Adversarial Robustness on In- and Out-Distribution Improves Explainability

Maximilian Augustin, Alexander Meinke, and Matthias Hein

University of Tübingen, Germany

Abstract. Neural networks have led to major improvements in image classification but suffer from being non-robust to adversarial changes, unreliable uncertainty estimates on out-distribution samples and their inscrutable black-box decisions. In this work we propose RATIO, a training procedure for Robustness via Adversarial Training on In- and Out-distribution, which leads to robust models with reliable and robust confidence estimates on the out-distribution. RATIO has similar generative properties to adversarial training so that visual counterfactuals produce class specific features. While adversarial training comes at the price of lower clean accuracy, RATIO achieves state-of-the-art l_2-adversarial robustness on CIFAR10 and maintains better clean accuracy.

1 Introduction

Deep neural networks have shown phenomenal success in achieving high accuracy on challenging classification tasks [29]. However, they are lacking in terms of robustness against adversarial attacks [51], make overconfident predictions [20, 21] especially on out-of-distribution (OOD) data [41, 24] and their black box decisions are inscrutable [56]. Progress has been made with respect to all these aspects but there is currently no approach which is accurate, robust, has good confidence estimates and is explainable. Adversarial training (AT) [34] leads to models robust against adversarial attacks in a defined threat model and has recently been shown to produce classifiers with generative capabilities [46]. However, AT typically suffers from a significant drop in accuracy and is over-confident on OOD data as we show in this paper. Adversarial confidence enhanced training (ACET) [21] enforces low confidence in a neighborhood around OOD samples and can be seen as adversarial training on the out-distribution. ACET leads to models with good OOD detection performance even in an adversarial setting and suffers from a smaller loss in clean accuracy compared to AT. However, ACET models typically are significantly less robust than adversarially trained models.

In this paper we show that combining AT and ACET into RATIO, Robustness via Adversarial Training on In- and Out-distribution, inherits the good properties of adversarial training and ACET without, or at least with significantly reduced, negative effects, e.g. we get SOTA l_2-robustness on CIFAR10 and have better clean accuracy than AT. On top of this we get reliable confidence estimates on the out-distribution even in a worst case scenario. In particular AT

Table 1: *Summary:* We show clean and robust accuracy in an l_2-threat model with $\epsilon = 0.5$ and the expected calibration error (ECE). For OOD detection we report the mean of clean and worst case AUC over several out-distributions in an l_2-threat model with $\epsilon = 1.0$ as well as the mean maximal confidence (MMC) on the out-distributions. In light red we highlight failure cases for certain metrics. Only RATIO-0.25 ($R_{0.25}$) has good performance across all metrics.

CIFAR10	Plain	OE	ACET	$M_{0.5}$	$AT_{0.5}$	$AT_{0.25}$	JEM-0	$R_{0.5}$	$\mathbf{R}_{0.25}$
Acc. ↑	96.2	**96.4**	94.1	90.8	90.8	94.0	92.8	91.1	93.5
R. $Acc._{0.5}$ ↑	0.0	0.0	52.3	69.3	70.4	65.0	40.5	**73.3**	70.5
ECE (in %) ↓	**1.0**	2.9	2.8	2.6	2.2	2.2	3.9	2.8	2.7
AUC ↑	94.2	**96.5**	94.7	81.8	88.9	92.7	75.0	95.6	95.0
WC $AUC_{1.0}$ ↑	1.6	8.7	81.9	48.5	57.4	42.0	14.6	83.6	**84.3**
MMC ↓	62.0	**31.9**	39.1	62.7	55.8	55.2	69.7	**31.9**	33.9
SVHN	Plain	OE	ACET	$AT_{0.5}$	$AT_{0.25}$	$R_{0.5}$	$\mathbf{R}_{0.25}$		
Acc. ↑	97.3	97.6	**97.8**	94.4	96.7	94.3	96.8		
R. $Acc._{0.5}$ ↑	0.9	0.3	28.8	68.1	63.0	**68.4**	64.8		
ECE ↓	0.9	0.9	1.6	1.6	**0.8**	2.0	1.8		
AUC ↑	96.9	99.6	99.8	91.0	97.0	99.8	**99.9**		
WC $AUC_{1.0}$ ↑	8.5	18.2	96.0	51.1	48.3	**97.5**	**97.5**		
MMC ↓	61.5	16.3	11.8	67.1	49.1	12.1	**11.1**		

yields highly overconfident predictions on out-distribution images in the absence of class specific features whereas RATIO only yields high confident predictions if recognizable features are present. In summary, RATIO achieves high clean accuracy, is robust, calibrated and has generative properties which can be used to produce high-quality visual counterfactual explanations: see Table 1 for a summary of our results for CIFAR10 and SVHN and Table 2 for CIFAR100 and restricted ImageNet [54].

2 Related Work

Adversarial Robustness. Adversarial attacks are small changes of an image with respect to some distance measure, which change the decision of a classifier [51]. Many defenses have been proposed but with more powerful or adapted attacks most of them could be defeated [13, 8, 3, 38]. Adversarial training (AT) [34] is the most widely used approach that has not been broken. However, adversarial robustness comes at the price of a drop in accuracy [48, 50]. Recent variations are using other losses [60] and boost robustness via generation of additional training data [9, 1] or pre-training [26]. Another line of work are provable defenses, either deterministic [58, 12, 37, 17] or based on randomized smoothing [33, 30, 11]. However, provable defenses are still not competitive with the empirical robustness of adversarial training for datasets like CIFAR10 and have even worse accuracy. We show that using AT on the in-distribution and out-distribution leads to a smaller drop in clean accuracy and similar or better robustness.

Confidence on In- and Out-distribution. Neural networks have been shown to yield overly confident predictions far away from the training data [41, 24, 32] and this is even provably the case for ReLU networks [21]. Moreover, large neural networks are not calibrated on the in-distribution and have a bias to be overconfident [20]. The overconfidence on the out-distribution has been tackled in [31, 21, 25] by enforcing low-confidence predictions on a large out-distribution dataset e.g. using the 80 million tiny images dataset[25] leads to state-of-the-art results. However, if one maximizes the confidence in a ball around out-distribution-samples, most OOD methods are again overconfident [48, 21, 49, 35] and only AT on the out-distribution as in ACET [21] or methods providing guaranteed worst case OOD performance [35, 7] work in this worst-case setting. We show that RATIO leads to better worst case OOD performance than ACET.

