A ROTATION AND A TRANSLATION SUFFICE: FOOLING CNNS WITH SIMPLE TRANSFORMATIONS

Logan Engstrom, Ludwig Schmidt, Dimitris Tsipras, Aleksander Mądry
Massachusetts Institute of Technology
{engstrom,ludwigs,tsipras,madry}@mit.edu

ABSTRACT

Recent work has shown that deep neural networks are vulnerable to imperceptible, adversarial perturbations of their input. Subsequent progress in the area of adversarial robustness has led to reliable defense mechanisms against concrete classes of adversaries. While these developments constituted an important step, the overall picture of adversarial robustness turns out to be far from complete. Specifically, we provide evidence that precisely defining the set of attacks we wish classifiers to be robust against is a crucial issue. To this end, we show that very natural transformations such as rotations and translations *alone*, can be used to completely fool image classification models, even when the latter are robust against the canonical ℓ_∞-bounded adversary. This emphasizes that deciding on a specific class of allowed input perturbations is a critical design choice. It can have significant impact and requires further exploration and understanding.

1 INTRODUCTION

Deep neural network have become widely used for a variety of tasks in computer vision [14, 11], speech recognition [9], and text analysis [6]. However, machine learning systems cannot be part of security-critical applications until their performance is reliable and well-understood. A significant obstacle, in this context, is the existence of *adversarial examples* [17, 8], i.e., inputs that are almost indistinguishable from natural data for a human, yet cause state-of-the-art classifiers to give wrong predictions with high confidence. The existence of such inputs, as well as the fact that they can be easily and reliably produced, raises important security concerns about the reliability of neural networks. Additionally, it contradicts the folklore belief that deep neural networks achieve truly human-level performance on some tasks.

There is now a long line of work on developing defenses against adversarial examples [16, 20, 7, 19, 10]. However, most of the proposed mechanisms were subsequently shown to be vulnerable to more sophisticated attacks [3, 12, 5]. Recently, it has been demonstrated [15, 2] that adversarial training against a sufficiently strong adversary increases the robustness of neural network classifiers to concrete families of adversaries. This demonstrates that once the adversary is restricted to a well-specified class of attacks, one can develop principled defenses against them. Inevitably, as long as the right notion of input similarity is unclear, it is essentially impossible to argue that a model is truly secure (even for a well-defined test set of benign inputs).

A major challenge in this context is thus to precisely define what it means for an adversarial input to be visually similar to a benign input. In the context of image classification, the main notions of similarity studied so far are pixel-based distances with respect to an ℓ_p-norm. If two images are close in an ℓ_p-norm, they are usually visually similar as well. However, the contrapositive does not always hold. There exist simple image transformations that lead to a modified image that is far from the original image in all ℓ_p-norms, yet arguably very similar to it from a human perspective. This raises a fundamental question:

What is a more comprehensive notion of visual similarity for adversarial examples?

In this work, we address this question by studying two fundamental image transformations: translations and rotations. While appearing natural to a human, we show that these transformations can still cause neural networks to misclassify the resulting images, *without any ℓ_p-bounded perturbation*.

Our starting point is the MNIST model of Madry et al. [15], which was shown to be robust against a range of adversaries constrained to small perturbations in ℓ_∞- or ℓ_2-norm. We show that small rotations and translations *alone* can cause that model to perform wrong predictions on the entirety of the test set. We then investigate this phenomenon further by examining networks trained on random rotations and translations. We also consider the performance of an adversary that utilizes simple transformations combined with small ℓ_∞-bounded perturbations. In summary, we show that:

- A simple attack based purely on rotations and translations is very effective against neural network classifiers, even when they have been trained against ℓ_p-bounded adversaries.
- Augmenting the dataset with random transformations *does* increase its robustness, but does *not* completely alleviate the issue.
- Combining rotations and translations with small ℓ_∞-bounded changes provides the adversary with disproportionately more power than when using only one of these attacks.

2 Adversarial Rotations and Translations

Formally, a classification problem consists of an input distribution with labeled examples (in the case of MNIST, images of handwritten digits and their true label). A classifier C is a function that receives an input x and predicts a label $C(x)$. Assuming that the true label of x is y, we consider x' to be an *adversarial example* for x if $C(x) = y$, $C(x') \neq y$, and x' is "visually similar" to x. The precise notion of visual similarity is notoriously hard to define, and most previous work considered x and x' close if $\|x - x'\|_p \leq \varepsilon$ for some $p \in [1, \infty]$ and ε small enough.

