Exploring the Landscape of Spatial Robustness

Logan Engstrom [*1] **Brandon Tran** [*1] **Dimitris Tsipras** [*1] **Ludwig Schmidt** [1] **Aleksander Mądry** [1]

Abstract

The study of adversarial robustness has so far largely focused on perturbations bound in ℓ_p-norms. However, state-of-the-art models turn out to be also vulnerable to other, more natural classes of perturbations such as translations and rotations. In this work, we thoroughly investigate the vulnerability of neural network–based classifiers to rotations and translations. While data augmentation offers relatively small robustness, we use ideas from robust optimization and test-time input aggregation to significantly improve robustness. Finally we find that, in contrast to the ℓ_p-norm case, first-order methods cannot reliably find worst-case perturbations. This highlights spatial robustness as a fundamentally different setting requiring additional study.

1. Introduction

Neural networks are now widely embraced as dominant solutions in computer vision (Krizhevsky et al., 2012; He et al., 2016), speech recognition (Graves et al., 2013), and natural language processing (Collobert & Weston, 2008). While their accuracy scores often match (and sometimes go beyond) human-level performance on key benchmarks (He et al., 2015; Taigman et al., 2014), we still do not understand how robust neural networks are. A prominent issue in this context is the existence of so-called *adversarial examples*, i.e., inputs that are almost indistinguishable from natural data to a human but cause state-of-the-art classifiers to make incorrect predictions with high confidence (Szegedy et al., 2013; Goodfellow et al., 2014). This raises concerns about the use of neural networks in contexts where reliability, dependability, and security are important desiderata.

There is a long line of work on methods for constructing adversarial perturbations in various settings (Szegedy et al., 2013; Goodfellow et al., 2014; Kurakin et al., 2016a;b; Sharif et al., 2016; Moosavi-Dezfooli et al., 2016; Carlini & Wagner, 2016; Papernot et al., 2017; Madry et al., 2017; Athalye et al., 2017). However, these methods are quite sophisticated, and, since they often rely on having fine-tuned control over a large number of input pixels or audio samples, end up creating fairly contrived perturbations. As such, one may suspect that adversarial examples constitute a problem only in the presence of a truly malicious attacker and are unlikely to arise in more benign environments. However, recent work has shown that neural network–based vision classifiers are vulnerable to input images that have been *spatially transformed* through small rotations, translations, shearing, scaling, and other natural transformations (Fawzi & Frossard, 2015; Kanbak et al., 2017; Xiao et al., 2018; Tramèr & Boneh, 2017). The vulnerability of neural networks to such transformations raises a natural question:

How can we build classifiers robust to such naturally occurring transformations?

We address this question by first performing an in-depth study of neural network–based classifier robustness to two basic image transformations: translations and rotations. While these transformations appear natural to a human, we show that small rotations and translations *alone* can significantly degrade accuracy. These transformations are particularly relevant for computer vision applications since real-world objects do not always appear centered and can often be significantly rotated.

1.1. Our Methodology and Results

We start with standard, near state-of-the-art image classifiers for the MNIST (LeCun et al., 1998), CIFAR10 (Krizhevsky & Hinton, 2009), and ImageNet (Russakovsky et al., 2015) datasets. We find that small rotations and translations consistently and significantly degrade accuracy across these classifiers in a number of settings, as shown in Figure 1.

We then perform a thorough analysis comparing the abilities of various adversaries – first-order, random, and grid–based – to fool models with small rotations and translations. Our results suggest that classifiers are even more brittle than

[*] Equal contribution [1] EECS, MIT, Massachusetts, USA. Correspondence to: Logan Engstrom <engstrom@mit.edu>, Brandon Tran <btran115@mit.edu>, Dimitris Tsipras <tsipras@mit.edu>, Ludwig Schmidt <ludwigs@mit.edu>, Aleksander Mądry <madry@mit.edu>.

Proceedings of the 36^{th} *International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

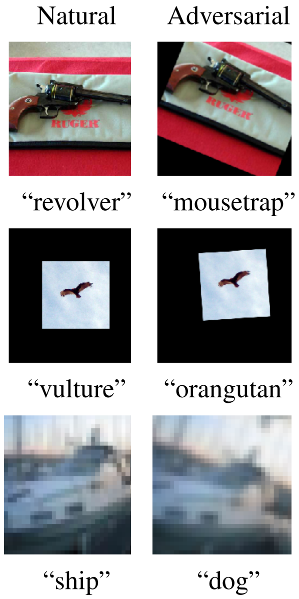

Figure 1. Examples of adversarial transformations and their predictions in the standard, "black canvas", and reflection padding settings.

previously believed, as we find that even small *random* transformations can degrade accuracy by up to 30%. More significantly, we find that grid adversaries are much more powerful than first-order adversaries. This is in stark contrast to results in the ℓ_p adversarial example literature, where first-order methods can consistently approximately worst-case inputs (Carlini & Wagner, 2016; Madry et al., 2017).

