Disentangling Adversarial Robustness and Generalization

David Stutz [1] **Matthias Hein** [2] **Bernt Schiele** [1]

Abstract

In an effort to clarify the relationship between robustness and generalization, we assume an underlying, low-dimensional data manifold and show: 1. regular adversarial examples leave the manifold; 2. adversarial examples constrained to the manifold, i.e., on-manifold adversarial examples, exist; 3. on-manifold adversarial examples are generalization errors; 4. regular robustness and generalization are not necessarily contradicting goals. These assumptions imply that *both* robust *and* accurate models are possible. We confirm our claims through extensive experiments on synthetic data (with known manifold) as well as on EMNIST and Fashion-MNIST. This is a short version of our CVPR'19 work (Stutz et al., 2019).

1. Introduction

Adversarial robustness describes a model's resilience to adversarial examples, imperceptibly perturbed images causing mis-classification. While many defenses against these attacks have been proposed – some of which have been shown to be ineffective (Carlini & Wagner, 2017; 2016; Athalye & Carlini, 2018; Athalye et al., 2018) – the problem of adversarial robustness is still poorly understood, even for simple datasets such as EMNIST (Cohen et al., 2017) and Fashion-MNIST (Xiao et al., 2017). Thus, the phenomenon of adversarial examples itself, i.e., their existence, has received considerable attention. Early explanations (Szegedy et al., 2013; Goodfellow et al., 2014) have recently been superseded by the manifold assumption (Gilmer et al., 2018; Tanay & Griffin, 2016; Song et al., 2018a): adversarial examples are assumed to leave the underlying, low-dimensional but usually unknown data manifold. Yet, on a simplistic toy dataset, Gilmer et al. (2018) also found adversarial examples on the manifold, as also tried on real datasets (Song et al., 2018b; Brown et al., 2017; Zhao et al., 2018), rendering the manifold assumption questionable.

[1]Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken [2]University of Tübingen, Tübingen. Correspondence to: David Stutz <david.stutz@mpi-inf.mpg.de>.

Presented at the ICML 2019 Workshop on Uncertainty and Robustness in Deep Learning.

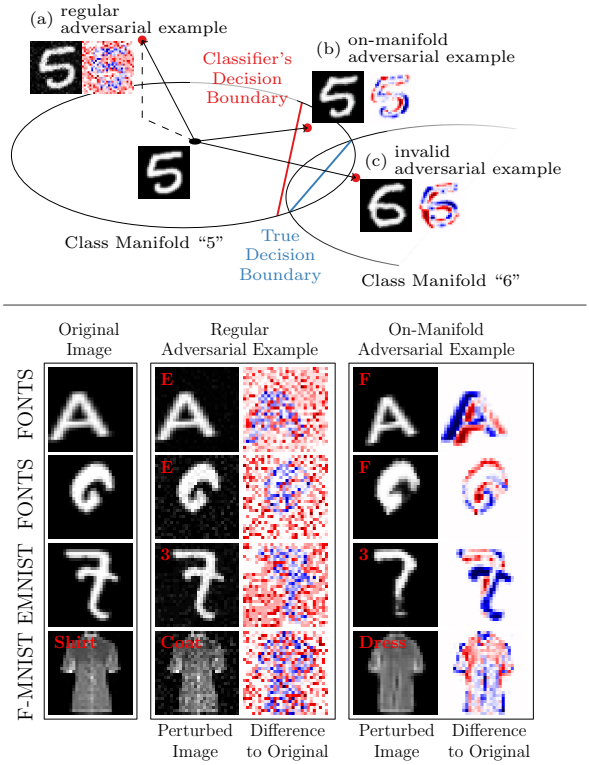

Figure 1. Adversarial examples (and normalized difference to the original image) in the context of the underlying manifold, e.g., class manifolds "5" and "6" on EMNIST (Cohen et al., 2017). Regular adversarial examples, *not* constrained to the manifold (a) result in (seemingly) random noise patterns. Adversarial examples constrained to the manifold (b) result in meaningful manipulations of the image content; however, care needs to be taken that the actual, true label wrt. the manifold does not change (c).

