Improving Model Robustness with Transformation-Invariant Attacks

Houpu Yao\(^1\), Zhe Wang\(^{1,2}\), Guangyu Nie\(^1\), Yassine Mazboudi\(^1\), Yezhou Yang\(^1\) and Yi Ren\(^1\)
\(^1\)Arizona State University
\(^2\)Beihang University
{hyao23,gniel,ymazboud,yz.yang,yiren}@asu.edu wangzhewz@buaa.edu.cn

Abstract—Vulnerability of neural networks under adversarial attacks has raised serious concerns and extensive research. Recent studies suggested that model robustness relies on the use of robust features, i.e., features with strong correlation with labels, and that data dimensionality and distribution affect the learning of robust features. On the other hand, experiments showed that human vision, which is robust against adversarial attacks, is invariant to natural input transformations. Drawing on these findings, this paper investigates whether constraints on transformation invariance, including image cropping, rotation, and zooming, will force image classifiers to learn and use robust features and in turn acquire better robustness. Experiments on MNIST and CIFAR10 show that transformation invariance alone has limited effect. Nonetheless, models adversarially trained on cropping-invariant attacks, in particular, can (1) extract more robust features, (2) have significantly better robustness than the state-of-the-art models from adversarial training, and (3) require less training data.

I. INTRODUCTION

While deep neural networks achieved near-human performance on various machine perception tasks, it is found that these models can be very sensitive to small but carefully designed input perturbations [1]–[3], thus allowing the attackers to fool a machine in targeted ways by reverse engineering the inputs. Recent studies have demonstrated potential risks in applying neural networks for classification [4]–[7], detection and segmentation [8], [9], image retrieval [10], and reinforcement learning [11], [12]. Furthermore, it has also been demonstrated that these attacks are successful under real-world settings [13]–[16], posing much threat to safety-critical applications.

The vulnerability of neural networks under adversarial attacks has led to rapid development of theories and algorithms for attack and defense mechanisms. However, most existing defense mechanisms are later shown to be vulnerable to stronger attacks [16], [17]. It is further revealed in Tsipras et al. [18] that there exists an intrinsic tradeoff between model robustness and standard accuracy, which may explain the limited robustness currently achieved, e.g., on CIFAR10 [18] and ImageNet [19].

Nonetheless, Tsipras et al. also suggested that model robustness relies on the use of robust features, i.e., the ones with strong correlation with labels; and that training driven by standard accuracy tends to bias the model towards non-robust features. Further, [20] revealed the dependency of model robustness on input dimensionality given a fixed data size, suggesting that improvements in robustness could be achieved through dimension-reducing transformations of the input space. This result is consistent with the more recent finding that transformations of input distributions (e.g., image saturation and edge detector) cause a significant change in model robustness within the transformed distributions [21]. Collectively, these studies suggested that avoiding the use of non-robust features in decision making may lead to improved model robustness, without sacrificing standard accuracy. On the other hand, experiments have shown that human vision, which is robust to adversaries, is invariant to mild natural input transformations [22]. Drawing on these findings, it is reasonable to question whether constraints on transformation invariance may force image classifiers to learn and rely on robust features only and, in turn, acquire better robustness.

This paper is thus motivated to investigate the effectiveness of a specific set of transformations, including cropping, rotation, and zooming, on model robustness. Our investigation is based on an image classifier that passes an ensemble of transformations of the input through shared copies of a network, before aggregating network outputs for decision making (Fig. 2). We hypothesize that the transformations under investigation are content preserving, i.e., the transformed images can still be correctly classified by human beings, and incorporate this hypothesis in adversarial training, where transformation-invariant adversaries are sought for during the attack phases. A comparison between the proposed method and the standard adversarial training is illustrated in Fig. 1.