Our goal is to examine whether these notions of similarity are complete from the point of view of adversarial training. That is, if a classifier is trained to be robust against adversaries constrained in some ℓ_p-norm, is the classifier also robust against other natural transformations of the input images? Here, we will focus on the simplest spatial transformations, namely translations and rotations.

By far the most successful approach for constructing adversarial examples has been the use of an optimization method on a suitable loss function [17, 8, 4]. Following this approach, we parametrize our attack method with a set of tunable parameters and then optimize over these parameters. We perform this optimization in two distinct ways:

- **Projected Gradient Descent (PGD):** Starting from a random choice of parameters, we iteratively take local steps in the direction that maximizes the loss of the classifier (as a surrogate for misclassification probability). 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 first discretize the parameter space and then exhaustively examine every possible parametrization of the attack to find a parametrization that causes the classifier to give a wrong prediction (if such a parametrization exists). Since our parameter space is small enough, this method is computationally feasible (in contrast to a grid search for ℓ_p-based adversaries).

It remains to define the class of attacks we 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 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 [13], which 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 apply a first-order optimization method to (locally) optimize for the worst-case transformation of the input. By defining the spatial transformation for some x as $T(x; \delta u, \delta v, \theta)$, we construct an adversarial perturbation for x by solving the problem

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

$\mathcal{L}$ is the loss function of the neural network[1], and y is the correct label for example x. Since this is a non-concave maximization problem, there are no guarantees for the quality of our solution.

3 Experiments

We evaluate the effectiveness of our attack by applying it on the MNIST models of Madry et al. [15]. They provide both a naturally trained model, as well as a model that has been adversarially trained against an ℓ_∞-bounded adversary with $\varepsilon = 0.3$.[2] The adversarially trained model has been shown to be highly robust ($\sim 89\%$ accuracy) against a wide range of white-box attacks. We attack the model without the ℓ_∞-norm constraint while ensuring that our attacks are visually similar to natural inputs.

In order to ensure that the perturbed images are not heavily distorted, we restrict our rotations and translations to a maximum of 30 degrees and 3 pixels per dimension, respectively. We visualize a random subset of successful attacks in Figure 2 of Appendix A. Our attacks are generated by solving the optimization problem of Equation 1 using 200 steps of PGD with a step size of 0.01. In order to reliably compare against the best possible attack, we perform an exhaustive grid search with a spacing of 1 pixel for translations and 100 equally distanced values in the $[-30, 30]$ degree interval. The results of our experiments are shown in the first two rows of Table 1. We also report the accuracy of the classifiers when the perturbations are chosen uniformly at random from the allowed interval.

	Natural	Random	PGD	Grid Search
Naturally Trained	99.17%	81.88%	63.78%	0.01%
Adversarially Trained	98.40%	81.33%	73.61%	0.00%
Model A (±3, ±30)	99.28%	99.09%	96.97%	86.21%
Model B (±4, ±40)	99.12%	99.10%	96.98%	92.18%

Table 1: Accuracy of different models against adversaries that only use rotations and translations. The allowed transformations are ±3 pixels translation and ±30 degrees rotation. The attack parameters are chosen through random sampling, projected gradient descent, or grid search (we also include the original accuracy). The models are the naturally and adversarially trained models of Madry et al. (first two rows) and the natural model trained using data augmentations with random rotations and translations (±3 pixels, ±30 degrees for model A and ±4 pixels, ±40 degrees for model B).

Despite the high accuracy of the models on natural examples and their reasonable performance on random perturbations, a grid search adversary can find an adversarial examples for (almost) every image in the test set for both classifiers. Our first order attack, despite being successful for a considerable number of examples (better than random) is not close to the ground truth. We conjecture that this is due to the problem being so low-dimensional that the non-concavity of the objective is a significant barrier. This is in contrast to the observations of [15] for pixel-based optimization where PGD is able to reliable find (almost) worst-case examples. We plan to further investigate this issue using more powerful first order adversaries.