To understand why such a difference occurs, we delve deeper into the classifiers to try and understand the failure modes induced by such natural transformations. We find that the loss landscape of classifiers with respect to rotations and translations is nonconvex and contains many spurious maxima. This is in contrast to the ℓ_p setting, in which, experimentally, the maxima tend to concentrate well (Madry et al., 2017). Our experiments with the landscapes demonstrate that any adversary relying on first order information might be unable to reliably find misclassifications. Furthermore, for most images, the set of fooling rotations and translations is highly irregular and nonconvex.

Using insights from our study, we next examine methods for alleviating these vulnerabilities. As a natural baseline, we augment the training procedure with rotations and translations. While this does largely mitigate the problem on MNIST, additional data augmentation only marginally adds robustness on CIFAR10 and ImageNet. We thus propose two natural methods for further increasing the robustness of these models. These methods are based on robust optimization and aggregation of random input transformations. They offer significant improvements in classification accuracy against both adaptive and random attackers when compared to both standardly trained models and those trained with additional data augmentation.

Finally, we examine the interplay between rotations / translations and the widely used ℓ_∞-based adversarial examples. We observe that robustness to these two classes of input perturbations is largely orthogonal to each other. In particular, pixel-based robustness does not imply spatial robustness, while combining spatial and ℓ_∞-bounded transformations seems to have a *cumulative* effect in reducing classification accuracy. This emphasizes the need to broaden the notions of image similarity in the adversarial examples literature beyond the common ℓ_p-balls.

1.2. Summary of Contributions

We perform extensive experiments that provide a fine-grained understanding of rotation / translation robustness on a wide spectrum of datasets and training regimes. In summary, we show that:

- Grid based adversaries performing an exhaustive, fine-grained search are a strong adversary for fooling models with small rotations and translations. Suprisingly, unlike in the corresponding ℓ_p setting, first-order based adversaries fail to reliably find fooling inputs and are significantly less effective than grid-based adversaries. Furthermore, an attacker using random rotations and translations can still significantly degrade accuracy, suggesting that there might be a large degree of misclassification stemming from these vulnerabilities even in benign settings.

- The optimization landscape of loss with respect to rotations and translations is nonconvex and contains many spurious local maxima. This could explain the failure of first-order methods to find fooling transformations. Consequently, rigorous evaluation of model robustness in this spatial setting requires techniques that that go beyond what was needed to induce ℓ_p adversarial robustness.

- Additional data augmentation is not sufficient to significantly increase robustness to rotations and translations, even in the benign case. However, robustness can be significantly increased using ideas from robust optimization and test-time input transformations; on ImageNet, our best model attains a top1 accuracy of **56%** against the strongest adversary, versus **34%** for a standard network with additional data augmentation.

- Robustness to ℓ_∞-bounded perturbations does not significantly affect spatial robustness. Instead, these two notions appear orthogonal to each other.

2. Related Work

The fact that small rotations and translation can fool neural networks on MNIST and CIFAR10 was first observed in (Fawzi & Frossard, 2015). They compute the minimum transformation required to fool the model and use it as a measure for a quantitative comparison of different architectures and training procedures. The main difference to our work is that we focus on the optimization aspect of the problem. We show that a few random queries usually suffice for a successful attack, while first-order methods are ineffective. Moreover, we go beyond standard data augmentation and evaluate the effectiveness of natural baseline defenses.

The concurrent work of (Kanbak et al., 2017) proposes a different first-order method to evaluate the robustness of classifiers based on geodesic distances on a manifold. This metric is harder to interpret than our parametrized attack space. Moreover, given our findings on the non-concavity of the optimization landscape, it is unclear how close their method is to the ground truth (exhaustive enumeration). While they perform a limited study of defenses (adversarial fine-tuning) using their method, it appears to be less effective than our baseline worst-of-10 training. We attribute this difference to the inherent obstacles first-order methods face in this optimization landscape.

