Disentangling Adversarial Robustness and Generalization

David Stutz[1] Matthias Hein[2] Bernt Schiele[1]
[1]Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken
[3]University of Tübingen, Tübingen
{david.stutz,schiele}@mpi-inf.mpg.de, matthias.hein@uni-tuebingen.de

Abstract

Obtaining deep networks that are robust against adversarial examples and *generalize well is an open problem. A recent hypothesis [102, 95] even states that* both *robust* and *accurate models are impossible, i.e., adversarial robustness and generalization are conflicting goals. In an effort to clarify the relationship between robustness and generalization, we assume an underlying, low-dimensional data manifold and show that: 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, and on-manifold adversarial training boosts generalization; 4. regular robustness and generalization are not necessarily contradicting goals. These assumptions imply that* both *robust* and *accurate models are possible. However, different models (architectures, training strategies etc.) can exhibit different robustness and generalization characteristics. To confirm our claims, we present extensive experiments on synthetic data (with known manifold) as well as on EMNIST [19], Fashion-MNIST [106] and CelebA [58].*

1. Introduction

Adversarial robustness describes a deep network's ability to defend against adversarial examples [97], imperceptibly perturbed images causing mis-classification. These adversarial attacks pose severe security threats, as demonstrated against Clarifai.com [57, 8] or Google Cloud Vision [37]. Despite these serious risks, defenses against such attacks have been largely ineffective; only adversarial training, i.e., training on adversarial examples [62, 30], has been shown to work well in practice [6, 5] – at the cost of computational overhead and reduced accuracy. Overall, the problem of adversarial robustness is left open and poorly understood – even for simple datasets such as EMNIST [19] and Fashion-MNIST [106].

The phenomenon of adversarial examples itself, i.e., their mere existence, has also received considerable attention. Recently, early explanations, e.g., attributing adversarial examples to "rare pockets" of the classification surface [97] or linearities in deep networks [30], have been superseded by the manifold assumption [27, 99]: adversarial examples are assumed to leave the underlying, low-dimensional but usually unknown data manifold. However, only [92] provide experimental evidence supporting this assumption. Yet, on a simplistic toy dataset, Gilmer et al. [27] also found adversarial examples on the manifold, as

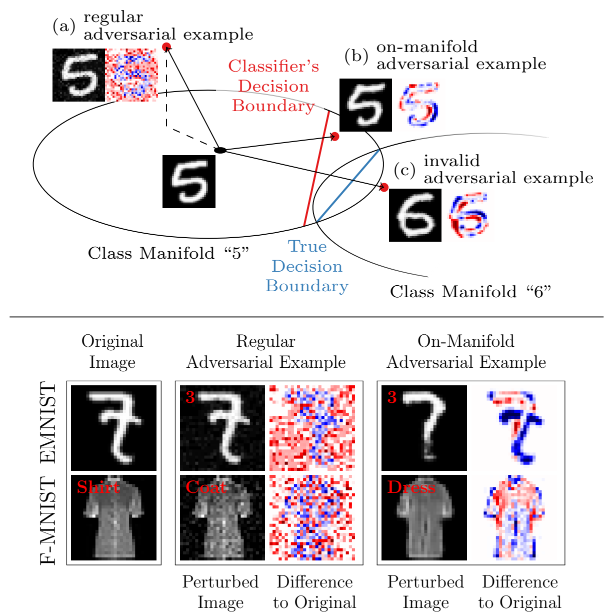

Figure 1: Adversarial examples, and their (normalized) difference to the original image, in the context of the underlying manifold, e.g., class manifolds "5" and "6" on EMNIST [19], allow to study their relation to generalization. Regular adversarial examples are not constrained to the manifold, cf. (a), and often result in (seemingly) random noise patterns; in fact, we show that they leave the manifold. However, adversarial examples on the manifold can be found as well, cf. (b), resulting in meaningful manipulations of the image content; however, care needs to be taken that the actual, true label wrt. the manifold does not change, cf. (c).

also tried on real datasets [93, 11, 110], rendering the manifold assumption questionable. Still, the manifold assumption fostered research on novel defenses [39, 72, 82].

Beyond the existence of adversarial examples, their relation to generalization is an important open problem. Recently, it has been argued [102, 95] that there exists an inherent trade-off, i.e., robust and accurate models seem impossible. While Tsipras et al. [102] provide a theoretical argument on a toy dataset, Su et al. [95] evaluate the robustness of different models on ImageNet [79]. However, these findings have to be questioned given the results in [27, 77] showing the opposite, i.e., better generalization helps robustness.

In order to address this controversy, and in contrast to [102, 96, 77], we consider adversarial robustness in the context of the underlying manifold. In particular, to break the hypothesis down, 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, adversarial examples are not guaranteed to lie on the manifold, cf. Fig. 1 (a). Adversarial examples can, however, also be constrained to the manifold, 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). For clarity, we refer to unconstrained adversarial examples, as illustrated in Fig. 1 (a), as *regular adversarial examples*; in contrast to adversarial examples constrained to the manifold, so-called *on-manifold adversarial examples*.

Contributions: Based on this distinction between regular robustness, i.e., against regular, unconstrained adversarial examples, and on-manifold robustness, i.e., against adversarial examples constrained to the manifold, 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 [19], Fashion-MNIST [106] and CelebA [58]. We will make our code and data publicly available.

2. Related Work

Attacks: Adversarial examples for deep networks were first reported in [97]; the problem of adversarial machine learning, however, has already been studied earlier [9]. Adversarial attacks on deep networks range from white-box attacks [97, 30, 49, 71, 66, 61, 14, 78, 21, 59], with full access to the model (weights, gradients etc.), to black-box attacks [17, 10, 96, 38, 80, 67], with limited access to model queries. White-box attacks based on first-order optimization, e.g., [61, 14], are considered state-of-the-art. Due to their transferability [57, 108, 70], these attacks can also be used in a black-box setting (e.g. using model stealing [87, 70, 101, 103, 69, 43]) and have, thus, become standard for evaluation. Recently, generative models have also been utilized to craft – or learn – more natural adversarial examples [93, 11, 110, 82]. Finally, adversarial examples have been applied to a wide variety of tasks, also beyond computer vision, e.g., [26, 18, 98, 48, 36, 55, 2, 16].

Defenses: Proposed defenses include detection and rejection methods [31, 25, 54, 60, 3, 63], pre-processing, quantization and dimensionality reduction methods [12, 73, 7], manifold-projection methods [39, 72, 82, 86], methods based on stochasticity/regularization or adapted architectures [109, 7, 68, 88, 34, 42, 76, 44, 50, 107], ensemble methods [56, 94, 33, 100], as well as adversarial training [109, 65, 35, 83, 90, 53, 61]; however, many defenses have been broken, often by considering "specialized" or novel attacks [13, 15, 5, 6]. In [6], only adversarial training, e.g., the work by Madry et al. [61], has been shown to be effective – although many recent defenses have not been studied extensively. Manifold-based methods, in particular, have received some attention lately: in [39, 72], generative adversarial networks [29] are used to project an adversarial example back to the learned manifold. Similarly, in [82], variational auto-encoders [47] are used to perform robust classification.