Experiments on MNIST [23] and CIFAR10 [24] lead to the following key findings:

1) Imposing transformation invariance on clean or adversarial training does not improve model robustness;

2) Adversarial training using transformation-invariant at-

Fig. 1. Comparison between different training methods: (a) Vanilla, (b) Transformation-invariant, (c) Adversarial, and (d) Proposed.
Adversarial training through robust optimization remains one of the few defense mechanisms that do not suffer gradient obfuscation, and so far achieves the best empirical robustness under $l_\infty$ attacks. Adversarial training can be formulated as the following min-max problem:

$$\min_{\theta} \mathbb{E}_{x,y \sim D} \left[ \max_{x' \in \mathcal{N}_r(x)} L(y, f(x', \theta)) \right].$$  \hspace{1cm} (1)

Eq. (1) is hard to solve due to the nonconcavity of the inner (attack) problem and the limited capacity of $f$. A common practice is to iteratively find adversaries and train the network on these adversaries [13], [25]. The state-of-the-art approaches use PGD attack during the attack phases due to its effectiveness.

II. PROPOSED METHOD

The proposed learning architecture passes an ensemble of transformations of an input through copies of a shared network, before aggregating the network outputs. Below we introduce details on the architecture and its implementation.

a) Decision making: Let the shared network be $f(\cdot, \theta)$, and the set of input transformations be $\mathcal{T}$. The probabilistic prediction of an input $x$ follows:

$$y_{pred} = \frac{1}{|T|} \sum_{T \in \mathcal{T}} f(T(x), \theta).$$  \hspace{1cm} (2)

b) Aggregated loss: The aggregated loss for data $(x, y)$ is defined as:

$$J(x, y; \theta) = \sum_{T \in \mathcal{T}} L(y, f(T(x), \theta))$$  \hspace{1cm} (3)

For classification tasks, $L(\cdot, \cdot)$ is the cross-entropy. It should be noted that an alternative loss is $L(y, y_{pred})$. We use Eq. (3) to avoid numerical issues in computing the log of sum of exponential terms. Our experiments showed that model robustness is not sensitive to the choice between these two loss definitions. For standard training, we solve $\min_{\theta} \mathbb{E}_{x,y \sim D} [J(x, y; \theta)].$
c) **Adversarial training under robust attacks:** We use robust untargeted PGD attacks for training, which follows the following min-max problem:

$$
\min_{\theta} \mathbb{E}_{x,y \sim D} \max_{a^{adv} \in \mathcal{N}(x)} J(x^{adv}, y; \theta)
$$

The inner problem in Eq. (4) is an attack robust against all transformations, and is similar to the Expectation of Transformation (EoT) method [15], [17]. However, to our best knowledge, EoT has not yet been incorporated into adversarial training. We solve Eq. (4) using PGD attacks. The attack parameters are summarized in Tab. I where $\alpha$ and $t$ denote the step size and number of gradient descent steps in one attack, respectively.

| Dataset   | $\epsilon$ | $\alpha$ | $t$ |
|-----------|-------------|----------|-----|
| MNIST     | 0.3         | 0.01     | 40  |
| CIFAR10   | 8/255       | 2/255    | 7   |

### d) Network architecture:
Our network for MNIST and CIFAR10 follow the wider networks in [25].

### e) Transformations:
The input transformations we tested in this paper include cropping, rotation, and zooming. For cropping, we drop out connections from the cropped pixels to the next layer. For rotation, we use the built-in rotation function from TensorFlow. For zooming, we first apply cropping and then resize the image back to the original size through bilinear interpolation.

## III. Results and Discussions

### A. Evaluation of Empirical Network Robustness

#### a) White-box robustness:
The white-box robustness of our model is compared with Madry’s approach on MNIST and CIFAR10 under FGM, BIM, PGD, and C&W attacks. For fair comparison, the attack parameters for test also follows [25]. For all attacks, we choose $\epsilon = 0.3$ for MNIST and $\epsilon = 8/255$ for CIFAR10, respectively. PGD and BIM attack parameters follow Tab. I. For C&W attack, we use learning rate 0.2, and 40 steps. For our model, we use 9 crops and a cropping size of 20 for MNIST, and 8 crops and a cropping size of 28 for CIFAR10. The locations of the crops are shown in Fig. 2b. The number of crops in the case of CIFAR10 is limited by the fact that we only have 8 GPUs in parallel, each of which handles one copy of the shared network.