Recently, (Xiao et al., 2018) and (Tramèr & Boneh, 2017) observed independently that it is possible to use various spatial transformations to construct adversarial examples for naturally and adversarially trained models. The main difference from our work is that we show even very simple transformations (translations and rotations) are sufficient to break a variety of classifiers, while the transformations employed in (Xiao et al., 2018) and (Tramèr & Boneh, 2017) are more involved. The transformation in (Xiao et al., 2018) is based on performing a displacement of individual pixels in the original image constrained to be globally smooth and then optimized for misclassification probability. (Tramèr & Boneh, 2017) consider an ℓ_∞-bounded pixel-wise perturbation of a version of the original image that has been slightly rotated and in which a few random pixels have been flipped. Both of these methods require direct access to the attacked model (or a surrogate) to compute (or at least estimate) the gradient of the loss function with respect to the model's input. In contrast, our attacks can be implemented using only a small number of random, non-adaptive inputs.

3. Adversarial Rotations and Translations

Recall that in the context of image classification, an *adversarial example* for a given input image x and a classifier C is an image x' that satisfies two properties: (i) on the one hand, the adversarial example x' causes the classifier C to output a different label on x' than on x, i.e., we have $C(x) \neq C(x')$. (ii) On the other hand, the adversarial example x' is "visually similar" to x.

Clearly, the notion of visual similarity is not precisely defined here. In fact, providing a precise and rigorous definition is extraordinarily difficult as it would require formally capturing the notion of human perception. Consequently, previous work largely settled on the assumption that x' is a valid adversarial example for x if and only if $\|x - x'\|_p \leq \varepsilon$ for some $p \in [0, \infty]$ and ε small enough. This convention is based on the fact that two images are indeed visually similar when they are close enough in some ℓ_p norm. However, the converse is not necessarily true. A small rotation or translation of an image usually appears visually similar to a human, yet can lead to a large change when measured in an ℓ_p norm. We aim to expand the range of similarity measures considered in the adversarial examples literature by investigating robustness to small rotations and translations.

Attack methods. Our first goal is to develop sufficiently strong methods for generating adversarial rotations and translations. In the context of pixel-wise ℓ_p perturbations, the most successful approach for constructing adversarial examples so far has been to employ optimization methods on a suitable loss function (Szegedy et al., 2013; Goodfellow et al., 2014; Carlini & Wagner, 2016). Following this approach, we parametrize our attack method with a set of tunable parameters and then optimize over these parameters.

First, we define the exact range of attacks we want to optimize over. For the case of rotation and translation attacks, we wish to find parameters $(\delta u, \delta v, \theta)$ such that rotating the original image by θ degrees around the center and then translating it by $(\delta u, \delta v)$ pixels causes the classifier to make a wrong prediction. Formally, the pixel at position (u, v) is moved to the following position (assuming the point $(0, 0)$ is the center of the image):

$$\begin{bmatrix} u' \\ v' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \cdot \begin{bmatrix} u \\ v \end{bmatrix} + \begin{bmatrix} \delta u \\ \delta v \end{bmatrix}.$$

We implement this transformation in a differentiable manner using the spatial transformer blocks of (Jaderberg et al., 2015) [1]. In order to handle pixels that are mapped to non-integer coordinates, the transformer units include a differentiable bilinear interpolation routine. Since our loss function is differentiable with respect to the input and the transformation is in turn differentiable with respect to its parameters, we can obtain gradients of the model's loss function w.r.t. the perturbation parameters. This enables us to apply a first-order optimization method to our problem.

By defining the spatial transformation for some x as

[1] We used the open source implementation found here: `https://github.com/tensorflow/models/tree/master/research/transformer`.

$T(x; \delta u, \delta v, \theta)$, we construct an adversarial perturbation for x by solving the problem

$$\max_{\delta u, \delta v, \theta} \mathcal{L}(x', y), \quad \text{for } x' = T(x; \delta u, \delta v, \theta) \ , \tag{1}$$

where $\mathcal{L}$ is the loss function of the neural network[2], and y is the correct label for x.

We compute the perturbation from Equation 1 in three distinct ways:

- **First-Order Method (FO):** Starting from a random choice of parameters, we iteratively take steps in the direction of the gradient of the loss function. This is the direction that locally maximizes the loss of the classifier (as a surrogate for misclassification probability). Since the maximization problem we are optimizing is non-concave, there are no guarantees for global optimality, but the hope is that the local maximum solution closely approximates the global optimum. Note that unlike the ℓ_p-norm case, we are not optimizing in the pixel space but in the latent space of rotation and translation parameters.
- **Grid Search:** We discretize the parameter space and exhaustively examine every possible parametrization of the attack to find one that causes the classifier to give a wrong prediction (if such a parametrization exists). Since our parameter space is low-dimensional enough, this method is computationally feasible (in contrast to a grid search for ℓ_p-based adversaries).
- **Worst-of-k:** We randomly sample k different choices of attack parameters and choose the one on which the model performs worst. As we increase k, this attack interpolates between a random choice and grid search.