Similarly, the relation between robustness and generalization is of interest. Recently, it has been argued (Tsipras et al., 2018; Su et al., 2018) that there exists an inherent trade-off, i.e., robust and accurate models seem impossible. However, these findings have to be questioned given the results in (Gilmer et al., 2018; Rozsa et al., 2016) showing the opposite, i.e., better generalization helps robustness. In order to address this controversy, we consider adversarial robustness in the context of the underlying manifold: we explicitly ask whether adversarial examples leave, or stay on, the manifold. On EMNIST, for example, considering the class manifolds for "5" and "6", as illustrated in Fig. 1, *regular adversarial examples* are not guaranteed to lie on

the manifold, cf. Fig. 1 (a). Adversarial examples can, however, also be constrained to the manifold, referred to as *on-manifold adversarial examples*, cf. Fig. 1 (b); in this case, it is important to ensure that the adversarial examples do not actually change their label, i.e., are more likely to be a "6" than a "5", as in Fig. 1 (c).

Contributions: Based on this distinction between regular robustness and on-manifold robustness we show:

1. regular adversarial examples leave the manifold;
2. adversarial examples constrained to the manifold, i.e., on-manifold adversarial examples, exist and can be computed using an approximation of the manifold;
3. on-manifold robustness is essentially generalization;
4. and regular robustness and generalization are not necessarily contradicting goals, i.e., for any arbitrary but fixed model, better generalization through additional training data does not worsen robustness.

We conclude that both robust and accurate models are possible and can, e.g., be obtained through adversarial training on larger training sets. Additionally, we propose on-manifold adversarial training to boost generalization in settings where the manifold is known, can be approximated, or invariances of the data are known. We present experimental results on a novel MNIST-like, synthetic dataset with known manifold, as well as on EMNIST (Cohen et al., 2017) and Fashion-MNIST (Xiao et al., 2017).

This paper is a short version of our work presented at CVPR'19; while this paper is self-contained, we refer to (Stutz et al., 2019) and its supplementary material for further details and experimental results.

2. Disentangling Adversarial Robustness and Generalization

Datasets: We use EMNIST (Cohen et al., 2017) and F(ashion)-MNIST (Xiao et al., 2017) and learn class-specific VAE-GANs (Larsen et al., 2016; Rosca et al., 2017) to approximate the underlying manifold. Our synthetic dataset, FONTS, consists of 28×28 images of the letters "A" to "J" of 1000 Google Fonts uniformly transformed over translation, shear, scale and rotation using a (differentiable) spatial transformer network (Jaderberg et al., 2015). The manifold, i.e., transformation parameters, font and class, is known.

Networks: We consider classifiers with three convolutional layers (4×4 kernels; stride 2; 16, 32, 64 channels), followed by ReLU activations and batch normalization (Ioffe & Szegedy, 2015), and two fully connected layers. To control their generalization, we use $250 \leq N \leq 40$k training images; for each N, we train 5 models with random weight initialization (Glorot & Bengio, 2010) and report averages.

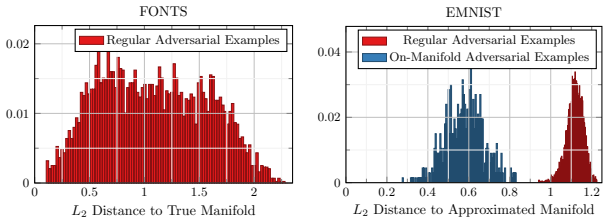

Figure 2. Distance of adversarial examples to the true, on FONTS (left), or approximated, on EMNIST (right), manifold. We show normalized histograms of the L_2 distance of adversarial examples to their projections onto the manifold. Regular adversarial examples exhibit a significant distance to the manifold; on EMNIST, clearly distinguishable from on-manifold adversarial examples.

Attack: Given an image-label pair (x, y) from an unknown data distribution p and a classifier f, an adversarial example is a perturbed image $\tilde{x} = x + \delta$ which is mis-classified by the model, i.e., $f(\tilde{x}) \neq y$. We concentrate on the L_∞ white-box attack by Madry et al. (2018) that directly maximizes the cross-entropy loss, i.e. $\max_\delta \mathcal{L}(f(x+\delta), y)$, such that $\|\delta\|_\infty \leq \epsilon$ and $\tilde{x}_i \in [0, 1]$ using projected gradient descent. We use 40 iterations and consider 5 restarts, uniformly sampled in the ϵ-ball for $\epsilon = 0.3$; we attack 1000 test images.