We compared our model with Madry’s on standard and adversarial training. The white-box robustness for these models are summarized in Tab. II, together with accuracy on clean test data. With transformation-invariant adversarial training, our model is able to achieve comparable or higher accuracy than the baseline both with and without adversarial training. More importantly, our model is able to achieve significantly higher robustness under all tested attacks. In particular, the proposed method improves white-box PGD robustness from 93.2% to 95.7% on MNIST, and from 47.3% to 54.4% on CIFAR10.

In addition, we compare the white-box robustness of our model with Madry’s under PGD attacks beyond the attack bounds used during adversarial training. For MNIST, we test $\epsilon \in [0, 1]$ where $\epsilon = 1$ represents the maximal $l_{\infty}$ attack strength. For CIFAR10, we test $\epsilon \in [0, 35/255]$. The comparisons are shown in Fig. 3. The proposed method exhibits consistently higher robustness under all attack bounds.

![Fig. 3. Model robustness for a range of attack bounds. The attack bounds used for training are marked as black vertical lines.](image-url)

### b) Sensitivity of hyper-parameters on white-box robustness:
We study the influence of the number of input crops and the crop size on model robustness using MNIST: For the former, we fix the cropping size to 20 and vary the number of crops from 1 to 64. The locations of crops are shown in Fig. 2b. For the latter, we fix the number of crops to 9 and vary the size of the crops from 12 to 24. Both training and test use $\epsilon = 0.3$. The white-box PGD accuracy of these models are summarized in Tab. III.

### TABLE III

| croppingsize | 12 | 16 | 20 | 24 | 28 |
|--------------|----|----|----|----|----|
| clean testing| 98.3|98.0|99.2|99.1|98.4|
| PGD white-box | 92.1|95.1|95.7|96.1|96.1|

Increasing the number of crops helps to improve both clean and adversarial accuracy, with diminishing effect. On the other hand, a sweet spot exists for the crop size: Larger cropping sizes tend to improve clean test accuracy, yet inevitably lead to reduced number of crops and robustness, while making the size too small will reduce the clean test accuracy significantly, which upper bounds the robustness. For the remaining experiments, we use 9 crops each with size 20 for MNIST and 8 crops with size 28 for CIFAR10.

### c) Effect of transformation ensemble on model robustness:
To better understand the effect of transformation ensemble on model robustness, we further conduct an ablation study where we remove the ensemble effect during training and test phases separately and monitor how the robustness changes in each of the cases.
No ensemble in the test phase: Here we train the model as proposed, and test the white box robustness of each individual network copies in the ensemble. The copies are only different in their input transformation layers. The robustness of individual copies range from 88.7% to 95.3% on MNIST, and 52.6% to 54.1% on CIFAR10 (Fig. 4). These values are lower than the robustness with ensemble (95.7% and 54.4%), suggesting that ensemble is effective during test.

No ensemble in the training phase: We now investigate the effect of the ensemble in adversarial training on model robustness. Specifically, we remove the influence of ensemble on the generation of adversaries, in which case the training becomes standard adversarial training with training data replaced by their cropped copies. When using nine cropping windows, this leads to a 9-fold data augmentation. During adversarial training, this experiment setting results in adversaries that are not necessarily transformation invariant. In the test phase, we perform the same ensemble operation as in Eq. 2. On MNIST, these settings lead to a model robustness of 93.4% and clean test accuracy of 98.7%, which is comparable to Madry’s and worse than the proposed.

This experiment reveals the critical role of transformation-invariant attacks in improving robustness from standard adversarial training. We conjecture that when the model is attacked by individual transformations, the resulting adversaries may lead to contradictory gradient directions for model refinement, i.e., refining the model with respect to attacks for one particular transformation may not help (or even worsen) the robustness under other transformations.

d) Effect of rotation and zooming: Investigation on the effect of input rotation and zooming leads to mixed results (Table IV): On MNIST, rotation yields more robust models while zooming does not; on CIFAR10, model robustness improves with mild rotation angles (maximum 4 degrees); yet larger angles (maximum 30 degrees) and zooming show limited effects. It is worth noting that combining different transformations may lead to extra improvement in robustness. Specifically, we tested the combination of two models on MNIST, trained separately with cropping (cropping size of 20, 9 crops) and rotation (4 orientations). The model reaches 97.2% white-box robustness, while the individual models have 95.7% with cropping and 95.5% with rotation.