We remark that while a first-order attack requires full knowledge of the model to compute the gradient of the loss with respect to the input, the other two attacks do not. They only require the outputs corresponding to chosen inputs, which can be done with only query access to the target model.

4. Improving Invariance to Spatial Transformations

As we will see in Section 5, augmenting the training set with random rotations and translations does improve the robustness of the model against such random transformations. However, data augmentation does not significantly improve the robustness against worst-case attacks and sometimes leads to a drop in accuracy on unperturbed images. To address these issues, we explore two simple baselines that turn out to be surprisingly effective.

Robust Optimization. Instead of performing standard empirical risk minimization to train the classification model, we utilize ideas from robust optimization. Robust optimization has a rich history (Ben-Tal et al., 2009) and has recently been applied successfully in the context of defending neural networks against adversarial examples (Madry et al., 2017; Sinha et al., 2017; Raghunathan et al., 2018; Kolter & Wong, 2017). The main barrier to applying robust optimization for spatial transformations is the lack of an efficient procedure to compute the worst-case perturbation of a given example. Performing a grid search (as described in Section 3) is prohibitive as this would increase the training time by a factor close to the grid size, which can easily be a factor 100 or 1,000. Moreover, the non-convexity of the loss landscape prevents potentially more efficient first-order methods from discovering (approximately) worst-case transformations (see Section 5 for details).

Given that we cannot fully optimize over the space of translations and rotations, we instead use a coarse approximation provided by the worst-of-10 adversary (as described in Section 3). So each time we use an example during training, we first sample 10 transformations of the example uniformly at random from the space of allowed transformations. We then evaluate the model on each of these transformations and train on the one perturbation with the highest loss. This corresponds to approximately minimizing a min-max formulation of robust accuracy similar to (Madry et al., 2017). Training against such an adversary increases the overall time by a factor of roughly six.[3]

Aggregating Random Transformations. As Section 5 shows, the accuracy against a *random* transformation is significantly higher than the accuracy against the worst transformation in the allowed attack space. This motivates the following inference procedure: compute a (typically small) number of random transformations of the input image and output the label that occurs the most in the resulting set of predictions. We constrain these random transformations to be within 5% of the input image size in each translation direction and up to 15° of rotation. [4] The training procedure and model can remain unchanged while the inference time is increased by a small factor (equal to the number of transformations we evaluate on).

[2] The loss $\mathcal{L}$ of the classifier is a function from images to real numbers that expresses the performance of the network on the particular example x (e.g., the cross-entropy between predicted and correct distributions).

[3] We need to perform 10 forward passes and one backwards pass instead of one forward and one backward pass required for standard training.

[4] Note that if an adversary rotates an image by 30° (a valid attack in our threat model), we may end up evaluating the image on rotations of up to 45°.

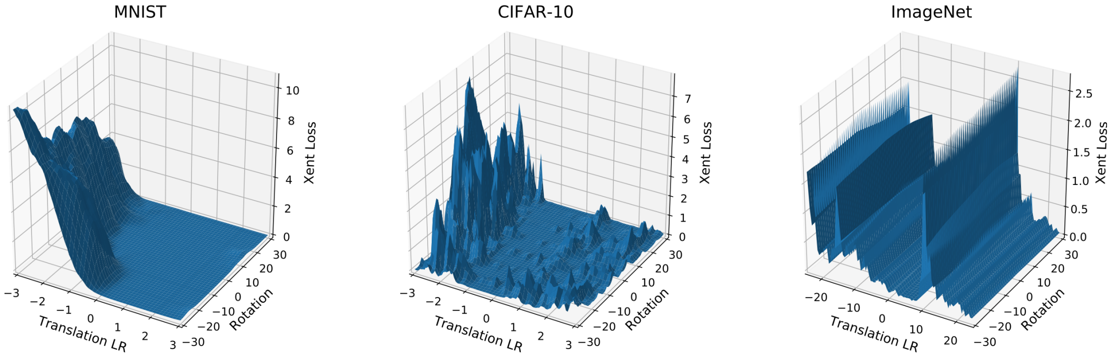

Figure 2. Loss landscape of a random example for each dataset when performing left-right translations and rotations. Translations and rotations are restricted to 10% of the image pixels and 30°, respectively. We observe that the landscape is significantly non-concave, rendering first-order methods to generate adversarial example ineffective. Figure 12 in the appendix shows additional examples.

