, as opposed to the CleverHans default of 64, because this was roughly the largest model which could be regularized s.t. it was unable to memorize training data examples, while maximizing clean test accuracy. We verify this by attempting to construct fooling images, and then evaluate our models on $L_\infty$, and two held-out $L_0$ counting “norm” attacks.

6.1 Fooling Images

Fooling images (Nguyen et al., 2015), also known as “rubbish examples” (Goodfellow et al., 2015), are a class of adversarial examples that begin with an out-of-distribution example, typically sampled from a Gaussian or uniform distribution, then maximizing the probability of a target class by accumulating several gradient updates in the input. This procedure has led to observations that we can easily cause DNNs to predict any chosen target class with high-confidence, even when the input remains completely unrecognizable to humans. This method is also motivated by model-based optimization, where one wants to generate examples that optimize some criteria, e.g. objects with a small drag coefficient, given a model that is only trained to distinguish, or regress over inputs.

When we use 64 filters in conv1 (and $2 \times$ this number in conv2, 3), we get significant filter sparsity, as both PGD and weight decay encourage this, yet the model can be coerced into reproducing some training data with fine detail. In addition to being a poor generalization strategy, this is a security vulnerability if the training data are confidential e.g., users’ faces or banking information, rather than publicly available images (see Carlini et al., 2018 for other ways to extract secrets). We limit the number of filters in conv1 to 32 to encourage models to learn a low-dimensional manifold capturing the data-generating distribution, but not the actual data. Although the model still has roughly $3 \times$ more parameters than CIFAR-10 has examples, it has been shown that we can
obtain parameter-independent capacity control with suitable norm-constraints for shallow models (Neyshabur et al., 2015).

Fooling image samples crafted with the model in Table 1 are shown in Figure 8. We sample uniformly from \(0.5 + U(0, 0.1) \in \mathbb{R}^{32 \times 32 \times 3}\) and do 50 iterations of gradient descent with a step size of 5e-3 on the loss using each class as the target label for a given trial.

(a)

(b)

Figure 8: Fooling images for each of the CIFAR-10 classes using (a) CNN and (b) CNN-L2. Noise is gradually transformed to resemble training data, but fine detail is omitted. Examples in (a) are classified with 32–63% confidence, and 41–81% for (b). Best viewed in colour.

To make the fooling images test more quantitative, we adopt the procedure suggested by Goodfellow (2015), of reporting the rate at which FGSM succeeds in causing the model to hallucinate a target class. This test is typically not reported in other works as robustness in this regime remains elusive, and machine learning models are thought to have a “thin manifold of accuracy” due to their linear behaviour which makes them overly confident when extrapolating (Goodfellow et al., 2015). We consider the attack a success if the argmax logit matches the target label, or equivalently, if the target softmax probability exceeds \(1/10\). This is conservative vs. choosing the threshold 0.5 per Goodfellow (2015). We sweep \(\epsilon\) over the range 0–255, and plot in Figure 9 the attack success rate (ASR), with the difference between the two largest softmax probabilities the margin (M). For each \(\epsilon\), ASR is an average over 1000 samples from \(U(0, 1) \in \mathbb{R}^{32 \times 32 \times 3}\), then taking a step in the direction that maximizes each of the 10 classes.

ASR peaks at around \(\epsilon = 20\) for CNN-L2, and then decays almost exponentially with increasing \(\epsilon\) as we move away from the manifold. Meanwhile, the softmax layer is close to uniform and the prediction margin remains small but slightly increasing. If we use 0.5 as the requisite prediction threshold, then the ASR is 0% for \(\epsilon = 16\) where it is nearly maximal in Figure 9 for the argmax case. WRN here is the “Public” model from the Madry et al. CIFAR-10 challenge. The ASR for WRN plateaus after an initial upward spike, as it is very non-linear, therefore a linear approximation of the loss gets “stuck”. Despite this, the WRN is more than twice as confident for all \(\epsilon\) than the CNN. It was previously shown that taking multiple steps results in maximum confidence predictions for the WRN, despite the resulting examples maintaining a uniform noise appearance (Galloway et al., 2018).

6.2 \(L_\infty\) Attacks

In Figure 10 we compare our CNNs to the “Public” and “Secret” pre-trained models that currently hold the benchmark for robustness to \(L_\infty\) attacks in the white-box and black-box setting respectively. These two models arise from independent runs of the same procedure, therefore their curves were almost identical over the full range of \(\epsilon\) for gradient sign attacks. We average the two series and denote “WRN” to reduce clutter. A baseline “WRN-Nat” trained by Madry et al. without PGD is included for context. As in their Figure 2, we sweep over \(\epsilon\) for white-box attacks, but continue on to \(\epsilon_{\text{max}} = 128\) until all predictions were reduced to chance.