Challenge with rotation and zooming on CIFAR10: It is interesting to note that rotation augmentation on CIFAR10 only works well under small rotation angles. The robustness and standard accuracy will be reduced to 48.1% and 78.9% if we increase the maximum rotation angle from 4 degrees to 30 degrees. We conjecture that this is because larger rotation leads to an increased number of robust features for CIFAR10 (e.g., the same image patterns of different angles). With the fixed network capacity, this may lead to unsuccessful learning. Such phenomenon, however, does not have significant influence on the robust features on MNIST (which are strokes of different orientations, see Fig. 5), and thus allows a network with fixed capacity to perform relatively equally across cropping, rotation, and zooming. We thus believe that increasing the network capacity or building in rotation invariance may help improving the robustness and accuracy at a larger rotation angles, but this investigation will be left as future work due to the resultant high training costs.

Convergence issue with saturation: A recent study discovered the sensitivity of model robustness to input distribution [21]. Inspired by this discovery, we tested the effect of saturation on model robustness in the original input space. However, our current implementation of robust adversarial training with image saturation suffers from explosion of the adversarial gradient, which is intrinsic to the transformation.

### Table II

| Network | MNIST | CIFAR10 |
|---------|-------|---------|
| Bnat    | 95.7  | 92.1 (98.3) |
| Cnat(Ours) | 95.6 | 99.3 |

| Network | MNIST | CIFAR10 |
|---------|-------|---------|
| Badv    | 95.2 | 95.2 |
| Cadv(Ours) | 95.6 | 99.3 |

### Table IV

| TRANSFORMATION | MNIST | CIFAR10 |
|----------------|-------|---------|
| cropping       | 95.1 (99.0) | 95.7 (99.2) |
| rotation       | 95.5 (99.1) | 96.1 (99.3) |
| zooming        | 93.6 (99.2) | 93.6 (99.2) |

| TRANSFORMATION | MNIST | CIFAR10 |
|----------------|-------|---------|
| cropping       | 53.3 (83.1) | 54.4 (87.9) |
| rotation       | 53.8 (87.8) | 53.8 (87.8) |

Fig. 4. White-box robustness of individual network copies within the ensemble under PGD attacks. Locations in the grid correspond to cropping locations.
A smooth approximation of the saturation operation needs to be introduced before a proper evaluation of its effect on model robustness can be performed.

e) Learning efficiency: We train the proposed and Madry’s models with increasing training data sizes and compare model robustness and accuracy along the data size. PGD attacks with parameters from Tab.III are applied. Results are shown in Tab.V. Our models consistently require less training data for the same robustness or accuracy levels. Considering that cropping is a dimension-reducing transformation, the result here is consistent with the analysis from [20] that the sample complexity (i.e., the data size) for reaching a certain level of model robustness is related to the input dimensionality.

f) Gradient obfuscation: On the last analysis on white-box attacks, we show that our model does not utilize gradient obfuscation. As discussed in [17], models with gradient obfuscation, i.e., gradient masking, shattering, and explosion or vanishing, can be attacked by tailored attacks. Our model does not create shattering since it has no randomness; it also does not create gradient explosion or vanishing since it does not contain long recurrence. Two pieces of evidence show that our model does not create gradient masking: First, the white-box accuracy of our model is higher than its black-box accuracy; second, from the loss landscapes around random test points shown in Fig. 5, the gradient directions, in comparison with random orthogonal directions, are informative and smooth. Therefore, the reported robustness improvements are reliable.

g) Black-box robustness: To investigate the black-box robustness of the proposed model, we perform PGD attacks on four source models: vanilla ($B_{nat}$) and robust ($B_{adv}$) versions of Madry’s model, and those ($C_{nat}$ and $C_{adv}$) of the proposed. We again use $\epsilon = 0.3$ and $\epsilon = 8/255$ for MNIST and CIFAR10, respectively. Test results are summarized in Tab.VI where rows and columns correspond to different test and source models, respectively. Diagonal elements are the white-box accuracy. The proposed model achieves significant improvements under all tested black-box attacks.