Counterfactual Explanations. Counterfactual explanations have been proposed in [56] as a tool for making classifier decisions plausible, since humans also justify decisions via counterfactuals "I would have decided for X, if Y had been true" [36]. Other forms are explanations based on image features [22, 23]. However, changing the decision for image classification in *image space* for non-robust models leads to adversarial samples [15] with changes that are visually not meaningful. Thus visual counterfactuals are often based on generative models or restrictions on the space of image manipulation [45, 42, 10, 18, 61, 57]. Robust models wrt l_2-adversarial attacks [54, 46] have been shown to change their decision when class-specific features appear in the image, which is a prerequisite for meaningful counterfactuals [6]. RATIO generates better counterfactuals, i.e. the confidence of the counterfactual images obtained by an l_2-adversarial attack tends to be high only after features of the alternative class have appeared. Especially for out-distribution images the difference to AT is pronounced.

Robust, reliable and explainable classifiers. This is the holy grail of machine learning. A model which is accurate and calibrated [20] on the in-distribution, reliably has low confidence on out-distribution inputs, is robust to adversarial manipulation and has explainable decisions. Up to our knowledge there is no model which claims to have all these properties. The closest one we are aware of is the JEM-0 of [19] which is supposed to be robust, detects out-of-distribution samples and has generative properties. They state "JEM does not confidently classify nonsensical images, so instead, ... natural image properties visibly emerge". We show that RATIO gets us closer to this ultimate goal and outperforms JEM-0 in all aspects: accuracy, robustness, (worst-case) out-of-distribution detection, and visual counterfactual explanations.

3 RATIO: Robust, Reliable and Explainable Classifier

In the following we are considering multi-class (image) classification. We have the logits of a classifier $f : [0,1]^d \to \mathbb{R}^K$ where d is the input dimension and K the number of classes. With $\Delta = \{p \in [0,1]^K \mid \sum_{i=1}^K p_i = 1\}$ we denote the predicted probability distribution of f over the labels by $\hat{p} : \mathbb{R}^d \to \Delta$ which is obtained using the softmax function: $\hat{p}_{f,s}(x) = \frac{e^{f_s(x)}}{\sum_{j=1}^K e^{f_j(x)}}$, $s = 1, \ldots, K$. We

further denote the training set by $(x_i, y_i)_{i=1}^N$ with $x_i \in [0,1]^d$ and $y_i \in \{1, \dots, K\}$. As loss we always use the cross-entropy loss defined as

$$L(p, \hat{p}_f) = \sum_{j=1}^{K} p_j \log(\hat{p}_{f,j}), \tag{1}$$

where $p \in \Delta$ is the true distribution and $\hat{p}_f$ the predicted distribution.

3.1 Robustness via Adversarial Training

An adversarial sample of x with respect to some threat model $T(x) \subset \mathbb{R}^d$ is a point $z \in T(x) \cap [0,1]^d$ such that the decision of the classifier f changes for z while an oracle would unambiguously associate z with the class of x. In particular this implies that z shows no meaningful class-associated features of any other class. Formally, let y be the correct label of x, then z is an adversarial sample if

$$\arg\max_{k \neq y} f_k(z) > f_y(x), \qquad z \in [0,1]^d \cap T(x), \tag{2}$$

assuming that the threat model is small enough such that no real class change occurs. Typical threat models are l_p-balls of a given radius ϵ, that is

$$T(x) = B_p(x, \epsilon) = \{z \in \mathbb{R}^d \mid \|z - x\|_p \leq \epsilon\}. \tag{3}$$

The robust test accuracy is then defined as the lowest possible accuracy when every test image x is allowed to be changed to some $z \in T(x) \cap [0,1]^d$. Plain models have a robust test accuracy close to zero, even for "small" threat models.

Several strategies for adversarial robustness have been proposed, but adversarial training (AT) [34] has proven to produce robust classifiers across datasets and network architectures without adding significant computational overhead during inference (compared to randomized smoothing [33, 30, 11]).

The objective of adversarial training for a threat model $T(x) \subset \mathbb{R}^d$ is:

$$\min_f \mathbb{E}_{(x,y)\sim p_{\text{in}}} \Big[\max_{z \in T(x)} L(\mathbf{e}_y, \hat{p}_f(z)) \Big], \tag{4}$$

where $\mathbf{e}_y$ is a one-hot encoding of label y and $p_{\text{in}}(x, y)$ is the training distribution. During training one approximately solves the inner maximization problem in equation 4 via projected gradient descent (PGD) and then computes the gradient wrt f at the approximate solution of the inner problem. The community has put emphasis on robustness wrt l_∞ but recently there is more interest in other threat models e.g. l_2-balls [53, 44, 46]. In particular, it has been noted [54, 46] that robust models wrt an l_2-ball have the property that "adversarial" samples generated within a sufficiently large l_2-ball tend to have image features of the predicted class. Thus they are not "adversarial" samples in the sense defined above as the true class has changed or is at least ambiguous.

The main problem of AT is that robust classifiers suffer from a significant drop in accuracy compared to normal training [54]. This trade-off [47, 50] can be mitigated e.g. via training 50% on clean samples and 50% on adversarial samples at the price of reduced robustness [50] or via semi-supervised learning [55, 39, 9].

3.2 Worst-case OOD detection via Adversarial Training on the Out-distribution

While adversarial training yields robust classifiers, similarly to plain models it suffers from overconfident predictions on out-of-distribution samples. Overconfident predictions are a problem for safety-critical systems as the classifier is not reliably flagging when it operates "out of its specification" and thus its confidence in the prediction cannot be used to trigger human intervention.

In order to mitigate over-confident predictions [21, 25] proposed to enforce low confidence on images from a chosen out-distribution $p_{\text{out}}(x)$. A generic out-distribution would be all natural images and thus [25] suggest the 80 million tiny images dataset [52] as a proxy for this. While [25] consistently reduce confidence on different out-of-distribution datasets, similar to plain training for the in distribution one can again get overconfident predictions by maximizing the confidence in a small ball around a given out-distribution image (adversarial attacks on the out-distribution [21, 35]).