Generalization: Research also includes independent benchmarks of attacks and defenses [13, 15, 5, 6, 85], their properties [57, 84], as well as theoretical questions [34, 42, 23, 99, 27, 88, 102, 104]. Among others, the existence of adversarial examples [97, 30, 99] raises many questions. While Szegedy et al. [97] originally thought of adversarial examples as "extremely" rare negatives and Goodfellow et al. [30] attributed adversarial examples to the linearity in deep networks, others argued against these assumptions [27, 99]. Instead, a widely accepted theory is the manifold assumption; adversarial examples are assumed to leave the data manifold [27, 99, 39, 72, 82].

This paper is particularly related to work on the connection of adversarial examples to generalization [102, 95, 27, 77]. Tsipras et al. [102] and Su et al. [95] argue that there exists an inherent trade-off between robustness and gen-

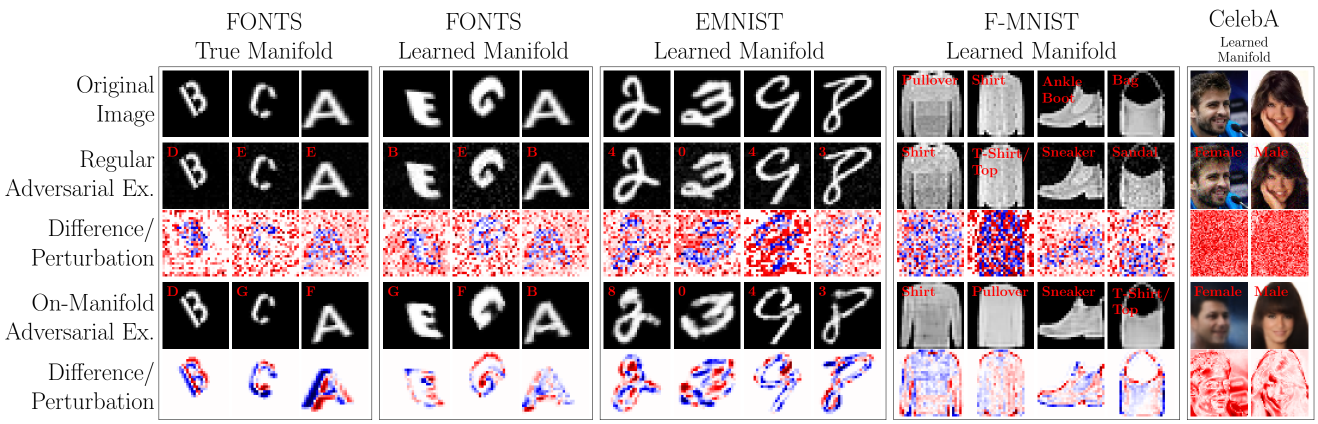

Figure 2: Regular and on-manifold adversarial examples on our synthetic dataset, FONTS, consisting of randomly transformed characters "A" to "J", EMNIST [19], F-MNIST [106] and CelebA [58]. On FONTS, the manifold is known by construction; in the other cases, the class manifolds have been approximated using VAE-GANs [51, 75]. The difference (normalized; or their magnitude on CelebA) to the original test image reveals the (seemingly) random noise patterns of regular adversarial examples in contrast to reasonable concept changes of on-manifold adversarial examples.

eralization. However, the theoretical argument in [102] is questionable as adversarial examples are allowed to change their actual, true label wrt. the data distribution, as illustrated Fig. 1 (c). The experimental results obtained in [95, 77] stem from comparing different architectures and training strategies; in contrast, we consider robustness and generalization for any arbitrary but fixed model. On a simple synthetic toy dataset, Gilmer et al. [27] show that on-manifold adversarial examples exist. We further show that on-manifold adversarial examples also exist on real datasets with unknown manifold, similar to [110]. In contrast to [27, 110], we utilize a gradient-based attack on the manifold, not in image space. Our work is also related to [24] and [65, 64] where variants of adversarial training are used to boost (semi-)supervised learning. While, e.g., Fawzi et al. [24], apply adversarial training to image transformations, we further perform adversarial training on adversarial examples constrained to the true, or approximated, manifold. This is also different from adversarial data augmentation schemes driven by GANs, e.g., [74, 91, 4, 20], where training examples are generated, but without the goal to be misclassified. Finally, [92] provide experimental evidence that adversarial examples have low probability under the data distribution; we show that adversarial examples have, in fact, zero probability.

3. Disentangling Adversarial Robustness and Generalization

To clarify the relationship between adversarial robustness and generalization, we explicitly distinguish between regular and on-manifold adversarial examples, as illustrated in Fig. 1. Then, the hypothesis [102, 95] that robustness and generalization are contradicting goals is challenged in four arguments: regular unconstrained adversarial examples leave the manifold; adversarial examples constrained to the manifold exist; robustness against on-manifold adversarial examples is essentially generalization; robustness against regular adversarial examples is not influenced by generalization when controlled through the amount of training data. Altogether, our results imply that adversarial robustness and generalization are not opposing objectives and both robust and accurate models are possible but require higher sample complexity.

3.1. Experimental Setup

Datasets: We use EMNIST [19], F(ashion)-MNIST [106] and CelebA [58] for our experiments (240k/40k, 60k/10k and 182k/20k training/test images); CelebA has been resized to 56×48 and we classify "Male" vs. "Female". Our synthetic dataset, FONTS, consists of letters "A" to "J" of 1000 Google Fonts randomly transformed (uniformly over translation, shear, scale, rotation in $[-0.2, 0.2]$, $[-0.5, 0.5]$, $[0.75, 1.15]$, $[-\pi/2, \pi/2]$) using a spatial transformer network [41] such that the generation process is completely differentiable. The latent variables correspond to the transformation parameters, font and class. We generated 960k/40k (balanced) training/test images of size 28×28.

We consider classifiers with three (four on CelebA) convolutional layers (4×4 kernels; stride 2; 16, 32, 64 channels), each followed by ReLU activations and batch normalization [40], and two fully connected layers. The networks are trained using ADAM [46], with learning rate 0.01 (decayed by 0.95 per epoch), weight decay 0.0001 and batch size 100, for 20 epochs. Most importantly, to control their generalization performance, we use N training images, with N between 250 and 40k; for each N, we train 5 models with random weight initialization [28] an report averages.