We do not equate accuracy to robustness for white-box attacks, since the original adversarial examples problem also considers the saliency of perturbations. Figure 8 depicts how gradient-based perturbations with our models tend to be highly recognizable and structured for targeted attacks. We expect the Madry et al. models to generally have higher accuracy for the \(L_\infty\) attacks characterized in Figure 10, since they allocate significant capacity for

https://github.com/MadryLab/cifar10_challenge
Figure 10: White-box $L_\infty$ optimized gradient-based attacks. The cyan-fill area is where CNN-L2 achieves higher accuracy than WRN (the Madry et al. defense), and the orange area is vice versa. The “PGD” attack is the $(20–2–\epsilon)$ variant using the cross entropy loss and random start. FGSM is only shown for context to compare the marginal improvement of the PGD attack.

this express purpose. However, it is interesting to note that our small models eventually do overtake the them for $\epsilon > 40$. It is also encouraging that the slopes remain similar between “CNN-L2 PGD” and “Madry et al. PGD”, with the offset staying roughly constant in the small $\epsilon$ regime. Neither model fails as catastrophically as the undefended WRN-Nat.

We provide curves for FGSM not as a meaningful attack, but to show that there was little marginal benefit of performing multiple attack iterations for CNN-L2, compared to WRN. This shows that CNN-L2 is more linear locally near natural data manifold—as well as globally with fooling images—than the WRN.

6.3 $L_0$ Attacks

The $L_0$ counting “norm” was previously proposed as a natural metric for measuring a model’s ability to resist perturbations (Papernot et al., 2016a). We now evaluate on two held out $L_0$ attacks: random pixel swapping, and the Jacobian-based Saliency Map Attack (JSMA) (Papernot et al., 2016b) as additional benchmarks. For both of these attacks, each RGB colour channel was perturbed separately.

6.3.1 Pixel Swapping

Randomly swapping pixels can be interpreted as a weak black-box attack that controls for the perturbations’ saliency. This type of perturbation may be less detectable than JSMA by some statistical methods, as it does not affect the data distribution globally, however most detection methods do not work well (Carlini and Wagner 2017a; Athalye et al. 2018; Smith and Gal 2018).

Figure 11: An $L_0$ counting “norm” black-box attack in which we randomly swap pixels, and R,G,B are considered separately. The cyan-fill area is where CNN-L2 is more robust than the Madry et al. defense, and the orange area is vice versa.

In Figure 11 we plot accuracy for four different models, with each series averaged over three runs. Despite the attack being random, variance was very low even when not setting the random seed. CNN-L2 overtakes the WRN beyond $\approx 11\%$ of features swapped, and WRN-Nat at $\approx 5\%$. For context, Figure 12 shows sample images with decreasing signal-to-noise (we did not normalize to an actual SNR). The noise is fairly easily detected by human vision when presented next to the original, but the overall image remains interpretable.

Figure 12: CIFAR-10 samples with ratio of pixels randomly swapped annotated (top) and original images (bottom). At 26% of pixels swapped, CNN-L2 achieves 63.1% accuracy, WRN–31.7%, and WRN-Nat–15.4%.

6.3.2 Jacobian-based Saliency Map Attack

Gradient-based attacks tuned for the $L_0$ metric have also been proposed, such as JSMA which uses a greedy iterative procedure that constructs a saliency map, and sets pairs of pixels to the full-scale value until the attack either succeeds, or some maximum number of pixels are modified. This maximum is usually expressed as a ratio of total pixels, as in Section 6.3.1.

In Table 2 we report JSMA targeted success rates when
Table 2: JSMA success rate (ASR), avg. rate of perturbed features overall (Pert. Rte), and when attack was successful (S Pert. Rte) for a number of source images N. Maximum ratio of features that may be perturbed denoted by $\gamma$ (as a percentage). For each source image, a targeted attack is performed using all non-true labels as the target.

| N   | $\gamma$ | ASR | Pert. Rte | S Pert. Rte |
|-----|----------|-----|-----------|-------------|
| 10  | 1%       | 33% | 0.8%      | 0.19%       |
|     | 2%       | 58% | 1.3%      | 0.51%       |
|     | 3%       | 81% | 1.5%      | 0.99%       |
|     | 4%       | 86% | 1.7%      | 1.14%       |
|     | 5%       | 89% | 1.8%      | 1.28%       |
| 100 | 1%       | 28.4% | 0.8% | 0.14% |

The adversary may perturb between 1–5% of features, denoted by $\gamma$. CNN-L2 limits the adversary to lower attack success rates for all $\gamma$ in this range, even when considering its lesser clean test accuracy relative to WRN, as we do not skip examples that are incorrectly classified originally.