Adversarial Training: An established defense is adversarial training, i.e., training on adversarial examples crafted during training (Madry et al., 2018). We follow common practice and train on 50% clean images and 50% adversarial examples (Szegedy et al., 2013). For $\epsilon = 0.3$, the attack is run for full 40 iterations, i.e., is not stopped at the first adversarial example found. Robustness of the obtained network is measured using attack **success rate**, i.e., the fraction of successful attacks on correctly classified test images; lower success rate indicates higher robustness of the network.

2.1. Adversarial Examples Leave the Manifold

On EMNIST, where particular background pixels are known to be constant, an adversarial example manipulating these pixels has zero probability under the data distribution; thus, the distance to its projection onto the manifold has to be non-zero. On FONTS, with known generative process in the form of a decoder, the projection can be obtained iteratively. On EMNIST, in contrast, the manifold is approximated using 50 nearest neighbors; the projection can be computed through least squares. Fig. 2 (left) shows that regular adversarial examples clearly exhibit non-zero distance to the manifold on FONTS. In fact, the projections of these adversarial examples to the manifold are almost always the original test images: the distance to the manifold is essentially the norm of the corresponding perturbation. This suggests that the adversarial examples leave the manifold in an almost orthogonal direction. On EMNIST, in Fig. 2 (right), these results can be confirmed in spite of the crude local

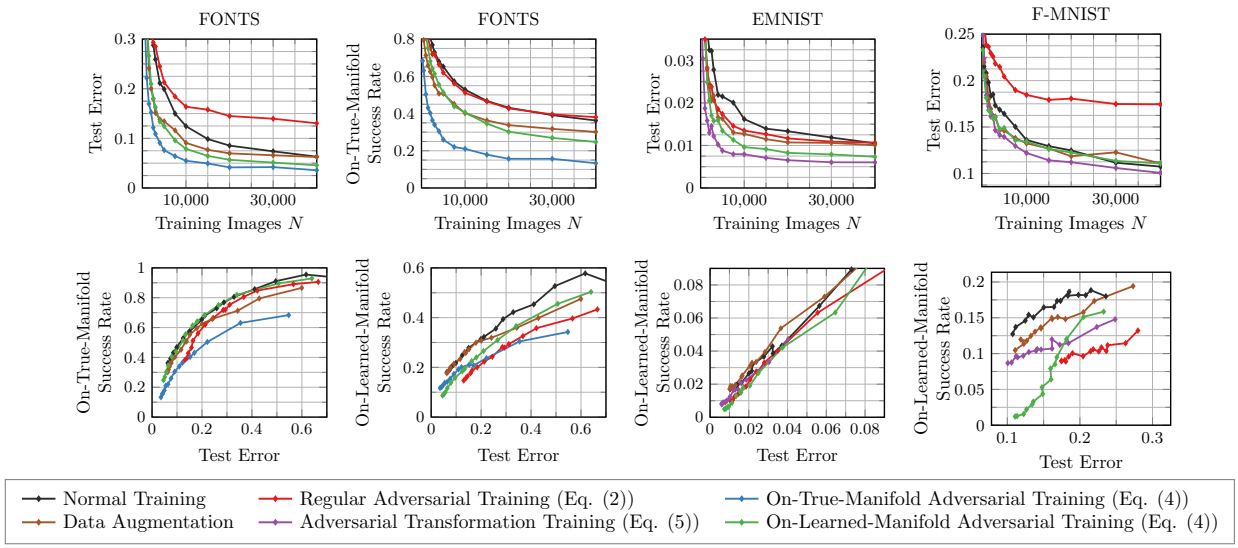

Figure 3. On-manifold robustness is strongly related to generalization, as shown on FONTS, EMNIST and F-MNIST considering on-manifold success rate and test error. Top: Generalization and on-manifold success rate in relation to the number of training examples. Bottom: On-manifold success rate plotted against test error.

approximation of the manifold. This shows that adversarial examples essentially *are* off-manifold adversarial examples; this is intuitive as for well-trained classifiers, leaving the manifold should be the "easiest" way to fool it.