Combining Both Methods. The two methods outlined above are orthogonal and in some sense complementary. We can therefore combine robust training (using a worst-of-k adversary) and majority inference to further increase the robustness of our models.

5. Experiments

We evaluate standard image classifiers for the MNIST (LeCun et al., 1998), CIFAR10 (Krizhevsky & Hinton, 2009) and ImageNet (Russakovsky et al., 2015) datasets. In order to determine the extent to which misclassification is caused by insufficient data augmentation during training, we examine various data augmentation methods. We begin with a description of our experimental setup.

Model Architecture. For MNIST, we use a convolutional neural network derived from the TensorFlow Tutorial (tft). In order to obtain a fully convolutional version of the network, we replace the fully-connected layer by two convolutional layers with 128 and 256 filters each, followed by a global average pooling. For CIFAR10, we consider a standard ResNet (He et al., 2016) model with 4 groups of residual layers with filter sizes [16, 16, 32, 64] and 5 residual units each. We use standard and ℓ_∞-adversarially trained models similar to those studied by (Madry et al., 2017).[5,6] For ImageNet, we use a ResNet-50 (He et al., 2016) architecture implemented in the `tensorpack` repository (Wu et al., 2016). We did not modify the model architectures or training procedures.

Attack Space. In order to maintain the visual similarity of images to the natural ones we restrict the space of allowed perturbations to be relatively small. We consider rotations of at most 30° and translations of at most (roughly) 10% percent of the image size in each direction. This corresponds to 3 pixels for MNIST (image size 28×28) and CIFAR10 (image size 32×32), and 24 pixels for ImageNet (image size 299×299). For grid search attacks, we consider 5 values per translation direction and 31 values for rotations, equally spaced. For first-order attacks, we use 200 steps of projected gradient descent of step size 0.01 times the parameter range. When rotating and translating the images, we fill the empty space with zeros (black pixels).

Data Augmentation. We consider five variants of training for our models.

- Standard training: The standard training procedure for the respective model architecture.
- ℓ_∞-bounded adversarial training: The classifier is trained on ℓ_∞-bounded adversarial examples that are generated with projected gradient descent.
- No random cropping: Standard training for CIFAR-10 and ImageNet includes data augmentation via random crops. We investigate the effect of this data augmentation scheme by also training a model without random crops.
- Random rotations and translations: At each training step, we perform a uniformly random perturbation from the attack space on each training example.
- Random rotations and translations from larger intervals: As before, we perform uniformly random perturbations, but now from a *superset* of the attack space (40°, ± 13% pixels).

[5] `https://github.com/MadryLab/cifar10_challenge`

[6] `https://github.com/MadryLab/mnist_challenge`

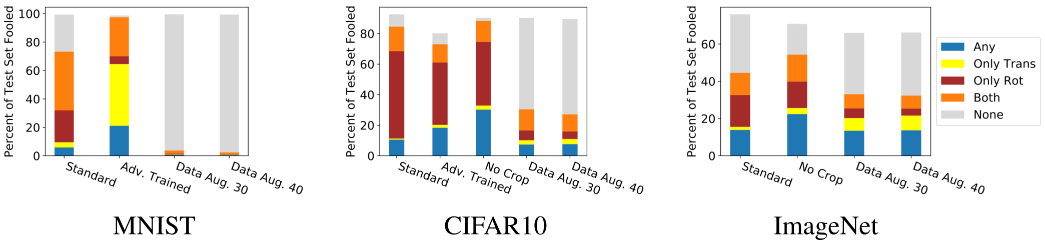

Figure 3. Fine-grained dataset analysis. For each model, we visualize what percent of the test set can be fooled via various methods. We compute how many examples can be fooled with either translations or rotations ("any"), how many can be fooled only by one of these, and how many require a combination to be fooled ("both").

5.1. Evaluating Model Robustness

We evaluate all models against random and grid search adversaries with rotations and translations considered both separately and together. We report the results in Table 1. We visualize a random subset of successful attacks in Figures 5, 6, and 7 of Appendix A.

Despite the high accuracy of standard models on unperturbed examples and their reasonable performance on random perturbations, a grid search can significantly lower the classifiers' accuracy on the test set. For the standard models, accuracy drops from 99% to 26% on MNIST, 93% to 3% on CIFAR10, and 76% to 31% on ImageNet (Top 1 accuracy).

The addition of random rotations and translations during training greatly improves both the random and adversarial accuracy of the classifier for MNIST and CIFAR10, but less so for ImageNet. For the first two datasets, data augmentation increases the accuracy against a grid adversary by 60% to 70%, while the same data augmentation technique adds less than 3% accuracy on ImageNet.