We learn class-specific VAE-GANs, similar to [51, 75],

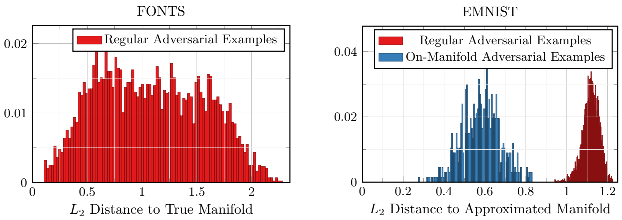

Figure 3: Distance of adversarial examples to the true, on FONTS (left), or approximated, on EMNIST (right), manifold. We show normalized histograms of the L_2 distance of adversarial examples to their projections onto the manifold (4377/3837 regular adversarial examples on FONTS/EMNIST; 667 on-manifold adversarial examples on EMNIST). Regular adversarial examples exhibit a significant distance to the manifold; on EMNIST, clearly distinguishable from on-manifold adversarial examples.

to approximate the underlying manifold; we refer to the supplementary material for details.

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., $f(\tilde{x}) \neq y$. While our results can be confirmed using other attacks and norms (see the supplementary material for [14] and transfer attacks), for clarity, we concentrate on the L_∞ white-box attack by Madry et al. [62] that directly maximizes the training loss,

$$\max_\delta \mathcal{L}(f(x+\delta), y) \quad \text{s.t.} \quad \|\delta\|_\infty \leq \epsilon, \tilde{x}_i \in [0, 1], \quad (1)$$

using projected gradient descent; where $\mathcal{L}$ is the cross-entropy loss and $\tilde{x} = x + \delta$. The ϵ-constraint is meant to ensure perceptual similarity. We run 40 iterations of ADAM [46] with learning rate 0.005 and consider 5 restarts, (distance and direction) uniformly sampled in the ϵ-ball for $\epsilon = 0.3$. Optimization is stopped as soon as the predicted label changes, i.e., $f(\tilde{x}) \neq y$. We attack 1000 test images.

Adversarial Training: An established defense is adversarial training, i.e., training on adversarial examples crafted during training [109, 65, 35, 83, 90, 53, 62]. Madry et al. [62] consider the min-max problem

$$\min_w \sum_{n=1}^N \max_{\|\delta\|_\infty \leq \epsilon} \mathcal{L}(f(x_n+\delta; w), y_n) \text{ s.t. } x_n+\delta \in [0, 1] \quad (2)$$

where w are the classifier's weights and x_n the training images. As shown in the supplementary material, we considered different variants [97, 30, 62]; in the paper, however, we follow common practice and train on 50% clean images and 50% adversarial examples [97]. For $\epsilon = 0.3$, the attack (for the inner optimization problem) is run for full 40 iterations, i.e., is not stopped at the first adversarial example found. Robustness of the obtained network is measured by computing the attack **success rate**, i.e., the fraction of successful attacks on correctly classified test images, as, e.g., in [14], for a fixed ϵ; lower success rate indicates higher robustness of the network.

3.2. Adversarial Examples Leave the Manifold

The idea of adversarial examples leaving the manifold is intuitive on EMNIST where particular background pixels are known to be constant, see Fig. 2. If an adversarial example $\tilde{x}$ manipulates these pixels, it has zero probability under the data distribution and its distance to the manifold, i.e., the distance to its projection $\pi(\tilde{x})$ onto the manifold, should be non-zero. On FONTS, with known generative process in the form of a decoder dec mapping latent variables z to images x, the projection is obtained iteratively: $\pi(\tilde{x}) = \text{dec}(\tilde{z})$ with $\tilde{z} = \text{argmin}_z \|\text{dec}(z) - \tilde{x}\|_2$ and z constrained to valid transformations (font and class, known from the test image x, stay constant). On EMNIST, as illustrated in Fig. 4 (right), the manifold is approximated using 50 nearest neighbors; the projection $\pi(\tilde{x})$ onto the subspace spanned by the x-centered nearest neighbors is computed through least squares. On both FONTS and EMNIST, the distance $\|\tilde{x} - \pi(\tilde{x})\|_2$ is considered to asses whether the adversarial example $\tilde{x}$ actually left the manifold.

On FONTS, Fig. 3 (left) shows that regular adversarial examples clearly exhibit non-zero distance to the manifold. In fact, the projections of these adversarial examples to the manifold are almost always the original test images; as a result, the distance to the manifold is essentially the norm of the corresponding perturbation: $\|\tilde{x} - \pi(\tilde{x})\|_2 \approx \|\tilde{x} - x\|_2 = \|\delta\|_2$. This suggests that the adversarial examples leave the manifold in an almost orthogonal direction. On EMNIST, in Fig. 3 (right), these results can be confirmed in spite of the crude local approximation of the manifold. Again, regular adversarial examples seem to leave the manifold almost orthogonally, i.e., their distance to the manifold coincides with the norm of the corresponding perturbations. These results show that regular adversarial examples essentially *are* off-manifold adversarial examples; this finding is intuitive as for well-trained classifiers, leaving the manifold should be the "easiest" way to fool it; results on F-MNIST as well as a more formal statement of this intuition can be found in the supplementary material.

3.3. On-Manifold Adversarial Examples

Given that regular adversarial examples leave the manifold, we intend to explicitly compute on-manifold adversarial examples. To this end, we assume our data distribution $p(x, y)$ to be conditional on the latent variables z, i.e., $p(x, y|z)$, corresponding to the underlying, low-dimensional manifold. On this manifold, however, there is no notion of "perceptual similarity" in order to ensure label invariance, i.e., distinguish valid on-manifold adversarial examples, Fig. 1 (b), from invalid ones that change the actual, true label, Fig. 1 (c):

Definition 1 (On-Manifold Adversarial Example)**.** Given the data distribution p, an on-manifold adversarial example for x with label y is a perturbed version $\tilde{x}$ such that $f(\tilde{x}) \neq y$ but $p(y|\tilde{x}) > p(y'|\tilde{x}) \forall y' \neq y$.

Note that the posteriors $p(y|\tilde{x})$ correspond to the true, unknown data distribution; any on-manifold adversarial example $\tilde{x}$ violating Def. 1 changed its actual, true label.