3.1 Data Augmentation

Since the network was not trained on rotations and translations of the training set, a natural question is whether this is the sole reason for its poor performance. We explore this issue by training two additional networks using data augmentation for rotations and translations. For each training example, we translate it by a random value in each direction and rotate it by another random value before training the classifier on it. In the following, let model A be a model trained with translations and rotations chosen uniformly at random from $[-3, 3]$ pixels and $[-30, 30]$ degrees respectively. Similarly, let model B be a model trained with translations in $[-4, 4]$ and rotations in $[-40, 40]$. We use the same set of attacks as before (±3, ±30). Note that model B has been trained on transformations of larger magnitude than those that appear during the evaluation. We repeat our attack experiments

[1] 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 examples x (e.g., the cross-entropy between predicted and correct distributions).

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

on these two networks and report our results in the last two rows of Table 1. We show some of the successful perturbations in Figure 3 of Appendix A. The addition of rotations and translations greatly improves both the random and adversarial accuracy of the classifier. Despite its good performance on these examples, we are still able to find transformations that fool the network on a sizeable portion of the test set (especially considering the simplicity of the task).

3.2 Combination with an ℓ_∞-bounded Adversary

Motivated by the fact that data augmentation provides a significant boost in the robustness of the classifier, we added an ℓ_∞-based adversarial training procedure on top of it. We found that even for small values of ε (say 0.1), the network did not learn any meaningful classifier. This behavior indicates that the ℓ_∞-bounded adversary is significantly more powerful when combined with random rotations and translations. We investigate this phenomenon further by examining how different networks perform against ℓ_∞-bounded adversaries of different ε in the following regimes: standard ℓ_∞-bounded attack, random rotation and translation followed by the ℓ_∞-bounded attack, grid search for the worst translation and rotation example, followed by ℓ_∞-bounded perturbations. The results are shown in Figure 1.

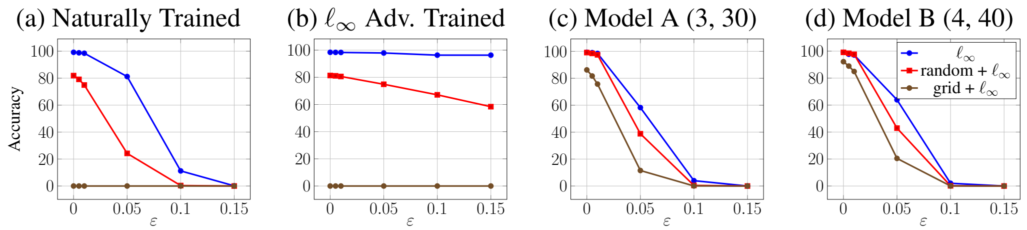
(a) Naturally Trained
(b) ℓ_∞ Adv. Trained
(c) Model A (3, 30)
(d) Model B (4, 40)

Figure 1: Accuracy of different classifiers against ℓ_∞ adversaries with various values of ε and spatial transformations. For each value of ε, we perform PGD to find the most adversarial ℓ_∞ bounded perturbations (ℓ_∞ plot). Additionally, we combine PGD with random rotations and translations and with a grid search over rotations and translations in order to find the transformation that combines with PGD in the most adversarial way.

On the ℓ_∞-based trained network, ℓ_∞-bounded perturbations cause a 2.1% drop in accuracy, while random translations and rotations drop of 20%. However when combined, these attacks cause a drop of 40% which is disproportionately high. For models A and B (trained with data augmentation), random translations and rotations do not cause a significant drop in accuracy ℓ_∞-bounded adversary. Again, when combined with an $\varepsilon = 0.05$ ℓ_∞-bounded attack, they cause an additional 20% of the test set to be classified incorrectly.

4 Related Work

Recently, [1] and [18] observed independently that it is possible to use spatial transformations to construct adversarial examples for naturally and adversarially trained models. The main difference to our work is that we show that very simple transformations (translations and rotations) are already sufficient to break a variety of classifiers, while [1] and [18] needed additional transformations.