We perform a fine-grained investigation of our findings:

- In Figure 3 we examine how many examples can be fooled by (i) rotations only, (ii) translations only, (iii) neither transformation, or (iv) both.
- We visualize the set of fooling angles for a random sample of the rotations-only grid in Figure 4 on ImageNet, and provide more examples in the appendix in Figure 10. We observe that the set of fooling angles is nonconvex and not contiguous.
- To investigate how many transformations are adversarial per image, we analyze the percentage of misclassified grid points for each example in Figure 11. While the majority of images has only a small number of adversarial transformations, a significant fraction of images is fooled by 20% or more of the transformations.

Padding Experiments. A natural question is whether the reduced accuracy of the models is due to the cropping applied during the transformation. We verify that this is not the case by applying zero and reflection padding to the image datasets. We note that the zero padding creates a "black canvas" version of the dataset, ensuring that no information from the original image is lost after a transformation. We show a random set of adversarial examples in this setting in Figure 8 and a full evaluation in Table 4. We also provide more details regarding reflection padding in Section B and provide an evaluation in Table 6. All of these are in Appendix A.

5.2. Comparing Attack Methods

In Table 2 we compare different attack methods on various classifiers and datasets. We observe that worst-of-10 is a powerful adversary despite its limited interaction with the target classifier. The first-order adversary performs significantly worse. It fails to approximate the ground-truth accuracy of the models and performs significantly worse than the grid adversary and even the worst-of-10 adversary.

Understanding the Failure of First-Order Methods. The fact that first-order methods fail to reliably find adversarial rotations and translations is in sharp contrast to previous work on ℓ_p-robustness (Carlini & Wagner, 2016; Madry et al., 2017). For ℓ_p-bounded perturbations parametrized directly in pixel space, prior work found the optimization landscape to be well-behaved which allowed first-order methods to consistently find maxima with high loss. In the case of spatial perturbations, we observe that the non-concavity of the problem is a significant barrier for first-order methods. We investigate this issue by visualizing the loss landscape. For a few random examples from the three datasets, we plot the cross-entropy loss of the examples as a function of translation and rotation. Figure 2 shows one example for each dataset and additional examples are visualized in Figure 12 of the appendix. The plots show that the loss landscape is indeed non-concave and contains many local maxima of low value. The low-dimensional problem structure seems

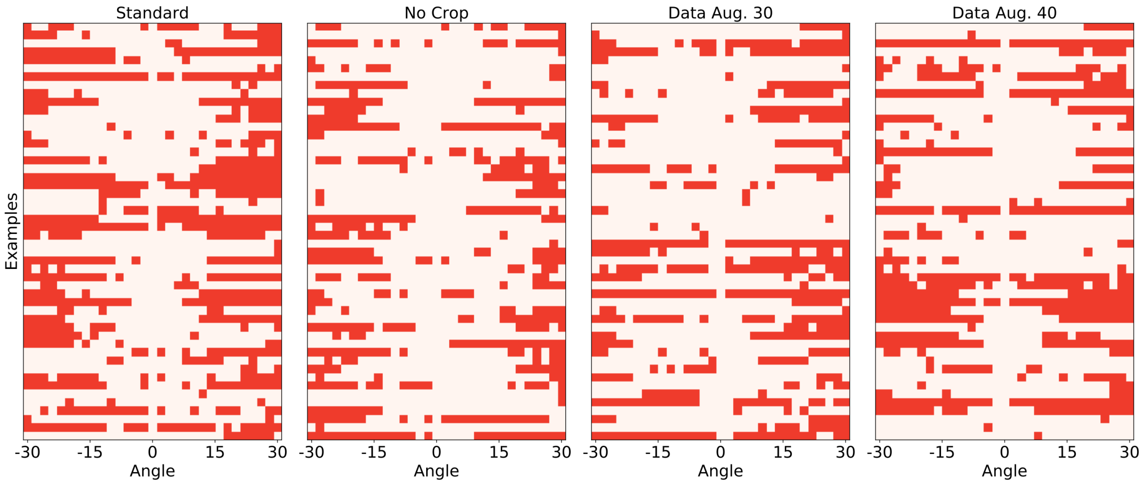

Figure 4. Visualizing which angles fool ImageNet classifiers for 50 random examples. For each dataset and model, we visualize one example per row. Red corresponds to *misclassification* of the images. We observe that the angles fooling the models form a highly non-convex set. Figure 10 in the appendix shows additional examples for CIFAR10 and MNIST.