In practice, we assume access to an encoder and decoder modeling the (class-conditional) distributions $p(z|x, y)$ and $p(x|z, y)$ – in our case, achieved using VAE-GANs [51, 75]. Then, given the encoder enc and decoder dec and as illustrated in Fig. 4 (left), we obtain the latent code $z = \text{enc}(x)$ and compute the perturbation ζ by maximizing:

$$\max_{\zeta} \mathcal{L}(f(\text{dec}(z + \zeta)), y) \quad \text{s.t.} \quad \|\zeta\|_\infty \leq \eta. \tag{3}$$

The image-constraint, i.e., $\text{dec}(z + \zeta) \in [0, 1]$, is enforced by the decoder; the η-constraint can, again, be enforced by projection; and we can additionally enforce a constraint on $z + \zeta$, e.g., corresponding to a prior on z. Label invariance, as in Def. 1, is ensured by considering only class-specific encoders and decoders, i.e., the data distribution is approximated per class. We use $\eta = 0.3$ and the same optimization procedure as for Eq. (1); 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.

On-manifold adversarial examples obtained through Eq. (3) are similar to those crafted in [27], [82], [6] or [110]. However, in contrast to [27, 82, 6], 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. Also note that enforcing any similarity constraint through a norm on the manifold is significantly more meaningful compared to using a norm on the image space, as becomes apparent when comparing the obtained on-manifold adversarial examples in Fig. 2 to their regular counterparts. Compared to [110], we find on-manifold adversarial examples using a gradient-based approach instead of randomly sampling the latent space.

Fig. 2 shows on-manifold adversarial examples for all datasets, which we found significantly harder to obtain compared to their regular counterparts. On FONTS, using the true, known class manifolds, on-manifold adversarial examples clearly correspond to transformations of the original test image – reflecting the true latent space. For the learned class manifolds, the perturbations are less pronounced, often manipulating boldness or details of the characters. Due to the approximate nature of the learned VAE-GANs, these adversarial examples are strictly speaking not always part of the true manifold – as can be seen for the irregular "A" (Fig. 2, 6th column). On EMNIST and F-MNIST, on-manifold adversarial examples represent meaningful manipulations, such as removing the tail of a hand-

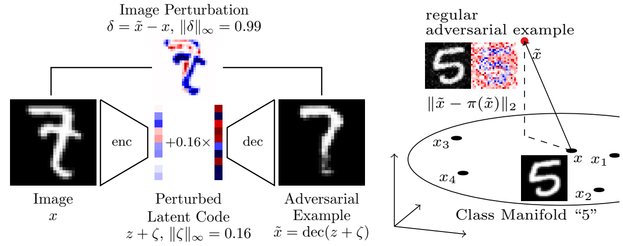

Figure 4: Left: On-manifold adversarial examples can be computed using learned, class-specific VAE-GANs [51, 75]. The perturbation ζ is obtained via Eq. (3) and added to the latent code $z = \text{enc}(x)$ yielding the adversarial example $\tilde{x} = \text{dec}(z + \zeta)$ with difference $\delta = \tilde{x} - x$ in image space. Right: The distance of a regular adversarial example $\tilde{x}$ to the manifold, approximated using nearest neighbors, is computed as the distance to its orthogonal projection $\pi(\tilde{x})$: $\|\tilde{x} - \pi(\tilde{x})\|_2$. Large distances indicate that the adversarial example likely left the manifold.

drawn "8" (Fig. 2, 10th column) or removing the collar of a pullover (Fig. 2, 11th column), in contrast to the random noise patterns of regular adversarial examples. However, these usually incur a smaller change in the images space; which also explains why regular, unconstrained adversarial examples almost always leave the manifold. Still, on-manifold adversarial examples are perceptually close to the original images. On CelebA, the quality of on-manifold adversarial examples is clearly limited by the approximation quality of our VAE-GANs. Finally, Fig. 3 (right) shows that on-manifold adversarial examples are closer to the manifold than regular adversarial examples – in spite of the crude approximation of the manifold on EMNIST.

3.4. 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. 5 (top left) where test error and on-manifold success rate on FONTS are shown. As expected, better generalization, i.e., using more training images N, also reduces on-manifold success rate. In order to make this relationship explicit, Fig. 5 (bottom) plots on-manifold success rate against test error. Then, especially for FONTS and EMNIST, the relationship of on-manifold robustness and generalization becomes apparent. 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, the experiments show a clear relationship between on-manifold robustness and generalization.

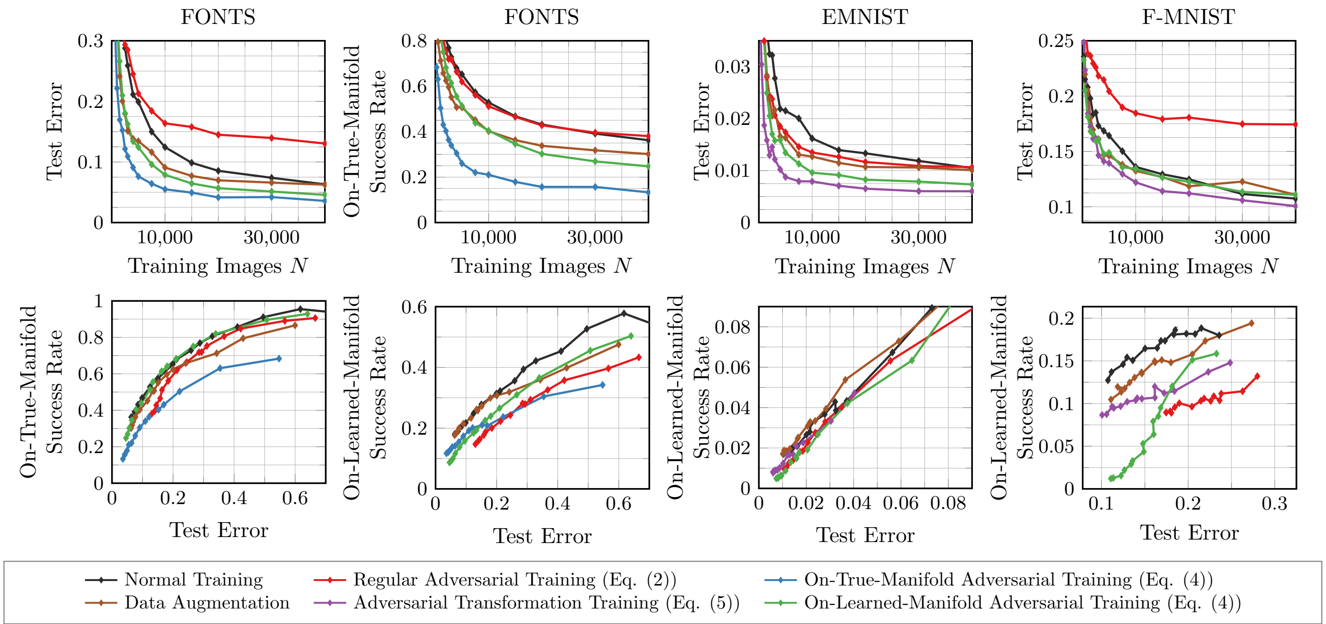

Figure 5: On-manifold robustness is strongly related to generalization, as shown on FONTS, EMNIST and F-MNIST considering on-manifold success rate and test error. Top: test error and on-manifold success rate in relation to the number of training images. As test error reduces, so does on-manifold success rate. Bottom: on-manifold success rate plotted against test error reveals the strong relationship between on-manifold robustness and generalization.