5 Conclusions

We examined the effectiveness of adversarial translations and rotations on naturally trained networks and on networks that were trained against an ℓ_∞-bounded adversary. We conclude that these transformations *alone* can completely fool such networks, providing evidence that training against ℓ_∞-bounded adversaries is not sufficient to ensure correct classification of examples derived from benign transformations of natural inputs. Additionally, we observe that augmenting the dataset with random translations and rotations provides a significant boost in the classification performance, but does not yield full robustness. Finally, we demonstrate that rotations and translations can significantly amplify the power of an ℓ_∞-bounded adversary. Overall, our findings suggest that rotation and translation transformations need to be incorporated into the design of robust image classifiers.

References

[1] Anonymous. Spatially transformed adversarial examples. *In submission to International Conference on Learning Representations*, 2018.

[2] Nicholas Carlini, Guy Katz, Clark Barrett, and David L Dill. Ground-truth adversarial examples. *arXiv preprint arXiv:1709.10207*, 2017.

[3] Nicholas Carlini and David Wagner. Defensive distillation is not robust to adversarial examples. *arXiv preprint arXiv:1607.04311*, 2016.

[4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. *arXiv preprint arXiv:1608.04644*, 2016.

[5] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. *arXiv preprint arXiv:1705.07263*, 2017.

[6] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[7] Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.

[8] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[9] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.

[10] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[12] Warren He, James Wei, Xinyun Chen, Nicholas Carlini, and Dawn Song. Adversarial example defenses: Ensembles of weak defenses are not strong. *arXiv preprint arXiv:1706.04701*, 2017.

[13] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[15] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[16] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 582–597, 2016.

[17] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[18] Florian Tramèr and Dan Boneh. Personal communication, 2017.

[19] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[20] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