Table 1. Accuracy of different classifiers against rotation and translation adversaries on MNIST, CIFAR10, and ImageNet. The allowed transformations are translations by (roughly) 10% of the image size and ±30° rotations. The attack parameters are chosen through random sampling or grid search with rotations and translations considered both together ("Rand.", "Grid") and separately ("Rand. T." and "Grid T." for transformations, "Rand R." and "Grid R." for rotations). We consider networks that are trained with (i) the respective standard setup, (ii) no data augmentation (if data augmentation is present in standard setup), (iii) with an ℓ_∞ adversary, (iv) with data augmentation corresponding to the attack space (±3px, ±30°) and an enlarged space (±4px, ±40°), and (v) with worst-of-10 training for both types of augmentations.

	Model	Nat.	Rand.	Grid	Rand. T.	Grid T.	Rand. R.	Grid R.
MNIST	Standard	99.31%	94.23%	**26.02%**	98.61%	89.80%	95.68%	70.98%
	ℓ_∞-Adv	98.65%	88.02%	**1.20%**	93.72%	34.13%	95.27%	72.03%
	Aug. 30	99.53%	99.35%	**95.79%**	99.47%	98.66%	99.34%	98.23%
	Aug. 40	99.34%	99.31%	**96.95%**	99.39%	98.65%	99.40%	98.49%
	W-10 (30)	99.48%	99.37%	**97.32%**	99.50%	99.01%	99.39%	98.62%
	W-10 (40)	99.42%	99.39%	**97.88%**	99.45%	98.89%	99.36%	98.85%
CIFAR10	Standard	92.62%	60.93%	**2.80%**	88.54%	66.17%	75.36%	24.71%
	No Crop	90.34%	54.64%	**1.86%**	81.95%	46.07%	69.23%	18.34%
	ℓ_∞-Adv	80.21%	58.33%	**6.02%**	78.15%	59.02%	62.85%	20.98%
	Aug. 30	90.02%	90.92%	**58.90%**	91.76%	79.01%	91.14%	76.33%
	Aug. 40	88.83%	91.18%	**61.69%**	91.53%	77.42%	91.10%	76.80%
	W-10 (30)	91.34%	92.35%	**69.17%**	92.43%	83.01%	92.33%	81.82%
	W-10 (40)	91.00%	92.11%	**71.15%**	92.28%	82.15%	92.53%	82.25%
ImageNet	Standard	75.96%	63.39%	**31.42%**	73.24%	60.42%	67.90%	44.98%
	No Crop	70.81%	59.09%	**16.52%**	66.75%	45.17%	62.78%	34.17%
	Aug. 30	65.96%	68.60%	**32.90%**	70.27%	45.72%	69.28%	47.25%
	Aug. 40	66.19%	67.58%	**33.86%**	69.50%	44.60%	68.88%	48.72%
	W-10 (30)	76.14%	73.19%	**52.76%**	74.42%	61.18%	73.74%	61.06%
	W-10 (40)	74.64%	71.36%	**50.23%**	72.86%	59.34%	71.95%	59.23%

Table 2. Comparison of attack methods across datasets and models. Worst-of-10 is very effective and significantly reduces the model accuracy despite the limited interaction. The first-order (FO) adversary performs poorly, despite the large number of steps allowed. We compare standard training to Augmentation ($\pm$3px, $\pm 30^\circ$). For the full table, see Figure 3 of Appendix A.

	MNIST		CIFAR-10		ImageNet	
	Standard	Aug.	Standard	Aug.	Standard	Aug.
Natural	99.31%	99.53%	92.62%	90.02%	75.96%	65.96%
Worst-of-10	73.32%	98.33%	20.13%	79.92%	47.83%	50.62%
First-Order	79.84%	98.78%	62.69%	85.92%	63.12%	66.05%
Grid	**26.02%**	**95.79%**	**2.80%**	**58.92%**	**31.42%**	**32.90%**

to make non-concavity a crucial obstacle. Even for MNIST, where we observe fewer local maxima, the large flat regions prevent first-order methods from finding transformations of high loss.

Relation to Black-Box Attacks. Given its limited interaction with the model, the worst-of-10 adversary achieves a significant reduction in classification accuracy. It performs only 10 *random*, *non-adaptive* queries to the model and is still able to find adversarial examples for a large fraction of the inputs (see Table 2). The low query complexity is an important baseline for black-box attacks on neural networks, which recently gained significant interest (Papernot et al., 2017; Chen et al., 2017; Bhagoji et al., 2017; Ilyas et al., 2017). Black-box attacks rely only function evaluations of the target classifier, without additional information such as gradients. The main challenge is to construct an adversarial example from a small number of queries. Our results show that it is possible to find adversarial rotations and translations for a significant fraction of inputs with very few queries.