B. The influence of input cropping-invariant attacks on the learning of robust features

From the experiments we observe that cropping is effective as an input transformation at improving model robustness. In the following, we provide preliminary explanations to this finding through two toy cases inspired by [18]: 1) We use a Gaussian data model to show that using cropped inputs for adversarial training leads to a higher probability of both dropping non-robust features and learning robust ones; 2) we then conduct a binary classification task on digits “5” and “7” from MNIST, and empirically show that incorporating cropping leads to more successful learning of robust features.

a) Preliminary analysis of the proposed method: A binary classification task: Consider a data model consisting of input-label pairs $(x, y)$ sampled from a distribution $G$:

$$y \overset{a,r}{\sim} \{-1, 1\}, \ x_{i,i,d} \sim \mathcal{N}(\eta y, 1),$$

where $\mathcal{N}(\mu, \sigma^2)$ is a normal distribution with mean $\mu$ and variance $\sigma^2$, and $\eta \in [0, 1]$ represents the correlation between $x_i$ (the $i$th element of $x$) and $y$. We consider a linear classifier with parameters $w$, $f(x) := \text{sign}(w^T x)$. With regularization, adversarial training solves:

$$\min_{w:||w||_2 \leq 1} \max_{\epsilon \leq \epsilon} \mathbb{E}_D \left[ \max(0, 1 - y w^T (x + \delta)) \right].$$

From [18], the solution to Eq. follows a simple rule: With a finite dataset $D$ drawn from $G$, if $|\mathbb{E}_D[y x_i]| \geq \epsilon$, $w_i$ is assigned a non-zero weight, or otherwise $w_i$ is zero. See Fig. 6 for an example. When $D$ is infinite, we derived the true robust features $R := \{x_i | \eta_i \geq \epsilon\}$. Without loss of generality, we will consider $\eta_1 = 1$ and $\eta_2 = \eta_3 = \ldots = \eta_{d+1} = \eta < \epsilon$ in the following analysis.

Robust features under the proposed model: We consider a simple aggregation over transformations $T := \{T_k\}_{k=1}^K$, which leads to the classifier $f(x) := \text{sign}(\theta^T z)$, where $z := (1/K) \sum_{k=1}^K T_k(x)$, and $T_k(x)$ is the $k$th transformed feature. Let $z_i$ be the $i$th element of $z$. With input cropping, we have $z_i = (1/K) \sum_{j \in N_i} x_{j,i}$, where $N_i$ is a set of image pixels for the $i$th aggregated feature. For simplicity, we assume that $N_i = \{i, i+1, \ldots, i+K\}$ for $i = 1, \ldots, d+1-K$. This makes $N_1$ the only set that contains $x_1$ (the robust feature). We have the following distribution in the aggregated feature space:

$$z_1 \sim \mathcal{N}(\rho, \frac{1}{K}), \ z_i \overset{i.i.d.}{\sim} \mathcal{N}(\eta \rho, \frac{1}{K}), \ \forall i \geq 2,$$

where $\rho = \frac{y^{(1+(T-1)\eta)}}{K}$. When $\rho \geq \epsilon$, and under infinite data size, adversarial training leads to $\theta_1 = 1$ and all other weights as 0. Thus $z_1$ is the robust (aggregated) feature.