A Omitted Figures

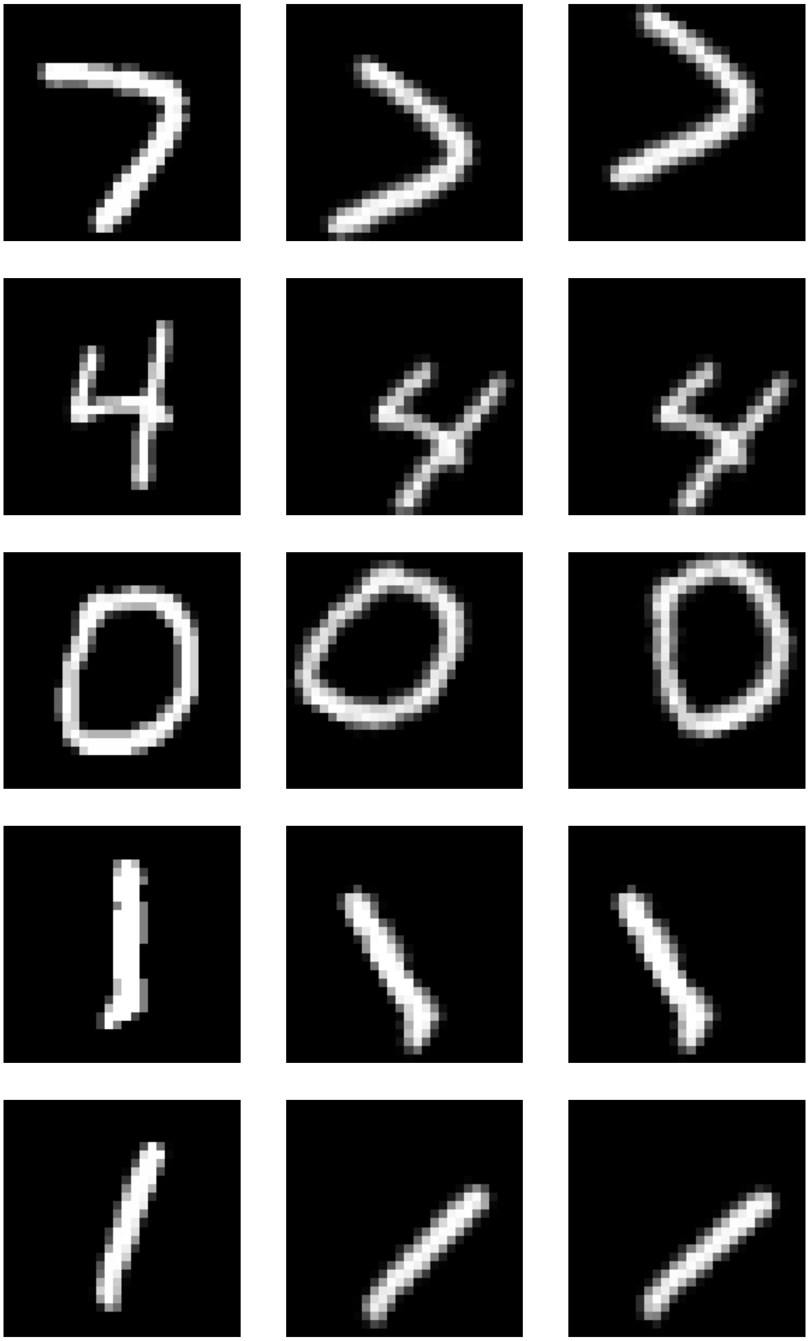

Figure 2: Successful adversarial examples for the models of Madry et al. using only rotations and translations. The first column corresponds to the natural (original example), while the next two columns correspond to the perturbed variant that fools the naturally and adversarially trained networks respectively.