Thus [21] proposed Adversarial Confidence Enhanced Training (ACET) which enforces low confidence in an entire neighborhood around the out-distribution samples which can be seen as a form of AT on the out-distribution:

$$\min_f \mathbb{E}_{(x,y)\sim p_{\text{in}}}\Big[L(\boldsymbol{e}_y, \hat{p}_f(x))\Big] + \lambda\, \mathbb{E}_{(x,y)\sim p_{\text{out}}}\Big[\max_{\|z-x\|_2\leq\epsilon} L(\mathbf{1}/K, \hat{p}_f(z))\Big], \tag{5}$$

where $\mathbf{1}$ is the vector of all ones (outlier exposure [25] has the same objective without the inner maximization for the out-distribution). Different from [21] we use the same loss for in-and out-distribution, whereas they used the maximal log-confidence over all classes as loss for the out-distribution. In our experience the maximal log-confidence is more difficult to optimize, but both losses are minimized by the uniform distribution over the labels. Thus the difference is rather small and we also denote this version as ACET.

3.3 RATIO: Robustness via Adversarial Training on In-and Out-distribution

We propose RATIO: adversarial training on in-and out-distribution. This combination leads to synergy effects where most positive attributes of AT and ACET are fused without having larger drawbacks. The objective of RATIO is given by:

$$\min_f \mathbb{E}_{(x,y)\sim p_{\text{in}}}\Big[\max_{\|z-x\|_2\leq\epsilon_i} L(\boldsymbol{e}_y, \hat{p}_f(z))\Big] + \lambda\, \mathbb{E}_{(x,y)\sim p_{\text{out}}}\Big[\max_{\|z-x\|_2\leq\epsilon_o} L(\mathbf{1}/K, \hat{p}_f(z))\Big], \tag{6}$$

where λ has the interpretation of $\frac{p_o}{p_i}$, the probability to see out-distribution p_o and in-distribution p_i samples at test time. Here we have specified an l_2-threat model for in-and out-distribution but the objective can be adapted to different threat models which could be different for in- and out-distribution. The surprising part of RATIO is that the addition of the out-distribution part can improve the results even on the in-distribution in terms of (robust) accuracy. The reason

is that adversarial training on the out-distribution ensures that spurious features do not change the confidence of the classifier. This behavior generalizes to the in-distribution and thus ACET (adversarial training on the out-distribution) is also robust on the in-distribution (52.3% robust accuracy for l_2 with $\epsilon = 0.5$ on CIFAR10). One problem of adversarial training is overfitting on the training set [43]. Our RATIO has seen more images at training time and while the direct goal is distinct (keeping one-hot prediction on the in-distribution and uniform prediction on out-distribution) both aim at constant behavior of the classifier over the l_2-ball and thus the effectively increased training size improves generalization (in contrast to AT, RATIO has its peak robustness at the end of the training). Moreover, RATIO typically only shows high confidence if class-specific features have appeared which we use in the generative process described next.

4 Visual Counterfactual Explanations

The idea of a counterfactual explanation [56] is to provide the smallest change of a given input such that the decision changes into a desired target class e.g. how would this X-ray image need to look in order to change the diagnosis from X to Y. Compared to sensitivity based explanations [5, 59] or explanations based on feature attributions [4] counterfactual explanations have the advantage that they have an "operational meaning" which couples the explanation directly to the decision of the classifier. On the other hand the counterfactual explanation requires us to specify a metric or a budget for the allowed change of the image which can be done directly in image space or in the latent space of a generative model. However, our goal is that the classifier directly learns what meaningful changes are and we do not want to impose that via a generative model. Thus we aim at visual counterfactual explanations directly in image space with a fixed budget for changing the image. As the decision changes, features of this class should appear in the image (see Figure 2). Normally trained models will not achieve this since non-robust models change their prediction for non-perceptible perturbations [51], see Figure 1. Thus robustness against (l_2-)adversarial perturbations is a necessary requirement for visual counterfactuals and indeed [54, 46] have shown "generative properties" of l_2-robust models.

A *visual counterfactual* for the original point x classified as $c = \underset{k=1,\dots,K}{\arg\max}\, f_k(x)$, a target class $t \in \{1, \dots, K\}$ and a budget ϵ is defined as

$$x^{(t)} = \underset{z \in [0,1]^d,\, \|x-z\|_2 \leq \epsilon}{\arg\max}\ \hat{p}_{f,t}(z), \tag{7}$$

where $\hat{p}_{f,t}(z)$ is the confidence for class t of our classifier for the image z. If $t \neq c$ it answers the counterfactual question of how to use the given budget to change the original input x so that the classifier is most confident in class t. Note that in our definition we include the case where $t = c$, that is we ask how to change the input x classified as c to get even more confident in class c. In Figure 2 we illustrate both directions and show how for robust models class specific image

Table 2: Summary for CIFAR100 and R. ImageNet (see Table 1 for details).

CIFAR100	Plain	OE	ACET	$AT_{0.5}$	$AT_{0.25}$	$R_{0.5}$	$\mathbf{R}_{0.25}$
Acc. ↑	**81.5**	81.4	-	70.6	75.8	69.2	74.4
R. $Acc._{0.5}$ ↑	0.0	0.0	-	43.2	37.3	**45.6**	42.4
ECE ↓	**1.2**	7.2	-	1.3	1.5	3.2	2.0
AUC ↑	84.0	**91.9**	-	75.6	79.4	87.0	86.9
WC $AUC_{1.0}$ ↑	0.4	14.6	-	29.9	24.8	**55.5**	54.5
MMC ↓	51.1	21.8	-	45.8	47.1	24.4	31.0
R.Imagenet	Plain	OE	ACET	$M_{3.5}$	$AT_{3.5}$	$AT_{1.75}$	$R_{3.5}$	$\mathbf{R}_{1.75}$
Acc. ↑	96.6	**97.2**	96.2	90.3	93.5	95.5	93.9	95.5
R. $Acc._{3.5}$ ↑	0.0	0.0	6.2	47.7	47.7	36.7	**49.2**	43.0
ECE ↓	0.6	1.8	0.9	0.7	0.9	0.5	**0.3**	0.7
AUC ↑	92.7	**98.9**	97.74	83.6	84.3	86.5	97.2	97.8
WC $AUC_{7.0}$ ↑	0.0	1.8	87.54	44.2	37.5	16.3	**90.9**	90.6
MMC ↓	67.9	**20.6**	34.85	69.2	75.2	81.8	33.6	32.3

features appear when optimizing the confidence of that class. This shows that the optimization of visual counterfactuals can be done directly in image space.