3.4.1 On-Manifold Adversarial Training Boosts Generalization

Given that generalization positively influences on-manifold robustness, we propose to adapt adversarial training to the on-manifold case in order to boost generalization:

$$\min_w \sum_{n=1}^{N} \max_{\|\zeta\|_\infty \leq \eta} \mathcal{L}(f(\text{dec}(z_n + \zeta); w), y_n). \quad (4)$$

with $z_n = \text{dec}(x_n)$ being the latent codes corresponding to training images x_n. Then, on-manifold adversarial training corresponds to robust optimization wrt. the true, or approximated, data distribution. For example, with the perfect decoder on FONTS, the inner optimization problem finds "hard" images irrespective of their likelihood under the data distribution. For approximate dec, the benefit of on-manifold adversarial training depends on how well the true data distribution is matched, i.e., how realistic the obtained on-manifold adversarial examples are; in our case, this depends on the quality of the learned VAE-GANs.

Instead of approximating the manifold using generative models, we can exploit known invariances of the data. Then, adversarial training can be applied to these invariances, assuming that they are part of the true manifold. In practice, this can, for example, be accomplished using adversarial deformations [1, 105, 22], i.e., adversarially crafted transformations of the image. For example, as on FONTS, we consider 6-degrees-of-freedom transformations corresponding to translation, shear, scaling and rotation:

$$\min_w \sum_{n=1}^{N} \max_{\|t\|_\infty \leq \eta, t \in \mathbb{R}^6} \mathcal{L}(f(T(x_n; t); w), y_n). \quad (5)$$

where $T(x; t)$ denotes the transformation of image x with parameters t and the η-constraint ensures similarity and label invariance. Again, the transformations can be applied using spatial transformer networks [41] such that T is differentiable; t can additionally be constrained to a reasonable space of transformations. We note that a similar approach has been used by Fawzi et al. [24] to boost generalization on, e.g., MNIST [52]. However, the approach was considered as an adversarial variant of data augmentation and not motivated through the lens of on-manifold robustness. We refer to Eq. (5) as adversarial transformation training and note that, on FONTS, this approach is equivalent to on-manifold adversarial training as the transformations coincide with the actual, true manifold by construction. We also include a data augmentation baseline, where the transformations t are applied randomly.

We demonstrate the effectiveness of on-manifold adversarial training in Fig. 5 (top). 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 the benefit of on-manifold adversarial training. On EMNIST and F-MNIST, the benefit reduces with the difficulty of approximating the manifold; this is the "cost" of imperfect approx-

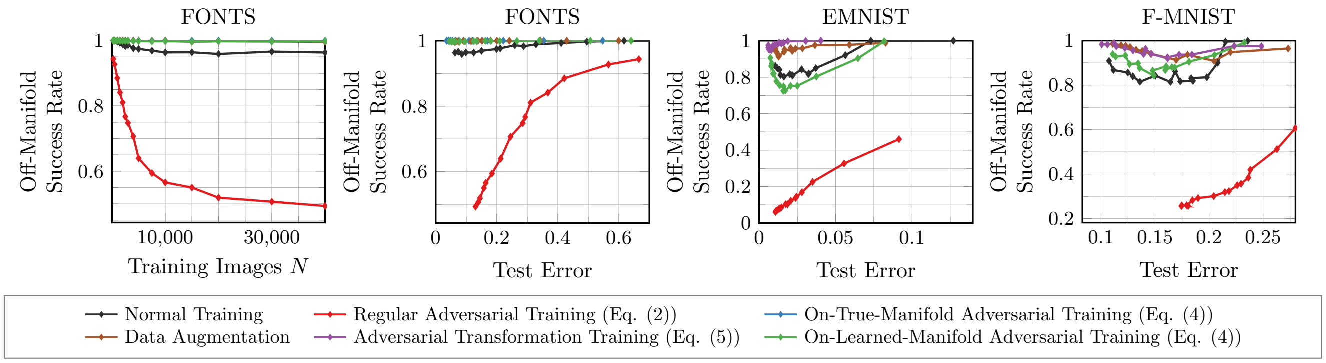

Figure 6: Regular robustness is not related to generalization, as demonstrated on FONTS, EMNIST and F-MNIST considering test error and (regular) success rate. On FONTS (left), success rate is not influenced by test error, except for adversarial training. Plotting success rate against test error highlights the independence of robustness and generalization; however, different training strategies exhibit different robustness-generalization characteristics.

imation. While the benefit is still significant on EMNIST, it diminishes on F-MNIST. However, both on EMNIST and F-MNIST, identifying invariances and utilizing adversarial transformation training recovers the boost in generalization; especially in contrast to the random data augmentation baseline. Overall, on-manifold adversarial training is a promising tool for improving generalization and we expect its benefit to increase with better generative models.

3.5. Regular Robustness is Independent of Generalization

We argue that generalization, as 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. 6 (left) for FONTS, it can be observed that – except for adversarial training – the success rate is invariant to the test error. This can best be seen when plotting the success rate against test error for different numbers of training examples, cf. Fig. 6 (middle left): only for adversarial training there exists a clear relationship; for the remaining training schemes success rate is barely influenced by the test error. In particular, better generalization does not worsen robustness. Similar behavior can be observed on EMNIST and F-MNIST, see Fig. 6 (right). Here, it can also be seen that different training strategies exhibit different characteristics wrt. robustness and generalization. Overall, regular robustness and generalization are not necessarily contradicting goals.

As mentioned in Section 1, these findings are in contrast to related work [102, 95] claiming that an inherent trade-off between robustness and generalization exists. For example, Tsipras et al. [102] use a synthetic toy dataset to theoretically show that no model can be both robust and accurate (on this dataset). However, they allow the adversary to produce perturbations that change the actual, true label wrt. the data distribution, i.e., the considered adversarial examples are not adversarial examples according to Def. 1. Thus, it is unclear whether the suggested trade-off actually exists for real datasets; our experiments, at least, as well as further analysis in the supplementary material seem to indicate the contrary. Similarly, Su et al. [95] experimentally show a trade-off between adversarial robustness and generalization by studying different models on ImageNet [79]. However, Su et al. compare the robustness and generalization characteristics of different models (i.e., different architectures, training strategies etc.), while we found that the generalization performance does not influence robustness for any *arbitrary, but fixed* model.