Distribution of sample means: Denote the sample means of the correlations as $\bar{\eta}_i$ for $x_i$, and $\bar{\gamma}_i$ for $z_i$. Under a finite
data size $N$, $\hat{\eta}_i$ and $\hat{\gamma}_1$ follow the following distributions

$$
\hat{\eta}_i \sim \mathcal{N}(y_i, \frac{1}{N}), \quad \hat{\eta}_i \overset{i.i.d.}{\sim} \mathcal{N}(\eta y_i, \frac{1}{N}), \quad \forall i \geq 2,
$$

$$
\hat{\gamma}_1 \sim \mathcal{N}(\rho, \frac{1}{NK}), \quad \hat{\gamma}_i \overset{i.i.d.}{\sim} \mathcal{N}(\eta y_i, \frac{1}{NK}), \quad \forall i \geq 2,
$$

(8)

**Effectiveness of dropping non-robust features:** The probabilities of correctly dropping a non-robust feature are $\Phi((\epsilon - \eta)N)$ and $\Phi((\epsilon - \eta)NK)$ before and after applying the transformation, respectively, where $\Phi$ is the cumulative distribution function for $\mathcal{N}(0, 1)$. Since $\epsilon - \eta > 0$ for non-robust features and the number of transformations $K > 1$, the transformation improves the effectiveness of dropping non-robust features.

**Effectiveness of learning robust features:** The probabilities of correctly learning the robust feature are $p(\hat{\eta}_1 > \epsilon) = 1 - \Phi((\epsilon - 1)N)$ and $p(\hat{\gamma}_1 > \epsilon) = 1 - \Phi((\epsilon - \rho)NK)$ before and after applying the transformation, respectively. We can derive from here that $p(\hat{\eta}_1 > \epsilon) < p(\hat{\gamma}_1 > \epsilon)$ if

$$
2 \log K + Nh(\rho) > 0,
$$

(9)

where $h(\rho) = \epsilon^2(K - 1) + (\rho^2K - 1) - 2\epsilon(\rho K - 1)$ is a quadratic function of $\rho$. Given $K$, $\epsilon$, and $\rho$, Eq. (9) sets a condition on $N$ for which a model that incorporates transformation will have a higher probability of correctly extracting the robust feature. Specifically, we can derive the following conclusions: (1) When $\rho$ is close to 1, $p(\hat{\eta}_1 > \epsilon) < p(\hat{\gamma}_1 > \epsilon)$ for any $N > 0$. We can show in particular that when $\rho = 1$,

$$
\epsilon^2(K - 1) + (\rho^2K - 1) - 2\epsilon(\rho K - 1) = (K - 1)(1 - \epsilon)^2 > 0,
$$

(10)

thus any $N > 0$ satisfies Eq. (9). (2) For $\rho \in [1/K, \rho^*)$, where $\rho^*$ is the larger root of $h(\rho) = 0$, $p(\hat{\eta}_1 > \epsilon) < p(\hat{\gamma}_1 > \epsilon)$ if

$$
N < -\frac{2 \log K}{h(\rho)}.
$$

(11)

This preliminary analysis shows that the probability for learning an aggregated feature with high correlation (e.g., a set of neighbouring pixels that are unique to one class) is higher than that for learning an individual feature with equally high correlation (with the label). On the other hand, for aggregated features that are moderately correlated with the label (e.g., a pixel unique to a class surrounded by noises), there is an upper bound on the data size for which incorporating transformation in adversarial training will gain an advantage.

**b) Empirical examination:** Here we formulate a binary classification problem using digits “5” and “7” from MNIST. We use a cropping size of 26 and a model that aggregates 9 cropped inputs. Correlation values for all features with and without cropping are computed (Fig. 5a). With $\epsilon = 0.2$ and all training data (10k in total), the robust features are identified for a linear model defined on the original images and a separate linear model defined on the features aggregated from cropped images. The resultant categorization of robust and non-robust features is considered as the ground truth.

We then perform the same computation on random batches of the data (400 for each), mimicking adversarial training under small dataset. For each batch, the learned robust features are recorded. By combining results from 200 random batches, we compute the percentage of batches that each feature is being correctly classified as robust or non-robust. A comparison is shown in Fig. 5b. The model with input cropping achieves higher success rate at learning robust features and dropping non-robust ones: Without cropping, there are 16 robust features not always regarded as robust, and 32 non-robust features regarded as robust in some cases. With cropping, these numbers reduce to 9 and 10, respectively.