Combining Spatial and ℓ_∞-Bounded Perturbations Table 1 shows that models trained to be robust to ℓ_∞ perturbations do not achieve higher robustness to spatial perturbations. This provides evidence that the two families of perturbation are orthogonal to each other. We further investigate this possibility by considering a combined adversary that utilizes ℓ_∞ bounded perturbations on top of rotations and translations. The results are shown in Figure 13. We indeed observe that these combined attacks reduce classification accuracy in an (approximately) additive manner.

5.3. Evaluating Our Defense Methods.

As we see in Table 1, training with a worst-of-10 adversary significantly increases the spatial robustness of the model, also compared to data augmentation with random transformations. We conjecture that using more reliable methods to compute the worst-case transformations will further improve these results. Unfortunately, increasing the number of random transformations per training example quickly becomes computationally expensive. And as pointed out above, current first-order methods also appear to be insufficient for finding worst-case transformations efficiently.

Our results for majority-based inference are presented in Table 5 of Appendix A. By combining these two defenses, we improve the worst-case performance of the models from 26% to 98% on MNIST, from 3% to 82% on CIFAR10, and from 31% to 56% on ImageNet (Top 1).

6. Conclusions

We examined the robustness of state-of-the-art image classifiers to translations and rotations. We observed that even a small number of randomly chosen perturbations of the input are sufficient to considerably degrade the classifier's performance.

The fact that common neural networks are vulnerable to simple and naturally occurring spatial transformations (and that these transformations can be found easily from just a few random tries) indicates that adversarial robustness should be a concern not only in a fully worst-case security setting. We conjecture that additional techniques need to be incorporated in the architecture and training procedures of modern classifiers to achieve worst-case spatial robustness. Also, our results underline the need to consider broader notions of similarity than only pixel-wise distances when studying adversarial misclassification attacks. In particular, we view combining the pixel-wise distances with rotations and translations as a next step towards the "right" notion of similarity in the context of images.

Acknowledgements

Dimitris Tsipras was supported in part by the NSF grant CCF-1553428 and the NSF Frontier grant CNS-1413920. Aleksander Mądry was supported in part by an Alfred P. Sloan Research Fellowship, a Google Research Award, and the NSF grants CCF-1553428 and CNS-1815221.

References

TensorFlow tutorial: Deep MNIST for experts. URL https://www.tensorflow.org/versions/r0.12/tutorials/mnist/pros/.

Athalye, A., Engstrom, L., Ilyas, A., and Kwok, K. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*, 2017.

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. *Robust Optimization*. Princeton University Press, 2009.

Bhagoji, A. N., He, W., Li, B., and Song, D. Exploring the space of black-box attacks on deep neural networks. *arXiv preprint arXiv:1712.09491*, 2017.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. *arXiv preprint arXiv:1608.04644*, 2016.

Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C.-J. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 15–26. ACM, 2017.

Collobert, R. and Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 160–167. ACM, 2008.

Fawzi, A. and Frossard, P. Manitest: Are classifiers really invariant? In *British Machine Vision Conference (BMVC)*, 2015.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Graves, A., Mohamed, A.-r., and Hinton, G. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pp. 6645–6649. IEEE, 2013.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

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

Jaderberg, M., Simonyan, K., Zisserman, A., et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pp. 2017–2025, 2015.

Kanbak, C., Moosavi-Dezfooli, S.-M., and Frossard, P. Geometric robustness of deep networks: analysis and improvement. *arXiv preprint arXiv:1711.09115*, 2017.

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

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. 2009.

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

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

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

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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

Moosavi-Dezfooli, S., Fawzi, A., and Frossard, P. Deepfool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2574–2582, 2016.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pp. 506–519. ACM, 2017.

Raghunathan, A., Steinhardt, J., and Liang, P. Certified defenses against adversarial examples. *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Bys4ob-Rb.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of*

Computer Vision (IJCV), 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Sharif, M., Bhagavatula, S., Bauer, L., and Reiter, M. K. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, pp. 1528–1540, 2016.

Sinha, A., Namkoong, H., and Duchi, J. Certifiable distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

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

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1701–1708, 2014.

Tramèr, F. and Boneh, D. Personal communication, 2017.

Wu, Y. et al. Tensorpack. https://github.com/tensorpack/, 2016.

Xiao, C., Zhu, J.-Y., Li, B., He, W., Liu, M., and Song, D. Spatially transformed adversarial examples. *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HyydRMZC-.