3.6. Discussion

Our results imply that robustness and generalization are not necessarily conflicting goals, as believed in related work [102, 95]. This means, in practice, for any arbitrary but fixed model, better generalization will not worsen regular robustness. Different models (architectures, training strategies etc.) might, however, exhibit different robustness and generalization characteristics, as also shown in [95, 77]. For adversarial training, on regular adversarial examples, the commonly observed trade-off between robustness and generalization is explained by the tendency of adversarial examples to leave the manifold. As result, the network has to learn (seemingly) random, but adversarial, noise patterns *in addition* to the actual task at hand; rendering the learning problem harder. On simple datasets, such as EMNIST, these adversarial directions might avoid overfitting; on harder tasks, e.g., FONTS or F-MNIST, the discrepancy in test error between normal and adversarial training increases. Our results also support the hypothesis that regular adversarial training has higher sample complexity [81, 45]. In fact, on FONTS, adversarial training can reach the same accuracy as normal training with roughly twice the amount of training data, as demonstrated in Fig. 7 (top). Furthermore, as illustrated in Fig. 7 (bottom), the trade-off between

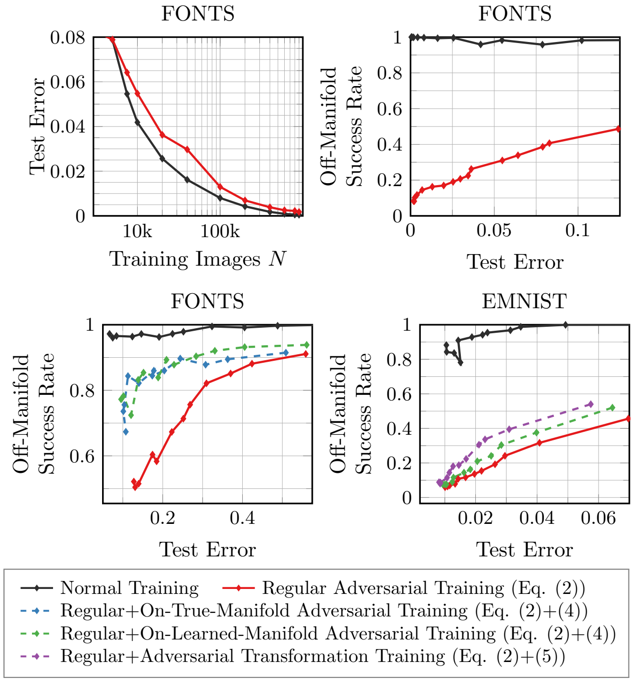

Figure 7: Adversarial training on regular adversarial examples, potentially leaving the manifold, renders the learning problem more difficult. Top: With roughly 1.5 to 2 times the training data, adversarial training can still reach the same accuracy as normal training; results for ResNet-13 [32]. Bottom: Additionally, the trade-off can be controlled by combining regular and on-manifold adversarial training; results averaged over 3 models.

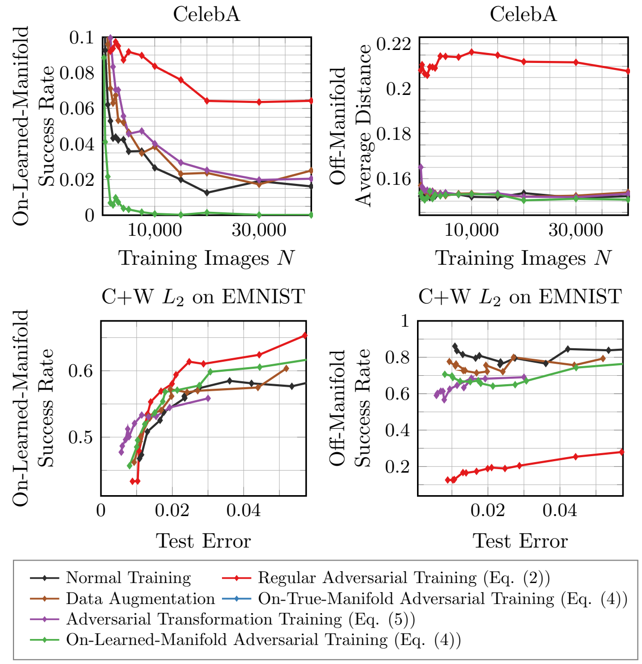

Figure 8: Results on CelebA and using the L_2 Carlini and Wagner [14] attack. On CelebA, as the class manifolds are significantly harder to approximate, the benefit of on-manifold adversarial training diminishes. For [14], we used 120 iterations; our hypotheses are confirmed, although [14] does not use the training loss as attack objective and the L_2 norm changes the similarity-constraint for regular and on-manifold adversarial examples.

regular robustness and generalization can be controlled by combining regular and on-manifold adversarial training, i.e. boost generalization while reducing robustness.

The presented results can also be confirmed on more complex datasets, such as CelebA, and using different threat models, i.e., attacks. On CelebA, where VAE-GANs have difficulties approximating the manifold, Fig. 8 (top left) shows that on-manifold robustness still improves with generalization although most on-manifold adversarial examples are not very realistic, see Fig. 2. Similarly, regular robustness, see Fig. 8 (top right), is not influenced by generalization; here, we also show that the average distance of the perturbation, i.e., average $\|\delta\|_\infty$, when used to asses robustness leads to the same conclusions. Similarly, as shown in Fig. 8 (bottom), our findings are confirmed using Carlini and Wagner's attack [14] with L_2-norm – to show that the results generalize across norms. However, overall, we observed lower success rates using [14] and the L_2 norm. Finally, our results can also be reproduced using transfer attacks (i.e., black-box attacks, which are generally assumed to be subsumed by white-box attacks [6]) as well as and different architectures such as multi-layer perceptrons, ResNets [32] and VGG [89], as detailed in the supplementary material.