**c) Network Visualization:** Lastly, we qualitatively evaluate the efficacy of our model at learning robust features, by considering such features as human-interpretable image patterns. We first visualize the adversarial gradients from $C_{nat}$ and $C_{adv}$ in Fig. 7. The result shows that incorporating cropping invariance alone does not yield meaningful adversarial gradients. However, combined with adversarial training, the adversarial gradients become interpretable.

We further investigate network filters from MNIST. Consistent with [25], the first-layer filters for Madry’s model

| TABLE V |
| --- |
| **COMPARISON ON LEARNING EFFICIENCY. A: CLEAN TEST ACCURACY, R: WHITE-BOX ROBUSTNESS UNDER PGD ATTACKS** |
| CIFAR10 |
| MNIST |
| $A(B_{adv})$ | $A(C_{adv})$ | $R(B_{adv})$ | $R(C_{adv})$ |
| 91.7 | 96.6 | 86.2 | 86.1 |
| 93.4 | 96.3 | 97.6 | 98.5 |
| 86.1 | 98.8 | 98.7 | 99.2 |
| 69.7 | 10 | 10 | 10 |

| $B_{nat}$ | $C_{nat}$ | $B_{adv}$ | $C_{adv}$ | worst case |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 8.8 | 0.6 | 89.5 | 8.6 | 0.6 |
| 85.6 | 89.5 | 93.2 | 93.2 | 86.1 |
| 94.8 | 86.1 | 47.4 | 66.9 | 47.4 |

| $C_{adv}(Ours)$ | $97.9$ | $97.4$ | $98.2$ | $95.7$ |
| $C_{adv}(Ours)$ | $96.7$ | $96.5$ | $93.2$ | $95.3$ |

| TABLE VI |
| --- |
| **ROBUSTNESS ON MNIST AND CIFAR10 AGAINST BLACK-BOX PGD ATTACKS** |
| CIFAR10 |
| MNIST |
| $B_{nat}$ | $C_{nat}$ | $B_{adv}$ | $C_{adv}$ | worst case |
| $B_{nat}$ | $C_{nat}$ | $B_{adv}$ | $C_{adv}$ | worst case |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 8.8 | 0.6 | 89.5 | 8.6 | 0.6 |
| 85.6 | 89.5 | 93.2 | 93.2 | 86.1 |
| 94.8 | 86.1 | 47.4 | 66.9 | 47.4 |
| 97.9 | 97.4 | 98.2 | 95.7 | 96.2 | 86.3 | 69.5 | 54.4 | 54.4 |
Fig. 6. (a) Correlation of each feature to labels w/ and w/o cropping aggregation. Dots are classifier weights derived from adversarial training based on Eq. (6). (b) The ratio that a feature is regarded as robust across 200 random batches. Dashed lines in black and yellow separate robust and non-robust features for models w/ and w/o cropping aggregation, respectively. The ideal ratio curve would be constantly 1 to the left of the dashed line and 0 to the right.

Fig. 7. Visualization of adversarial gradients. Images from top to bottom are random clean samples, adversarial gradients from $C_{\text{nat}}$, and those from $C_{\text{adv}}$. are sparse. In comparison, our model increases the number of non-zero filters from 3 to 22. Considering that these filters serve as denoisers with learned thresholds, more filters would allow richer information to be passed to the next layer of the network. We also compare the second-layer filters from Madry’s and our model in Fig. 8. Our model is able to produce more human-interpretable strokes, which may explain its improvement in robustness.

Fig. 8. Visualization of the second-layer filters from networks trained on MNIST (left: Madry’s, right: proposed). The proposed method is able to learn more meaningful strokes.