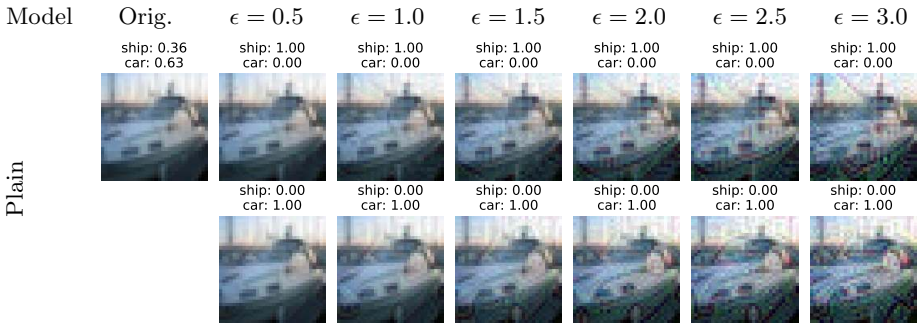


Fig. 1: Failure of a visual counterfactual for a plain model. The targeted attack immediately produces very high confidence in both classes but instead of class features only high-frequency noise appears because plain models are not robust.

5 Experiments

Comparison, Training and Attacks. We validate our approach on SVHN [40], CIFAR10/100 [28] and restricted ImageNet [46]. On CIFAR10 we compare RATIO to a pretrained JEM-0 [19] and the AT model [16] with $l_2 = 0.5$ ($M_{0.5}$) (both not available on the other datasets). As an ablation study of RATIO we train a plain model, outlier exposure (OE) [25], ACET [21] and AT with $l_2 = 0.5$

($\mathrm{AT}_{0.5}$) and $l_2 = 0.25$ ($\mathrm{AT}_{0.25}$), using the same hyperparameters as for our RATIO training. On SVHN we use a ResNet18 architecture for all methods and on the other datasets we use ResNet50, both with standard input normalization. For ACET on CIFAR10 we use ResNet18 since for ResNet50 we could not obtain a model with good worst case OOD performance as the attack seemed to fail at some point during training (on CIFAR100 this was even the case for ResNet18 and thus we omit it from comparison). In general ACET is difficult to train. For RATIO the additional adversarial training on the in-distribution seems to stabilize the training and we did not encounter any problems. As out-distribution for SVHN and CIFAR we use 80 million tiny images [52] as suggested in [25] and for restricted ImageNet the remaining ImageNet classes. For the out-distribution we always use l_2-attacks with radius $\epsilon_o = 1$ for SVHN/CIFAR and $\epsilon_o = 7$ on restricted ImageNet (both ACET and RATIO) whereas on the in-distribution we use $\epsilon_i = 0.25$ and $\epsilon_i = 0.5$ and $\epsilon_i = 1.75$ and $\epsilon_i = 3.5$, respectively (both AT and RATIO). Therefore RATIO/AT models are labeled by ϵ_i. For further training details see the Appendix. For the adversarial attacks on in- and out-distribution we use the recent Auto-Attack [13] which is an ensemble of four attacks, including the black-box Square Attack [2] and three white-box attacks (FAB-attack [14] and AUTO-PGD with different losses). For each of the white-box attacks, a budget of 100 iterations and 5 restarts is used and a query limit of 5000 for Square attack. In [13] they show that Auto-Attack consistently improves the robustness evaluation for a large number of models (including JEM-x).

Calibration on the in-distribution. With RATIO we aim for reliable confidence estimates, in particular no overconfident predictions. In order to have comparable confidences for the different models we train, especially when we check visual counterfactuals or feature generation, we first need to "align" their confidences. We do this by minimizing the expected calibration error (ECE) via temperature rescaling [20]. Note that this rescaling does not change the classification and thus has no impact on (robust) accuracy and only a minor influence on the (worst case) AUC values for OOD-detection. For details see the Appendix.

(Robust) Accuracy on the in-distribution. Using Auto-Attack [13] we evaluate robustness on the full test set for both CIFAR and r. Imagenet and 10000 test samples for SVHN. Tables 1 and 2 contain (robust l_2) accuracy, detailed results, including l_∞ attacks, can be found in the Appendix. On CIFAR10, RATIO achieves significantly higher robust accuracy than AT for l_2-and l_∞-attacks. Thus the additional adversarial training on the out-distribution with radius $\epsilon_o = 1$ boosts the robustness on the in-distribution. In particular, $\mathrm{RATIO}_{0.25}$ achieves better l_2-robustness than $\mathrm{AT}_{0.5}$ and $\mathrm{M}_{0.5}$ at $\approx 2.7\%$ higher clean accuracy. In addition, $\mathrm{R}_{0.5}$ yields new state-of-the-art l_2-robust accuracy at radius 0.5 (see [13] for a benchmark) while having higher test accuracy than $\mathrm{AT}_{0.5}$, $\mathrm{M}_{0.5}$. Moreover, the l_2-robustness at radius 1.0 and the l_∞-robustness at 8/255 is significantly better. Interestingly, although ACET is not designed to yield adversarial robustness on the in-distribution, it achieves more than 50% robust accuracy for $l_2 = 0.5$ and outperforms JEM-0 in all benchmarks. However, as our goal is to have a model which is both robust and accurate, we recommend to use $R_{0.25}$ for

CIFAR10 which has a drop of only 2.6% in test accuracy compared to a plain model while having similar robustness to $M_{0.5}$ and $AT_{0.5}$. Similar observations as for CIFAR10 hold for CIFAR100 and for Restricted ImageNet, see Table 2, even though for CIFAR100 AT and RATIO suffer a higher loss in accuracy. On SVHN, RATIO outperforms AT in terms of robust accuracy trained with the same l_2-radius but the effect is less than for CIFAR10. We believe that this is due to the fact that that the images obtained from the 80 million tiny image dataset (out distribution) do not reflect the specific structure of SVHN numbers which makes (worst case) outlier detection an easier task. This is supported by the fact that ACET achieves better clean accuracy on SVHN than both OE and the plain model while it has worse clean accuracy on CIFAR10.