2.2. On-Manifold Adversarial Examples

Given that regular adversarial examples leave the manifold, we explicitly compute adversarial examples constrained to the manifold. Here, we assume our data distribution $p(x, y)$ to be conditional on latent variables z, i.e., $p(x, y|z)$, corresponding to the underlying, low-dimensional manifold. On FONTS, we know the (class-conditional) distributions $p(z|x, y)$ and $p(x|z, y)$ by construction; on EMNIST and F-MNIST, we obtain approximations using VAE-GANs (Larsen et al., 2016; Rosca et al., 2017). Then, given encoder enc and decoder dec with $z = \text{enc}(x)$, we solve $\max_\zeta \mathcal{L}(f(\text{dec}(z + \zeta)), y)$ such that $\|\zeta\|_\infty \leq \eta$; the image-constraint, i.e., $\text{dec}(z + \zeta) \in [0, 1]$, is enforced by the decoder and the η-constraint can, again, be enforced by projection. Label invariance is ensured by considering only class-specific encoders and decoders. We use $\eta = 0.3$ and the same optimization procedure as for regular adversarial examples; on approximated manifolds, the perturbation $z+\zeta$ is additionally constrained to $[-2, 2]^{10}$, corresponding to a truncated normal prior from the class-specific VAE-GANs; we attack 2500 test images.

Fig. 1 (bottom) shows on-manifold adversarial examples for all datasets. On FONTS, using the true, known class manifolds, on-manifold adversarial examples clearly reflect the transformations of the latent space (1st row). For the learned class manifolds, the perturbations are less pronounced, often manipulating boldness or details of the characters (2nd row). On EMNIST and F-MNIST, on-manifold adversarial examples represent meaningful manipulations, such as removing the horizontal line of the hand-drawn "7" (3rd row) or removing the collar and buttons of a shirt (4th row). Finally, Fig. 2 (right) shows that on-manifold adversarial examples are closer to the manifold than regular adversarial examples. Finally, we note that these on-manifold adversarial examples are similar to those crafted in (Gilmer et al., 2018; Schott et al., 2018; Athalye et al., 2018). However, we directly compute the perturbation ζ on the manifold instead of computing the perturbation δ in the image space and subsequently projecting $x + \delta$ to the manifold.

2.3. On-Manifold Robustness is Essentially Generalization

We argue that on-manifold robustness is nothing different than generalization: as on-manifold adversarial examples have non-zero probability under the data distribution, they are merely generalization errors. This is shown in Fig. 3 where test error and on-manifold success rate are shown. On FONTS and EMNIST, better generalization, i.e., using more training images N, also reduces on-manifold success rate. On F-MNIST, the relationship is less pronounced because on-manifold adversarial examples, computed using our VAE-GANs, are not close enough to real generalization errors. However, even on F-MNIST, there is a clear relationship between on-manifold robustness and generalization.

To exploit this relationship between on-manifold robustness and generalization, we perform on-manifold adversarial training, i.e., training $50\%/50$ on clean examples and