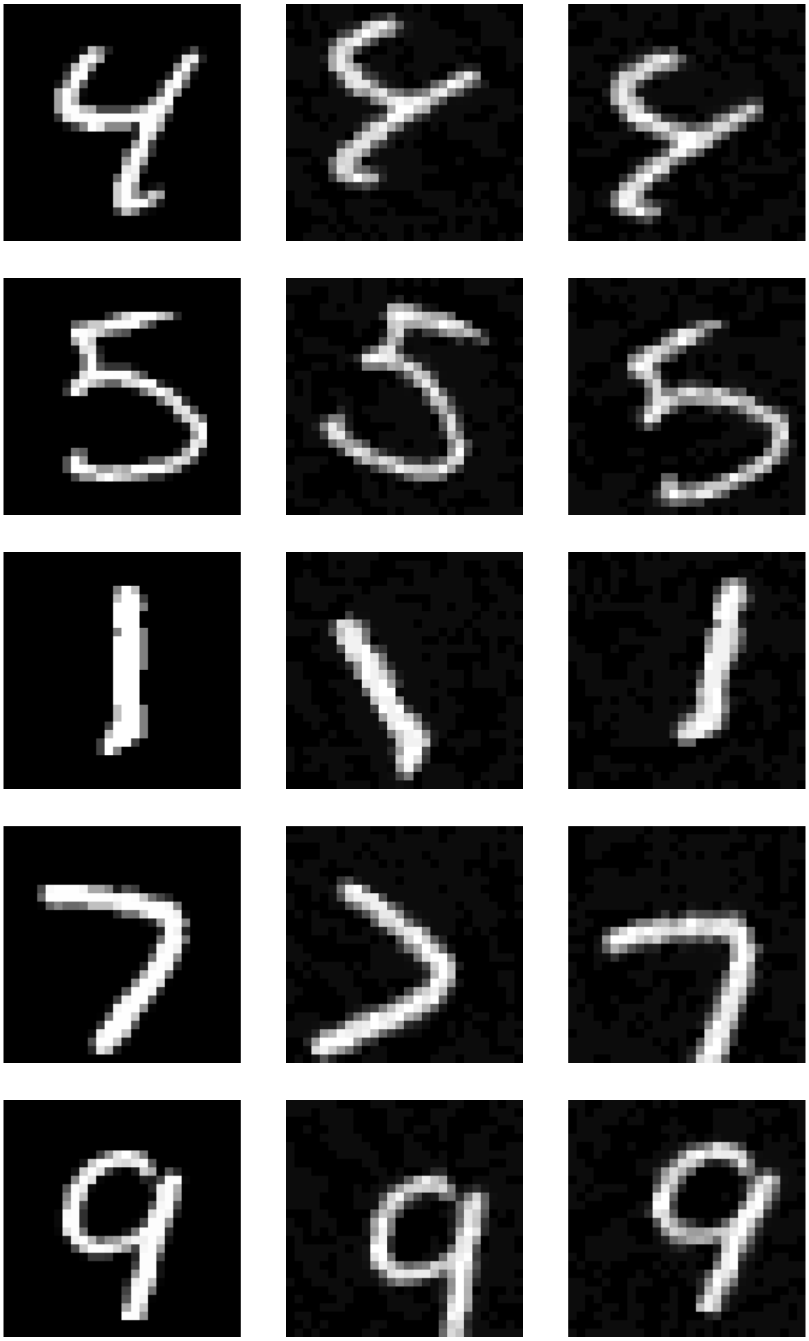

Figure 3: Successful adversarial examples for models A and B (trained with random translations/rotations of $\pm 3, \pm 30$ and $\pm 4, \pm 40$ respectively). The examples were generated by translating the image by at most 3 pixels and rotating it by at most 30 degrees, followed by a change of at most 0.05 to each pixel. The first column corresponds to the natural (original example), while the next two columns correspond to the perturbed variant that fools model A and B respectively. We plot the frequency of attack parameters in Figure 4.

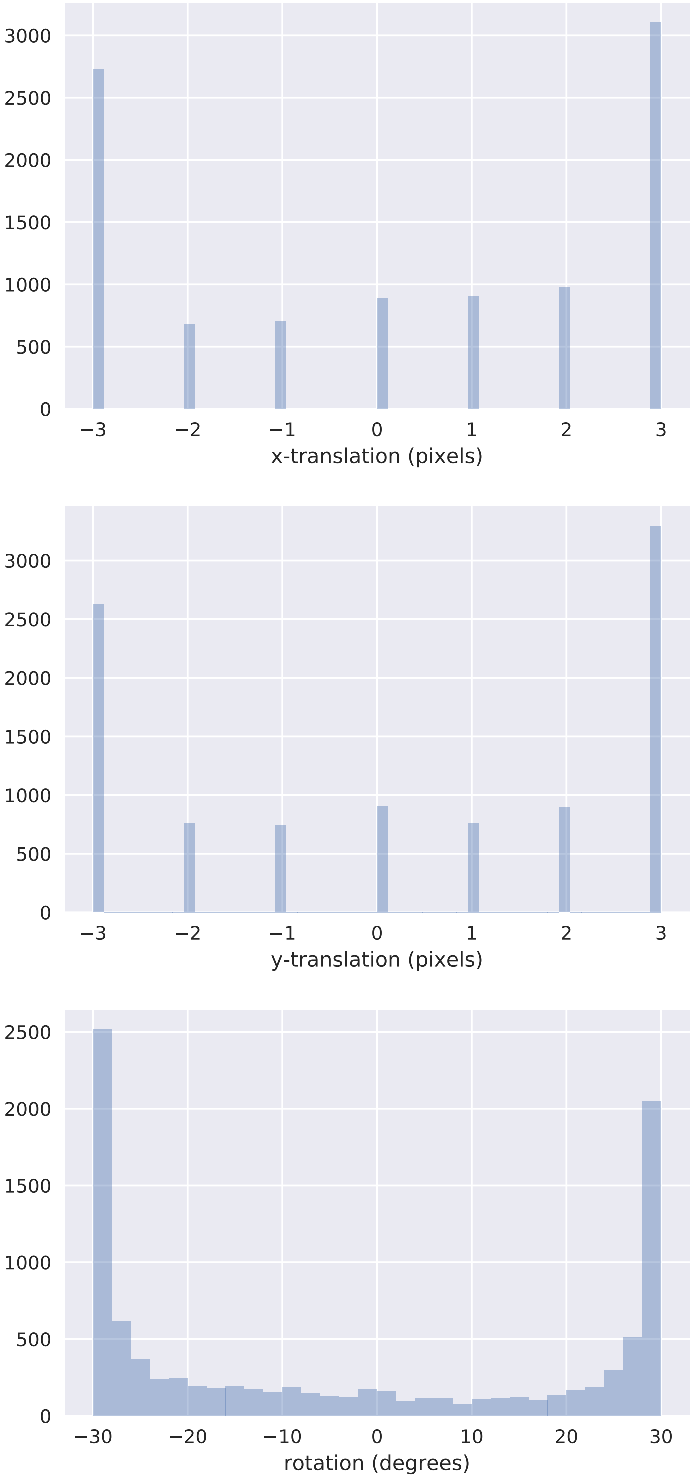

Figure 4: Histograms of the parameters chosen by grid search for the ℓ_∞ adversary combined with translations and rotations attacking Model B $(\pm 4, \pm 40)$. We observe that while most attacks choose extremal values this is not the case for every example.