Visual Counterfactual Generation. We use 500 step Auto-PGD [13] for a targeted attack with the objective in equation 7. However, note that this non-convex optimization problem has been shown to be NP-hard [27]. In Figure 2, 3 and 4 and in the Appendix we show generated counterfactuals for all datasets. For CIFAR10 $AT_{0.5}$ performs very similar to $RATIO_{0.25}$ in terms of the emergence of class specific image features. In particular, we often see the appearance of characteristic features such as pointed ears for cats, wheels for cars and trucks, large eyes for both cats and dogs and the antlers for deers. JEM-0 and ACET perform worse but for both of them one observes the appearance of image features. However, particularly the images of JEM-0 have a lot of artefacts. For SVHN $RATIO_{0.25}$ on average performs better than $AT_{0.25}$ and ACET. It is interesting to note that for both datasets class-specific features emerge already for an l_2-radius of 1.0. Thus it seems questionable if l_2-adversarial robustness beyond a radius of 1.0 should be enforced. Due to the larger number of classes, CIFAR100 counterfactuals are of slightly lower quality. For Restricted ImageNet the visual counterfactuals show class-specific features but can often be identified as synthetic due to misaligned features.

Reliable Detection of (Worst-case) Out-of-Distribution Images. A reliable classifier should assign low confidence to OOD images. This is not the case for plain models and AT. As the 80 million tiny image dataset has been used for training for ACET and RATIO (respectively other ImageNet classes for Restricted ImageNet), we evaluate the discrimination of in-distribution versus out-distribution on other datasets as in [35], see the Appendix for details. We use $\max_k \hat{p}_{f,k}(x)$ as feature to discriminate in-and out-distribution (binary classification) and compute the AUC. However, it has been shown that even state-of-the-art methods like outlier exposure (OE) suffer from overconfident predictions if one searches for the most confident prediction in a small neighborhood around the the out-distribution image [35]. Thus we also report the worst-case AUC by maximizing the confidence in an l_2-ball of radius 1.0 (resp. 7.0 for R. ImageNet) around OOD images via Auto-PGD [13] with 100 steps and 5 random restarts. Figure 5 further shows that while RATIO behaves similar to AT around samples from the data distribution, which explains similar counterfactuals, it has a flatter confidence profile around out-distribution samples.

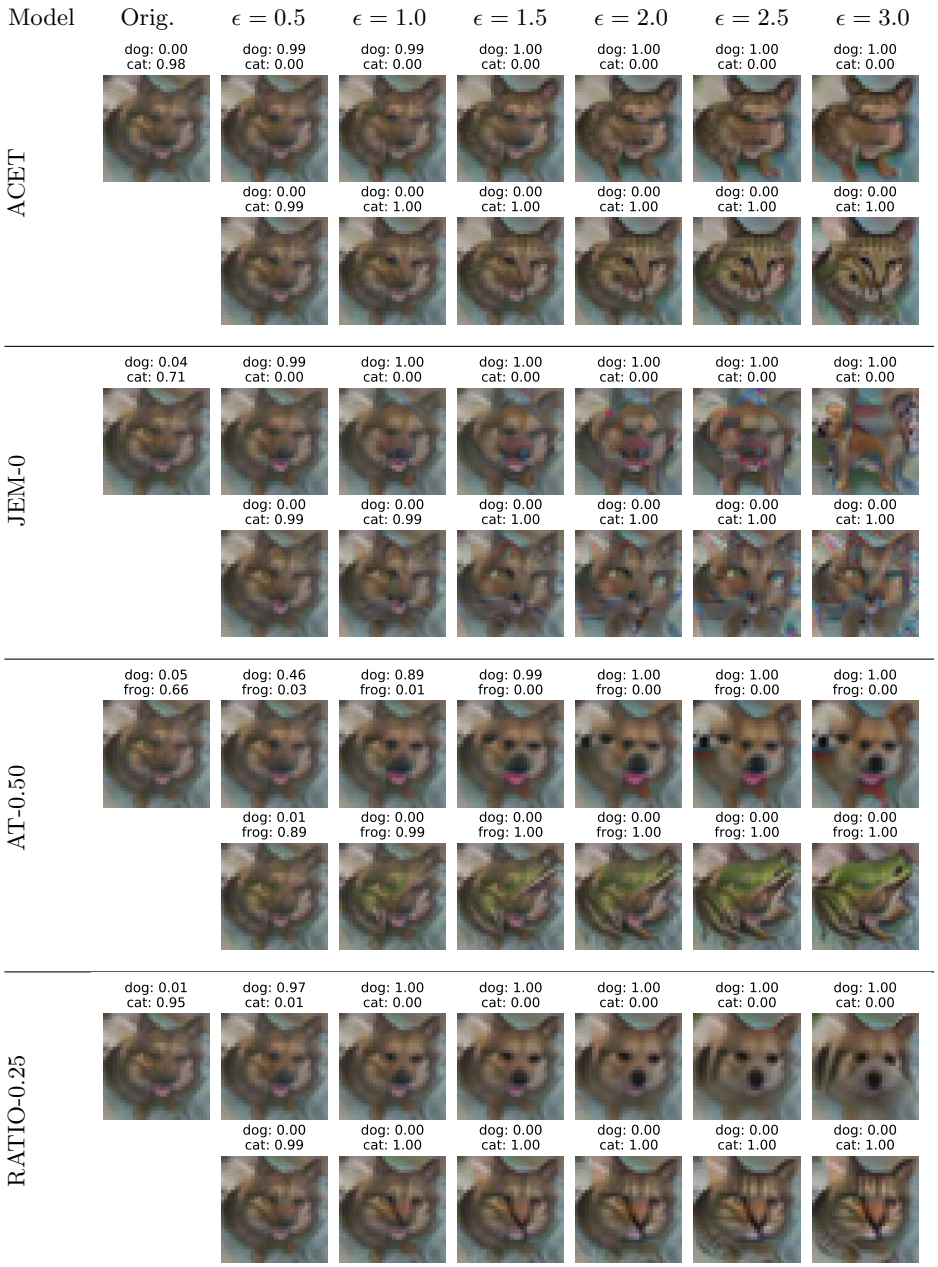


Fig. 2: **Visual Counterfactuals (CIFAR10):** The dog image on the left is misclassified by all models (confidence for true and predicted class are shown). The top row shows visual counterfactuals for the correct class (how to change the image so that it is classified as dog) and the bottom row shows how to change the image in order to increase the confidence in the wrong prediction for different budgets of the l_2-radius ($\epsilon = 0.5$ to $\epsilon = 3$). More examples are in the appendix.