6.4 Discussion

It can be observed in Figure 11 that $L_2$ weight decay had a beneficial effect in defending against the $L_0$ attack for all $\epsilon$, with the benefits being most pronounced when the ratio of pixels swapped was less than 0.2, or between 0.4 and 0.9.

We omitted “CNN” in Figure 10 to keep the figure clear, and offer a comparison between with “CNN-L2” here instead. For the PGD (20–2–$\epsilon$) attack, the curves have similar slope for $\epsilon < 16$, with that of CNN being translated downward relative to CNN-L2. The difference was usually between 3–6% for this region, with the greatest difference of 6.8% at $\epsilon = 5$. Beyond $\epsilon = 16$, CNN was better, but accuracy was already quite low at this point around $\approx 15\%$. We found that the more a model underfit w.r.t. clean test accuracy (e.g. by optimizing with SGD instead of Adam, or binarizing hidden representations) the more gradual the roll-off when attacking with PGD, and earlier the cross-over point vs. the WRN.

Visualizing the filters of CNN-L2 in Figure 13 further speaks to the generality of this model. Given the fundamentally limited diversity and finite number of examples in CIFAR-10, it is desirable to learn a distributed representation comprising simple image primitives such as edge and colour detectors. These types of filters emerge naturally as a result of training a small architecture for a long time (e.g. 250 epochs), however models trained with weight decay (Figure 13(b)) have smoother filters, with some unnecessary ones becoming exactly zero.

The random swap experiment is sufficient to demonstrate that our models are significantly more robust to at least one attack defined w.r.t. the $L_0$ metric. We believe it is unlikely that the Madry et al. “Public” or “Secret” models will perform in a more compelling manner than “CNN-L2” over a similar range for any $L_0$ attack, but we invite others to attempt to show this by attacking our pre-trained model. Note that no specific action was taken to confer performance for $L_0$ attacks during training, other than attempting to tune a model to generalize well.

Models trained on PGD adversarial examples artificially increase the margin between class centroids by translating them with respect to the loss function, as PGD is an equivalent $L_\infty$ analogue to FGSM for deeper models. This reduces data separability by $\approx \epsilon \| w \|_1$ when the signed gradient is used, which encourages the model to fit some dimensions of low-variance in the training data to fully separate the classes. This results in the highly textured effect seen in Figure 13(a), supporting previous observations that CNNs have a tendency to latch on to surface statistics to fulfill the ERM optimization objective, resulting in a narrow generalization (Jo and Bengio, 2017).

The tendency of adversarial training to encourage fitting a specific metric, to a given magnitude, is even more pronounced with the harsh cliff in Madry et al. (their Figures 2 a and b) for MNIST. We omit an MNIST comparison for brevity, but refer readers to examples of targeted

![Figure 13: Visualizing filters from layer conv1 for the model outlined in Table 1. Filters arising from training without weight decay in (a) have a highly textured effect, while unimportant filters go to zero in (b) with overall $L_2$ weight decay. Both models were trained with the same random seed, hence why similar filters occur at the same coordinates in the grid. Best viewed in colour.](image-url)

(a) (b)
hand-draw attacks in a mobile application.

7 Conclusion

As we may not anticipate how our systems are likely to be compromised, a reasonable choice is to generalize to the best of our ability given finite data. Adversarial training has a limited “sweet-spot” in which it is effective at increasing robustness to some attacks, whereas weight decay reduces all generalization errors and encourages underfitting in the worst-case. Adversarial training is effective when the perturbation size is small, but remains challenging to apply without exposing other vulnerabilities. By significantly reducing the hypothesis space, our models strike compelling trade-offs when subject to unseen attacks. This provides a more complete baseline for future robustness comparisons which may involve more sophisticated learning objectives, and regularization techniques that do not degenerate with depth. Incorporating prior expert knowledge and interpretability measures seem promising for limiting the impact of adversarial examples, given sparsely sampled datasets in high-dimensions. We suggest future works sweep over the full perturbation range for several unseen attacks before drawing robustness conclusions.

Acknowledgments

The authors wish to acknowledge the financial support of NSERC, CFI, CIFAR and EPSRC. The authors also acknowledge hardware support from NVIDIA and Compute Canada. We thank Colin Brennan, Brittany Reiche, and Lewis Griffin for helpful edits and suggestions that improved the clarity of our manuscript.

References

M. Alzantot, Y. Sharma, S. Chakraborty, and M. Srivastava. GenAttack: Practical Black-box Attacks with Gradient-Free Optimization. arXiv:1805.11090, 2018.