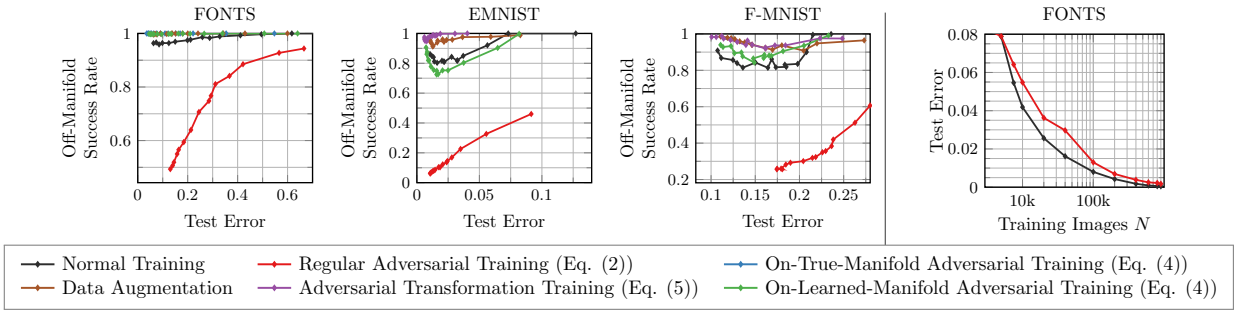

Figure 4. Left: Regular robustness and generalization are not contradictory, as demonstrated on FONTS, EMNIST and F-MNIST considering (regular) success rate plotted against test error. Except for adversarial training, success rate is hardly influenced by test error. Right: Adversarial training has higher sample complexity than normal training, as shown on FONTS.

on-manifold adversarial examples. Then, on-manifold adversarial training corresponds to robust optimization wrt. the true, or approximated, data distribution; i.e., training on "hard" examples. If the manifold cannot be approximated, we do adversarial training on known invariances of the data, e.g., using adversarial deformations (Alaifari et al., 2018; Xiao et al., 2018; Engstrom et al., 2017), referred to as adversarial transformation training. As on FONTS, we consider 6-degrees-of-freedom transformations corresponding to translation, shear, scaling and rotation; an η-constraint on the transformation parameters ensures perceptual similarity. We note that a similar approach has been used by Fawzi et al. (2016) as adversarial variants of data augmentation to boost generalization on, e.g., MNIST (LeCun et al., 1998).

We demonstrate the effectiveness of on-manifold adversarial training in Fig. 3. On FONTS, with access to the true manifold, on-manifold adversarial training is able to boost generalization significantly, especially for low N, i.e., few training images. Our VAE-GAN approximation on FONTS seems to be good enough to preserve this benefit. On EMNIST and F-MNIST, the benefit reduces with the difficulty of approximating the manifold; this is the "cost" of imperfect approximation. However, both on EMNIST and F-MNIST, identifying invariances and utilizing adversarial transformation training recovers the boost in generalization. Overall, on-manifold adversarial training is a promising tool for improving generalization and we expect its benefit to increase with better generative models.

2.4. Regular Robustness is Independent of Generalization

We argue that generalization, measured *on* the manifold wrt. the data distribution, is mostly independent of robustness against regular, possibly off-manifold, adversarial examples when varying the amount of training data. Specifically, in Fig. 4 (left) on FONTS, it can be observed that – except for adversarial training – the success rate is invariant to the test error. Similar behavior can be observed on EMNIST and F-MNIST, see Fig. 4 (right). These results imply that robustness and generalization are not contradicting goals; in fact, robust and accurate models can be found, however, might require higher sample complexity (Schmidt et al., 2018; Khoury & Hadfield-Menell, 2018). This is confirmed in Fig. 4 (right) where adversarial training requires roughly twice the amount of training data to reach the same accuracy as normal training. The higher sample complexity might be justified by the difficulty of the task: as adversarial examples tend to leave the manifold, the network has to learn adversarial, random directions *in addition* to the actual task.

Our results are in contrast to related work (Tsipras et al., 2018; Su et al., 2018) claiming that an inherent trade-off between robustness and generalization exists. However, in their studied toy dataset, Tsipras et al. (2018) allow the adversary to produce perturbations that change the actual, true label wrt. the data distribution. Thus, it is unclear whether the suggested trade-off actually exists for real datasets; our experiments, at least, seem to indicate the contrary. And Su et al. (2018) experimentally show a trade-off between adversarial robustness and generalization by studying *different* models on ImageNet (Russakovsky et al., 2015), while we found that the generalization performance does not influence robustness for any *arbitrary, but fixed* model.

3. Conclusion

In this paper, we showed that regular adversarial examples indeed leave the manifold as widely assumed. Additionally, we demonstrated that adversarial examples can also be found on the manifold, even if the manifold has to be approximated, e.g., using VAE-GANs. Then, we established that robustness against on-manifold adversarial examples is clearly related to generalization and on-manifold adversarial training exploits this relationship to boost generalization. Finally, we provided evidence that robustness against regular, unconstrained adversarial examples and generalization are not necessarily contradicting goals: for any arbitrary but fixed model, better generalization, e.g., through more training data, does not reduce robustness.