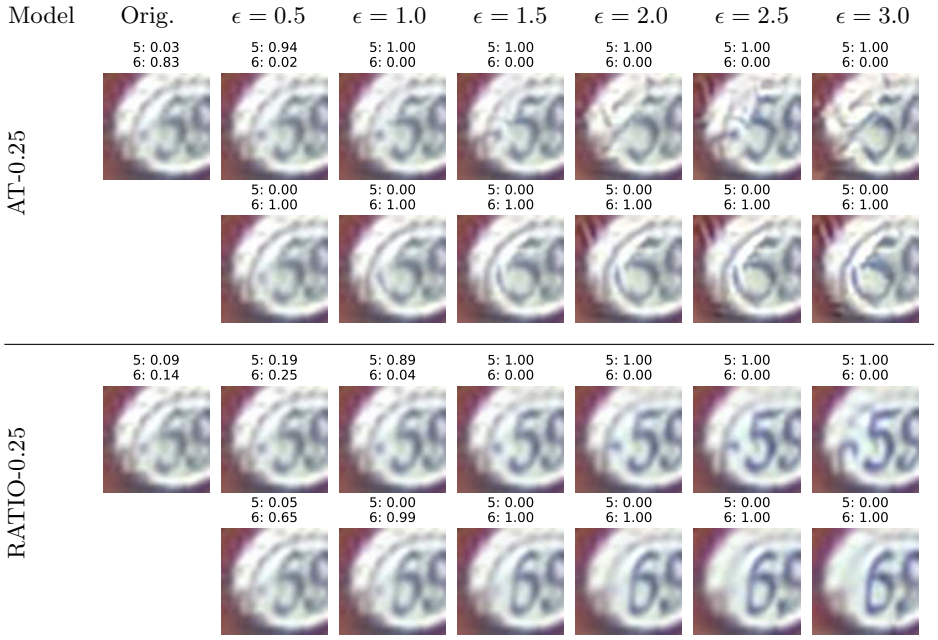


Fig. 3: **Visual Counterfactuals (SVHN):** The 5 on the left is misclassified by all models. We show counterfactuals for the true class the predicted class (see Figure 2). RATIO consistently produces samples with fewer artefacts than AT.

on 1024 points from each out-distribution (300 points for LSUN_CR). Using the worst case confidences of these points we find empirical upper bounds on the worst-case AUC under our threat model. We report both the average-case AUCs as well as the worst-case AUCs in the Appendix.The average AUC over all OOD datasets is reported in Tables 1 and 2. The AT-model of Madry et. al ($M_{0.5}$) perform worse than the plain model even on the average case task. However, we see that with our more aggressive data augmentation this problem is somewhat alleviated ($AT_{0.5}$ and $AT_{0.25}$). As expected ACET, has good worst-case OOD performance but is similar to the plain model for the average case. JEM-0 has bad worst-case AUCs and we cannot confirm the claim that "JEM does not confidently classify nonsensical images"[19]. As expected, OE has state-of-the-art performance on the clean task but has no robustness on the out-distribution, so it fails completely in this regime. Our RATIO models show strong performance on all tasks and even outperform the ACET model which shows that adversarial robustness wrt the in-distribution also helps with adversarial robustness on the out-distribution. On SVHN the average case OOD task is simple enough that several models achieve near perfect AUCs, but again only ACET and our RATIO models manage to retain strong performance in the worst case setting. The worst-case AUC of AT models is significantly worse than that of ACET and RATIO.

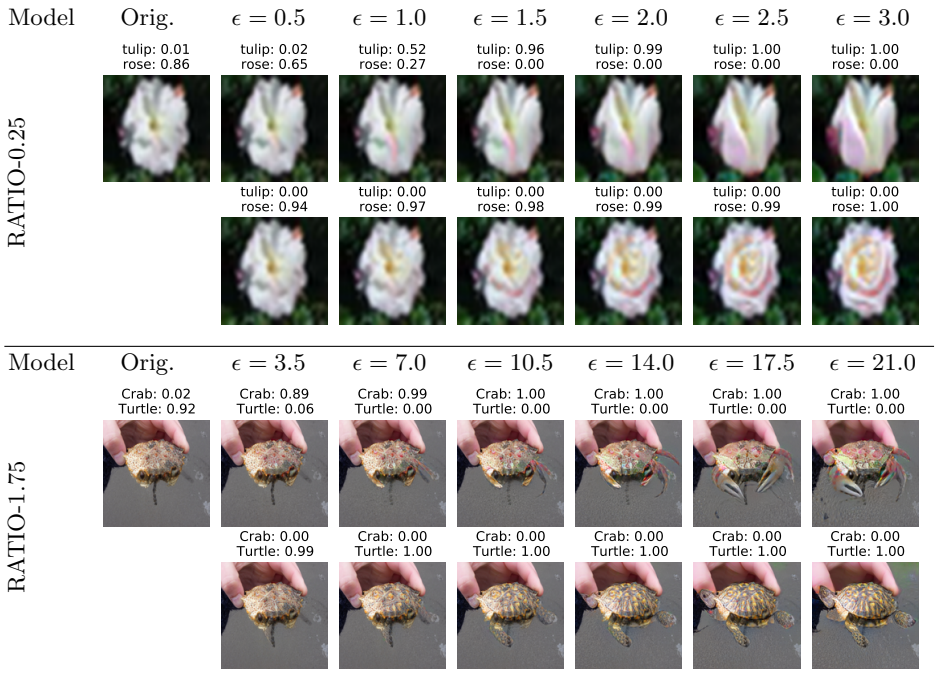
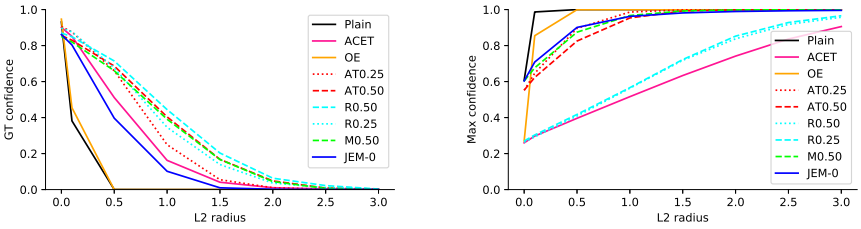


Fig. 4: **Visual Counterfactuals** top: RATIO-0.25 for CIFAR100 and bottom: RATIO-1.75 for RestrictedImageNet.