4. Conclusion

In this paper, we intended to disentangle the relationship between adversarial robustness and generalization by initially adopting the hypothesis that robustness and generalization are contradictory [102, 95]. By considering adversarial examples in the context of the low-dimensional, underlying data manifold, we formulated and experimentally confirmed four assumptions. First, we showed that regular adversarial examples indeed leave the manifold, as widely assumed in related work [27, 99, 39, 72, 82]. Second, we demonstrated that adversarial examples can also be found on the manifold, so-called on-manifold adversarial examples; even if the manifold has to be approximated, e.g., using VAE-GANs [51, 75]. Third, we established that robustness against on-manifold adversarial examples is clearly related to generalization. Our proposed on-manifold adversarial training exploits this relationship to boost generalization using an approximate manifold, or known invariances. Fourth, 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

[1] Rima Alaifari, Giovanni S. Alberti, and Tandri Gauksson. Adef: an iterative algorithm to construct adversarial deformations. *arXiv.org*, abs/1804.07729, 2018. 6

[2] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani B. Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. In *EMNLP*, 2018. 2

[3] Laurent Amsaleg, James Bailey, Dominique Barbe, Sarah M. Erfani, Michael E. Houle, Vinh Nguyen, and Milos Radovanovic. The vulnerability of learning to adversarial perturbation increases with intrinsic dimensionality. In *WIFS*, 2017. 2

[4] Antreas Antoniou, Amos J. Storkey, and Harrison Edwards. Augmenting image classifiers using data augmentation generative adversarial networks. In *ICANN*, 2018. 3

[5] Anish Athalye and Nicholas Carlini. On the robustness of the CVPR 2018 white-box adversarial example defenses. *arXiv.org*, abs/1804.03286, 2018. 1, 2

[6] Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv.org*, abs/1802.00420, 2018. 1, 2, 5, 8

[7] Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Dimensionality reduction as a defense against evasion attacks on machine learning classifiers. *arXiv.org*, abs/1704.02654, 2017. 2

[8] Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Exploring the space of black-box attacks on deep neural networks. *arXiv.org*, abs/1712.09491, 2017. 1

[9] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018. 2

[10] Wieland Brendel and Matthias Bethge. Comment on "biologically inspired protection of deep networks from adversarial attacks". *arXiv.org*, abs/1704.01547, 2017. 2

[11] Tom B. Brown, Nicholas Carlini, Chiyuan Zhang, Catherine Olsson, Paul Christiano, and Ian Goodfellow. Unrestricted adversarial examples. *arXiv.org*, abs/1809.08352, 2017. 2

[12] Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *ICLR*, 2018. 2

[13] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *AISec*, 2017. 2

[14] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *SP*, 2017. 2, 4, 8

[15] Nicholas Carlini and David A. Wagner. Defensive distillation is not robust to adversarial examples. *arXiv.org*, abs/1607.04311, 2016. 2

[16] Nicholas Carlini and David A. Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *SP*, 2018. 2

[17] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *AISec*, 2017. 2

[18] Moustapha M Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured visual and speech recognition models with adversarial examples. In *NIPS*, 2017. 2

[19] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: an extension of MNIST to handwritten letters. *arXiv.org*, abs/1702.05373, 2017. 1, 2, 3

[20] Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. *arXiv.org*, abs/1805.09501, 2018. 3

[21] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018. 2

[22] Logan Engstrom, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling CNNs with simple transformations. *arXiv.org*, abs/1712.02779, 2017. 6

[23] Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Fundamental limits on adversarial robustness. In *ICML Workshops*, 2015. 2

[24] Alhussein Fawzi, Horst Samulowitz, Deepak S. Turaga, and Pascal Frossard. Adaptive data augmentation for image classification. In *ICIP*, 2016. 3, 6

[25] Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. Detecting adversarial samples from artifacts. *arXiv.org*, abs/1703.00410, 2017. 2

[26] Volker Fischer, Mummadi Chaithanya Kumar, Jan Hendrik Metzen, and Thomas Brox. Adversarial examples for semantic image segmentation. *ICLR*, 2017. 2

[27] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S. Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *ICLR Workshops*, 2018. 1, 2, 3, 5, 8

[28] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010. 3

[29] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 2

[30] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv.org*, abs/1412.6572, 2014. 1, 2, 4

[31] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. On the (statistical) detection of adversarial examples. *arXiv.org*, abs/1702.06280, 2017. 2

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 8

[33] Warren He, James Wei, Xinyun Chen, Nicholas Carlini, and Dawn Song. Adversarial example defense: Ensembles of weak defenses are not strong. In *USENIX Workshops*, 2017. 2

[34] Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *NIPS*, 2017. 2

[35] Ruitong Huang, Bing Xu, Dale Schuurmans, and Csaba Szepesvári. Learning with a strong adversary. *arXiv.org*, abs/1511.03034, 2015. 2, 4

[36] Sandy H. Huang, Nicolas Papernot, Ian J. Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *ICLR*, 2017. 2

[37] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *ICML*, 2018. 1

[38] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. *arXiv.org*, abs/1807.07978, 2018. 2

[39] Andrew Ilyas, Ajil Jalal, Eirini Asteri, Constantinos Daskalakis, and Alexandros G. Dimakis. The robust manifold defense: Adversarial training using generative models. *arXiv.org*, abs/1712.09196, 2017. 2, 8

[40] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 3

[41] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NIPS*, 2015. 3, 6

[42] Daniel Jakubovitz and Raja Giryes. Improving DNN robustness to adversarial attacks using jacobian regularization. *arXiv.org*, abs/1803.08680, 2018. 2

[43] Mika Juuti, Sebastian Szyller, Alexey Dmitrenko, Samuel Marchal, and N. Asokan. PRADA: Protecting against DNN model stealing attacks. *arXiv.org*, abs/1805.02628, 2018. 2

[44] Harini Kannan, Alexey Kurakin, and Ian J. Goodfellow. Adversarial logit pairing. *arXiv.org*, abs/1803.06373, 2018. 2

[45] Marc Khoury and Dylan Hadfield-Menell. On the geometry of adversarial examples. *arXiv.org*, abs/1811.00525, 2018. 7

[46] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 3, 4

[47] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 2

[48] Jernej Kos, Ian Fischer, and Dawn Song. Adversarial examples for generative models. In *SP Workshops*, 2018. 2

[49] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv.org*, abs/1607.02533, 2016. 2

[50] Alex Lamb, Jonathan Binas, Anirudh Goyal, Dmitriy Serdyuk, Sandeep Subramanian, Ioannis Mitliagkas, and Yoshua Bengio. Fortified networks: Improving the robustness of deep networks by modeling the manifold of hidden representations. *arXiv.org*, abs/1804.02485, 2018. 2

[51] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *ICML*, 2016. 3, 5, 8

[52] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 6

[53] Hyeungill Lee, Sungyeob Han, and Jungwoo Lee. Generative adversarial trainer: Defense to adversarial perturbations with GAN. *arXiv.org*, abs/1705.03387, 2017. 2, 4

[54] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *CVPR*, 2018. 2

[55] Yen-Chen Lin, Zhang-Wei Hong, Yuan-Hong Liao, Meng-Li Shih, Ming-Yu Liu, and Min Sun. Tactics of adversarial attack on deep reinforcement learning agents. In *IJCAI*, 2017. 2

[56] Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. Towards robust neural networks via random self-ensemble. *arXiv.org*, abs/1712.00673, 2017. 2

[57] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *ICLR*, 2017. 1, 2

[58] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 1, 2, 3

[59] Bo Luo, Yannan Liu, Lingxiao Wei, and Qiang Xu. Towards imperceptible and robust adversarial example attacks against neural networks. In *AAAI*, 2018. 2