IV. RELATED WORK

a) Random input transformation: Guo et al. [36] proposed using random transformations to pre-processing the input images to improve model robustness. It was later shown, however, that this approach creates a gradient masking effect and can be broken by robust attacks [17]. Unlike [36], we consider the transformation as part of our model during the adversarial training process.

b) Bag of features: Studies on bag of features (BoF) [39]–[42] proposed the aggregation of clusters of local image features for classification. While our model also takes in multiple inputs as in BoF models, we assume that these images are all content preserving, and do not rely on an aggregation for classification.

c) Ensemble adversarial training: Attacks from an ensemble of black-box models have been used to effectively avoid gradient masking in one-step adversarial training [38]. While our model also uses an ensemble of attacks, these attacks are white-box and multi-step. Importantly, these attacks do not cause gradient masking.

V. LIMITATIONS AND FUTURE WORK

a) Limitations: Computational cost: The proposed model requires higher computational cost due to the computation of transformation-invariant attacks. As the effectiveness of the proposed method is positively correlated with the number of transformations (shown experimentally in Tab. III and theoretically in Sec. III-B for image cropping), the success of the method depends on the availability of parallel GPUs for computing attack gradients. Due to this limitation, our experiments on the restricted ImageNet dataset [18] have so far achieved limited success. Specifically, with cropping size 168 and four crops, the proposed model achieves 92.83% and 96.93% for robustness and standard accuracy, which are comparable to 92.75% and 96.83% for Madry’s model. We hypothesize that increasing the cropping size and the number of crops will further improve the model robustness.

b) Verification: Existing formal verification [43] and certification [44]–[46] methods rely on the linearity of the decision function. Our model aggregates softmax outputs (Eq. (2)), which makes existing tools not directly applicable. It will be necessary to investigate whether the aggregation can be performed before softmax.

c) Future work: CV-based transformations: Our analysis suggests that transformations that lead to features with high correlation with labels are desired. Considering that traditional CV-based features such as SIFT [47] and SURF [48] are readily available and designed to align well with human vision, it is reasonable to hypothesize that such filters can help improve model robustness through adversarial training.

VI. CONCLUSIONS

In this paper we investigated a learning architecture that incorporates input transformations into adversarial training, and showed that the model (1) is more effective at extracting robust features, (2) achieves better empirical robustness than the state of the art on MNIST and CIFAR10, and (3) is more data efficient, by using image cropping as an ensemble of transformations. Importantly, we showed that while constraining the model to be transformation invariant
References

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

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," pp. 1–11, 2014. [Online]. Available: http://arxiv.org/abs/1412.6572

[3] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," arXiv preprint arXiv:1801.00553, 2018.

[4] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 427–436.

[5] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2574–2582.

[6] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," ArXiv, no. pp., pp. 1–15, 2016. [Online]. Available: http://arxiv.org/abs/1607.02533

[7] K. Eykholt, I.Evtimov, E. Fernandes, B. Li, A. Rahman, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1625–1634.

[8] J. Hendrik Metzen, M. Chaitanya Kumar, T. Brox, and V. Fischer, "Universal adversarial perturbations against semantic image segmentation," in Proceedings of the IEEE Conference on Computer Vision, 2017, pp. 2755–2764.

[9] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, "Adversarial examples for semantic segmentation and object detection," in International Conference on Computer Vision. IEEE, 2017.

[10] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Analyzing and Harnessing Adversarial Examples," pp. 1–11, 2014. [Online]. Available: http://arxiv.org/abs/1412.6572

[11] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," arXiv preprint arXiv:1801.00553, 2018.

[12] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," arXiv preprint arXiv:1611.01236, 2016.
[45] E. Wong and Z. Kolter, “Provable defenses against adversarial examples via the convex outer adversarial polytope,” in International Conference on Machine Learning, 2018, pp. 5283–5292.

[46] E. Wong, F. Schmidt, J. H. Metzen, and J. Z. Kolter, “Scaling provable adversarial defenses,” arXiv preprint arXiv:1805.12514, 2018.

[47] P. C. Ng and S. Henikoff, “Sift: Predicting amino acid changes that affect protein function,” Nucleic acids research, vol. 31, no. 13, pp. 3812–3814, 2003.

[48] H. Bay, T. Tuytelaars, and L. Van Gool. “Surf: Speeded up robust features,” in European conference on computer vision. Springer, 2006, pp. 404–417.