Feature Generation on OOD images. Finally, we test the abilities to generate image features with a targeted attack on OOD images (taken from 80m tiny image dataset resp. ImageNet classes not belonging to R. ImageNet). The setting is similar to the visual counterfactuals. We take some OOD image and then optimize the confidence in the class which is predicted on the OOD image. The results can be found in Figure 7 and 6 and additional samples are attached in the Appendix. For CIFAR10 all methods are able to generate image features of the class but the predicted confidences are only reasonable for ACET and $\text{RATIO}_{0.25}$ whereas $\text{AT}_{0.5}$ and JEM-0 are overconfident when no strong class features are visible. This observation generalizes to SVHN and mostly CIFAR100 and r. Imagenet, i.e. RATIO generally has the best OOD-confidence profile.
Summary. In summary, in Table 1 and 2 we can see that $\text{RATIO}_{0.25}$ resp. $\text{RATIO}_{1.75}$ is except for CIFAR100 the only model which has no clear failure case. Here the subjective definition of a failure case (highlighted in red) is an entry which is "significantly worse" than the best possible in this metric. Thus we think that RATIO succeeds in being state-of-the-art in generating a model which is accurate, robust, has reliable confidence and is able to produce meaningful visual counterfactuals. Nevertheless RATIO is not perfect and we discuss failure cases of all models in the Appendix.



(a) ID worst-case confidence (b) OD worst-case confidence

Fig. 5: (a) Mean confidence in true label as a function of the attack l_2-radius around CIFAR10 test images. RATIO and AT0.5 have a reasonable decay of the confidence. (b) Mean of maximal confidence around OD-data (tiny images) over the attack l_2-radius. All methods except RATIO and ACET are overconfident.

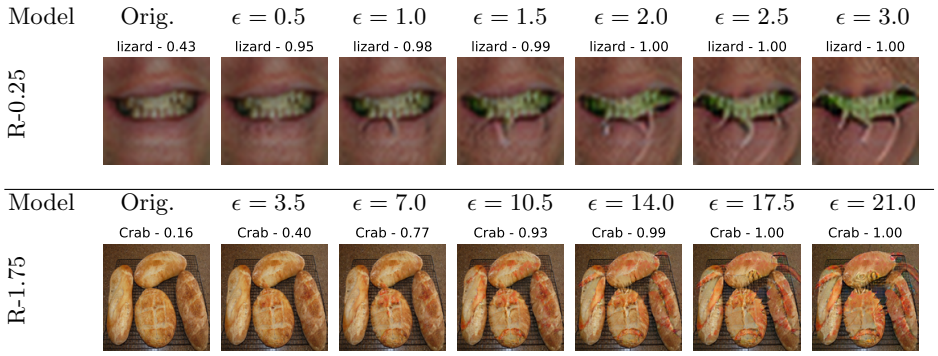


Fig. 6: **Feature Generation for out-distribution images** top: RATIO-0.25 for CIFAR100 and bottom: RATIO-1.75 for R.ImageNet

6 Conclusion and Outlook

We have shown that adversarial robustness on in-distribution and out-distribution (as a proxy of all natural images) gets us closer to a classifier which is accurate, robust, has reliable confidence estimates and is able to produce visual counterfactual explanations with strong class specific image features. For the usage in safety-critical in systems it would be ideal if these properties can be achieved in a provable way which remains an open problem.

Acknowledgements

M.H and A.M. acknowledge support by the BMBF Tübingen AI Center (FKZ: 01IS18039A) and by DFG TRR 248, project number 389792660 and the DFG Excellence Cluster Machine Learning -New Perspectives for Science, EXC 2064/1, project number 390727645. A.M. thanks the IMPRS for Intelligent Systems.

Fig. 7: **Feature Generation for out-distribution images (CIFAR10 (top), SVHN (bottom)):** targeted attacks towards the class achieving highest confidence on original image for different budgets of the l_2-radius ranging from $\epsilon = 0.5$ to $\epsilon = 3$. RATIO-0.25 generates the visually best images and in particular has reasonable confidence values for its decision. While AT-0.5/AT-0.25 generates also good images it is overconfident into the target class.

References

1. Alayrac, J.B., Uesato, J., Huang, P.S., Fawzi, A., Stanforth, R., Kohli, P.: Are labels required for improving adversarial robustness? In: NeurIPS (2019)
2. Andriushchenko, M., Croce, F., Flammarion, N., Hein, M.: Square attack: a query-efficient black-box adversarial attack via random search. In: ECCV (2020)
3. Athalye, A., Carlini, N., Wagner, D.A.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: ICML (2018)
4. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One **10**(7), e0130140 (2015)
5. Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., Müller, K.R.: How to explain individual classification decisions. Journal of Machine Learning Research (JMLR) **11**, 1803–1831 (2010)
6. Barocas, S., Selbst, A.D., Raghavan, M.: The hidden assumptions behind counterfactual explanations and principal reasons. In: FAT (2020)
7. Bitterwolf, J., Meinke, A., Hein, M.: Provable worst case guarantees for the detection of out-of-distribution data. arXiv:2007.08473 (2020)
8. Carlini, N., Wagner, D.: Adversarial examples are not easily detected: Bypassing ten detection methods. In: ACM Workshop on Artificial Intelligence and Security (2017)
9. Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J.C., Liang, P.S.: Unlabeled data improves adversarial robustness. In: NeurIPS (2019)
10. Chang, C.H., Creager, E., Goldenberg, A., Duvenaud, D.: Explaining image classifiers by counterfactual generation. In: ICLR (2019)
11. Cohen, J.M., Rosenfeld, E., Kolter, J.Z.: Certified adversarial robustness via randomized smoothing. In: NeurIPS (2019)
12. Croce, F., Andriushchenko, M., Hein, M.: Provable robustness of relu networks via maximization of linear regions. In: AISTATS (2019)
13. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: ICML (2020)
14. Croce, F., Hein, M.: Minimally distorted adversarial examples with a fast adaptive boundary attack. In: ICML (2020)
15. Dong, Y., Su, H., Zhu, J., Bao, F.: Towards interpretable deep neural networks by leveraging adversarial examples (2017), arXiv preprint, arXiv:1708.05493
16. Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D.: Robustness (python library) (2019), https://github.com/MadryLab/robustness
17. Gowal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelovic, R., Mann, T.A., Kohli, P.: On the effectiveness of interval bound propagation for training verifiably robust models (2018), preprint, arXiv:1810.12715v3
18. Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., Lee, S.: Counterfactual visual explanations. In: ICML (2019)
19. Grathwohl, W., Wang, K.C., Jacobsen, J.H., Duvenaud, D., Norouzi, M., Swersky, K.: Your classifier is secretly an energy based model and you should treat it like one. ICLR (2020)
20. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.: On calibration of modern neural networks. In: ICML (2017)
21. Hein, M., Andriushchenko, M., Bitterwolf, J.: Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In: CVPR (2019)