A. Athalye, N. Carlini, and D. Wagner. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. arXiv:1802.00420, 2018.

Y. Bengio, N. L. Roux, P. Vincent, O. Delalleau, and P. Marcotte. Convex neural networks. In Advances in Neural Information Processing Systems, pages 123–130, 2006.

B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In ECML-KDD. Springer, 2013.

N. Carlini and D. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In ACM Workshop on Artificial Intelligence and Security, pages 3–14, 2017a.

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

N. Carlini, C. Liu, J. Kos, . Erlingsson, and D. Song. The Secret Sharer: Measuring Unintended Neural Network Memorization & Extracting Secrets. arXiv:1802.08232, 2018.

M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier. Parseval networks: Improving robustness to adversarial examples. In International Conference on Machine Learning, 2017.

S. Dube. High dimensional spaces, deep learning and adversarial examples. arXiv:1801.00634, 2018.

A. Galloway, G. W. Taylor, and M. Moussa. Predicting adversarial examples with high confidence. arXiv:1802.04457, 2018.

I. Goodfellow. Adversarial examples. slide 41, 2015. URL http://www.iangoodfellow.com/slides/2015-08-09-adv.pdf.

I. Goodfellow, Y. Bengio, and A. Courville. Deep Learning. MIT Press, 2016. http://www.deeplearningbook.org.

I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. In International Conference on Machine Learning, 2013.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In International Conference on Learning Representations, 2015.

A. Ilyas. Circumventing the ensemble adversarial training defense. https://github.com/andrewlyas/ens-adv-train-attack, 2018.

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

J. Jo and Y. Bengio. Measuring the tendency of CNNs to learn surface statistical regularities. arXiv:1711.11561, 2017.

H. Kannan, A. Kurakin, and I. Goodfellow. Adversarial logit pairing. arXiv:1803.06373, 2018.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv:1412.6980, 2014.
A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

A. Kurakin, I. J. Goodfellow, and S. Bengio. Adversarial Machine Learning at Scale. In International Conference on Learning Representations, 2017.

Y. LeCun and C. Cortes. The mnist database of handwritten digits. http://yann.lecun.com/exdb/mnist/, 1998.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations, 2018.

T. Na, J. H. Ko, and S. Mukhopadhyay. Cascade adversarial machine learning regularized with a unified embedding. arXiv:1708.02582, 2017.

B. Neyshabur, R. Tomioka, and N. Srebro. Norm-based capacity control in neural networks. In Proceedings of The 28th Conference on Learning Theory, 2015.

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.

A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In Computer Vision and Pattern Recognition, 2015.

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

N. Papernot, P. D. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In IEEE European Symposium on Security and Privacy, pages 372–387, 2016b.

N. Papernot, N. Carlini, I. Goodfellow, R. Feinman, F. Faghri, A. Matyasko, K. Hambardzumyan, Y.-L. Juang, A. Kurakin, R. Sheatsley, A. Garg, and Y.-C. Lin. cleverhans v2.0.0: an adversarial machine learning library. arXiv:1610.00768, 2017.

G. Pereyra, G. Tucker, J. Chorowski, L. Kaiser, and G. E. Hinton. Regularizing neural networks by penalizing confident output distributions. In International Conference on Learning Representations (Workshop), 2017.

A. Rozsa, M. Günther, and T. E. Boul. Towards robust deep neural networks with BANG. arXiv:1612.00138, 2016.

S. Sabour, Y. Cao, F. Faghri, and D. J. Fleet. Adversarial Manipulation of Deep Representations. arXiv:1511.05122, 2015.

Y. Sharma and P.-Y. Chen. Attacking the Madry Defense Model with L1-based Adversarial Examples. arXiv:1710.10733, 2017.

A. Sinha, H. Namkoong, and J. Duchi. Certifiable distributional robustness with principled adversarial training. International Conference on Learning Representations, 2018.

L. Smith and Y. Gal. Understanding measures of uncertainty for adversarial example detection. arXiv:1803.08533, 2018.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In International Conference on Learning Representations, 2014.

C. Szegedy, V. Vanhoucke, 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, pages 2818–2826, 2016.

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

F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. In International Conference on Learning Representations, 2018.

Y. Tsuzuku, I. Sato, and M. Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. arXiv:1802.04034, 2018.

J. Uesato, B. O’Donoghue, A. v. d. Oord, and P. Kohli. Adversarial Risk and the Dangers of Evaluating Against Weak Attacks. arXiv:1802.05666, 2018.

L. Xie, J. Wang, Z. Wei, M. Wang, and Q. Tian. Disturbabl: Regularizing cnn on the loss layer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4753–4762, 2016.

S. Zagoruyko and N. Komodakis. Wide residual networks. In BMVC, 2016.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In International Conference on Learning Representations, 2017.