References

Alaifari, R., Alberti, G. S., and Gauksson, T. Adef: an iterative algorithm to construct adversarial deformations. *arXiv.org*, abs/1804.07729, 2018.

Athalye, A. and Carlini, N. On the robustness of the CVPR 2018 white-box adversarial example defenses. *arXiv.org*, abs/1804.03286, 2018.

Athalye, A., Carlini, N., and Wagner, D. A. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv.org*, abs/1802.00420, 2018.

Brown, T. B., Carlini, N., Zhang, C., Olsson, C., Christiano, P., and Goodfellow, I. Unrestricted adversarial examples. *arXiv.org*, abs/1809.08352, 2017.

Carlini, N. and Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. In *AISec*, 2017.

Carlini, N. and Wagner, D. A. Defensive distillation is not robust to adversarial examples. *arXiv.org*, abs/1607.04311, 2016.

Cohen, G., Afshar, S., Tapson, J., and van Schaik, A. EMNIST: an extension of MNIST to handwritten letters. *arXiv.org*, abs/1702.05373, 2017.

Engstrom, L., Tsipras, D., Schmidt, L., and Madry, A. A rotation and a translation suffice: Fooling CNNs with simple transformations. *arXiv.org*, abs/1712.02779, 2017.

Fawzi, A., Samulowitz, H., Turaga, D. S., and Frossard, P. Adaptive data augmentation for image classification. In *ICIP*, 2016.

Gilmer, J., Metz, L., Faghri, F., Schoenholz, S. S., Raghu, M., Wattenberg, M., and Goodfellow, I. Adversarial spheres. *ICLR Workshops*, 2018.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv.org*, abs/1412.6572, 2014.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

Jaderberg, M., Simonyan, K., Zisserman, A., and Kavukcuoglu, K. Spatial transformer networks. In *NIPS*, 2015.

Khoury, M. and Hadfield-Menell, D. On the geometry of adversarial examples. *arXiv.org*, abs/1811.00525, 2018.

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O. Autoencoding beyond pixels using a learned similarity metric. In *ICML*, 2016.

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. *ICLR*, 2018.

Rosca, M., Lakshminarayanan, B., Warde-Farley, D., and Mohamed, S. Variational approaches for auto-encoding generative adversarial networks. *arXiv.org*, abs/1706.04987, 2017.

Rozsa, A., Günther, M., and Boult, T. E. Are accuracy and robustness correlated. In *ICMLA*, 2016.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Li, F. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. *CoRR*, arXiv.org, 2018.

Schott, L., Rauber, J., Brendel, W., and Bethge, M. Towards the first adversarially robust neural network model on mnist. *arXiv.org*, abs/1805.09190, 2018.

Song, Y., Kim, T., Nowozin, S., Ermon, S., and Kushman, N. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *ICLR*, 2018a.

Song, Y., Shu, R., Kushman, N., and Ermon, S. Generative adversarial examples. *arXiv.org*, abs/1805.07894, 2018b.

Stutz, D., Hein, M., and Schiele, B. Disentangling adversarial robustness and generalization. *CVPR*, 2019.

Su, D., Zhang, H., Chen, H., Yi, J., Chen, P.-Y., and Gao, Y. Is robustness the cost of accuracy? – a comprehensive study on the robustness of 18 deep image classification models. *arXiv.org*, abs/1808.01688, 2018.

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

Tanay, T. and Griffin, L. A boundary tilting persepective on the phenomenon of adversarial examples. *arXiv.org*, abs/1608.07690, 2016.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. *arXiv.org*, abs/1805.12152, 2018.

Xiao, C., Zhu, J., Li, B., He, W., Liu, M., and Song, D. Spatially transformed adversarial examples. *ICLR*, 2018.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv.org*, abs/1708.07747, 2017.

Zhao, Z., Dua, D., and Singh, S. Generating natural adversarial examples. *ICLR*, 2018.