22. Hendricks, L.A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., Darrell, T.: Generating visual explanations. In: ECCV (2016)
23. Hendricks, L.A., Hu, R., Darrell, T., Akata, Z.: Grounding visual explanations. In: ECCV (2018)
24. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: ICLR (2017)
25. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. In: ICLR (2019)
26. Hendrycks, D., Lee, K., Mazeika, M.: Using pre-training can improve model robustness and uncertainty. In: ICML. pp. 2712–2721 (2019)
27. Katz, G., Barrett, C., Dill, D., Julian, K., Kochenderfer, M.: Reluplex: An efficient smt solver for verifying deep neural networks. In: CAV (2017)
28. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
29. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521** (2015)
30. Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., Jana, S.: Certified robustness to adversarial examples with differential privacy. In: IEEE Symposium on Security and Privacy (SP) (2019)
31. Lee, K., Lee, H., Lee, K., Shin, J.: Training confidence-calibrated classifiers for detecting out-of-distribution samples. In: ICLR (2018)
32. Leibig, C., Allken, V., Ayhan, M.S., Berens, P., Wahl, S.: Leveraging uncertainty information from deep neural networks for disease detection. Scientific Reports **7** (2017)
33. Li, B., Chen, C., Wang, W., Carin, L.: Certified adversarial robustness with additive noise. In: NeurIPS (2019)
34. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Valdu, A.: Towards deep learning models resistant to adversarial attacks. In: ICLR (2018)
35. Meinke, A., Hein, M.: Towards neural networks that provably know when they don't know. In: ICLR (2020)
36. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1 – 38 (2019)
37. Mirman, M., Gehr, T., Vechev, M.: Differentiable abstract interpretation for provably robust neural networks. In: ICML (2018)
38. Mosbach, M., Andriushchenko, M., Trost, T., Hein, M., Klakow, D.: Logit pairing methods can fool gradient-based attacks. In: NeurIPS 2018 Workshop on Security in Machine Learning (2018)
39. Najafi, A., Maeda, S.i., Koyama, M., Miyato, T.: Robustness to adversarial perturbations in learning from incomplete data. In: NeurIPS (2019)
40. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning (2011)
41. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: CVPR (2015)
42. lvaro Parafita, Vitri, J.: Explaining visual models by causal attribution. In: ICCV Workshop on XCAI (2019)
43. Rice, L., Wong, E., Kolter, J.Z.: Overfitting in adversarially robust deep learning. In: ICML (2020)
44. Rony, J., Hafemann, L.G., Oliveira, L.S., Ayed, I.B., Sabourin, R., Granger, E.: Decoupling direction and norm for efficient gradient-based L2 adversarial attacks and defenses. In: CVPR (2019)

45. Samangouei, P., Saeedi, A., Nakagawa, L., Silberman, N.: Explaingan: Model explanation via decision boundary crossing transformations. In: ECCV (2018)
46. Santurkar, S., Tsipras, D., Tran, B., Ilyas, A., Engstrom, L., Madry, A.: Computer vision with a single (robust) classifier. In: NeurIPS (2019)
47. Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., Madry, A.: Adversarially robust generalization requires more data. In: NeurIPS (2018)
48. Schott, L., Rauber, J., Bethge, M., Brendel, W.: Towards the first adversarially robust neural network model on mnist. In: ICLR (2019)
49. Sehwag, V., Bhagoji, A.N., Song, L., Sitawarin, C., Cullina, D., Chiang, M., Mittal, P.: Better the devil you know: An analysis of evasion attacks using out-of-distribution adversarial examples. preprint, arXiv:1905.01726 (2019)
50. Stutz, D., Hein, M., Schiele, B.: Disentangling adversarial robustness and generalization. In: CVPR (2019)
51. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: ICLR. pp. 2503–2511 (2014)
52. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: A large data set for nonparametric object and scene recognition. IEEE transactions on pattern analysis and machine intelligence **30**(11), 1958–1970 (2008)
53. Tramr, F., Boneh, D.: Adversarial training and robustness for multiple perturbations. In: NeurIPS (2019)
54. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., Madry, A.: Robustness may be at odds with accuracy. In: ICLR (2019)
55. Uesato, J., Alayrac, J.B., Huang, P.S., Stanforth, R., Fawzi, A., Kohli, P.: Are labels required for improving adversarial robustness? In: NeurIPS (2019)
56. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harvard Journal of Law and Technology **31**(2), 841–887 (2018)
57. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: CVPR (2018)
58. Wong, E., Schmidt, F., Metzen, J.H., Kolter, J.Z.: Scaling provable adversarial defenses. In: NeurIPS (2018)
59. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV (2014)
60. Zhang, H., Yu, Y., Jiao, J., Xing, E.P., Ghaoui, L.E., Jordan, M.I.: Theoretically principled trade-off between robustness and accuracy. In: ICML (2019)
61. Zhu, J.Y., Krähenbühl, P., Shechtman, E., Efros, A.A.: Generative visual manipulation on the natural image manifold. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV (2016)