[60] Xingjun Ma, Bo Li, Yisen Wang adn Sarah M. Erfani, Sudanthi Wijewickrema, Michael E. Houle, Grant Schoenebeck, Dawn Song, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv.org*, abs/1801.02613, 2018. 2

[61] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv.org*, abs/1706.06083, 2017. 2

[62] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ICLR*, 2018. 1, 4

[63] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. *arXiv.org*, abs/1702.04267, 2017. 2

[64] Takeru Miyato, Shin-ichi Maeda, Shin Ishii, and Masanori Koyama. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *PAMI*, 2018. 3

[65] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. Distributional smoothing with virtual adversarial training. *ICLR*, 2016. 2, 3, 4

[66] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *CVPR*, 2016. 2

[67] Nina Narodytska and Shiva Prasad Kasiviswanathan. Simple black-box adversarial attacks on deep neural networks. In *CVPR Workshops*, 2017. 2

[68] Aran Nayebi and Surya Ganguli. Biologically inspired protection of deep networks from adversarial attacks. *arXiv.org*, abs/1703.09202, 2017. 2

[69] Seong Joon Oh, Max Augustin, Mario Fritz, and Bernt Schiele. Towards reverse-engineering black-box neural networks. *ICLR*, 2018. 2

[70] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *AsiaCCS*. ACM, 2017. 2

[71] Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *SP*, 2016. 2

[72] Rama Chellappa Pouya Samangouei, Maya Kabkab. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. *ICLR*, 2018. 2, 8

[73] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James A. Storer. Protecting JPEG images against adversarial attacks. In *DCC*, 2018. 2

[74] Alexander J. Ratner, Henry R. Ehrenberg, Zeshan Hussain, Jared Dunnmon, and Christopher Ré. Learning to compose domain-specific transformations for data augmentation. In *NIPS*, 2017. 3

[75] Mihaela Rosca, Balaji Lakshminarayanan, David Warde-Farley, and Shakir Mohamed. Variational approaches for auto-encoding generative adversarial networks. *arXiv.org*, abs/1706.04987, 2017. 3, 5, 8

[76] Andrew Slavin Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *AAAI*, 2018. 2

[77] Andras Rozsa, Manuel Günther, and Terrance E. Boult. Are accuracy and robustness correlated. In *ICMLA*, 2016. 2, 3, 7

[78] Andras Rozsa, Manuel Günther, and Terrance E. Boult. Adversarial robustness: Softmax versus openmax. In *BMVC*, 2017. 2

[79] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 2, 7

[80] Sayantan Sarkar, Ankan Bansal, Upal Mahbub, and Rama Chellappa. UPSET and ANGRI : Breaking high performance image classifiers. *arXiv.org*, abs/1707.01159, 2017. 2

[81] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *CoRR*, arXiv.org, 2018. 7

[82] Lukas Schott, Jonas Rauber, Wieland Brendel, and Matthias Bethge. Towards the first adversarially robust neural network model on mnist. *arXiv.org*, abs/1805.09190, 2018. 2, 5, 8

[83] Uri Shaham, Yutaro Yamada, and Sahand Negahban. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing*, 307:195–204, 2018. 2, 4

[84] Mahmood Sharif, Lujo Bauer, and Michael K. Reiter. On the suitability of l_p-norms for creating and preventing adversarial examples. *arXiv.org*, abs/1802.09653, 2018. 2

[85] Yash Sharma and Pin-Yu Chen. Attacking the madry defense model with l1-based adversarial examples. *arXiv.org*, abs/1710.10733, 2017. 2

[86] Shiwei Shen, Guoqing Jin, Ke Gao, and Yongdong Zhang. Ape-gan: Adversarial perturbation elimination with gan. *arXiv.org*, abs/1707.05474, 2017. 2

[87] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *SP*, 2017. 2

[88] Carl-Johann Simon-Gabriel, Yann Ollivier, Bernhard Schölkopf, Lon Bottou, and David Lopez-Paz. Adversarial vulnerability of neural networks increases with input dimension. *arXiv.org*, abs/1802.01421, 2018. 2

[89] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv.org*, abs/1409.1556, 2014. 8

[90] Aman Sinha, Hongseok Namkoong, and John C. Duchi. Certifiable distributional robustness with principled adversarial training. *ICLR*, 2018. 2, 4

[91] Leon Sixt, Benjamin Wild, and Tim Landgraf. Rendergan: Generating realistic labeled data. *Frontiers in Robotics and AI*, 2018, 2018. 3

[92] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *ICLR*, 2018. 1, 3

[93] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Generative adversarial examples. *arXiv.org*, abs/1805.07894, 2018. 2

[94] Thilo Strauss, Markus Hanselmann, Andrej Junginger, and Holger Ulmer. Ensemble methods as a defense to adversarial perturbations against deep neural networks. *arXiv.org*, abs/1709.03423, 2017. 2

[95] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy? – a comprehensive study on the robustness of 18 deep image classification models. *arXiv.org*, abs/1808.01688, 2018. 1, 2, 3, 7, 8

[96] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *arXiv.org*, abs/1710.08864, 2017. 2

[97] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv.org*, abs/1312.6199, 2013. 1, 2, 4

[98] Pedro Tabacof, Julia Tavares, and Eduardo Valle. Adversarial images for variational autoencoders. *arXiv.org*, abs/1612.00155, 2016. 2

[99] Thomas Tanay and Lewis Griffin. A boundary tilting persepective on the phenomenon of adversarial examples. *arXiv.org*, abs/1608.07690, 2016. 1, 2, 8

[100] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. *ICLR*, 2018. 2

[101] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In *USENIX*, 2016. 2

[102] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv.org*, abs/1805.12152, 2018. 1, 2, 3, 7, 8

[103] Binghui Wang and Neil Zhenqiang Gong. Stealing hyper-parameters in machine learning. In *SP*, 2018. 2

[104] Yizhen Wang, Somesh Jha, and Kamalika Chaudhuri. Analyzing the robustness of nearest neighbors to adversarial examples. In *ICML*, 2018. 2

[105] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. *ICLR*, 2018. 6

[106] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv.org*, abs/1708.07747, 2017. 1, 2, 3

[107] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan L. Yuille. Mitigating adversarial effects through randomization. *ICLR*, 2018. 2

[108] Cihang Xie, Zhishuai Zhang, Jianyu Wang, Yuyin Zhou, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. *CoRR*, abs/1803.06978, 2018. 2

[109] Valentina Zantedeschi, Maria-Irina Nicolae, and Ambrish Rawat. Efficient defenses against adversarial attacks. In *AISec*, 2017. 2, 4

[110] Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. *ICLR*, 2018. 2, 3, 5