Test-time Detection and Repair of Adversarial Samples via Masked Autoencoder

Yun-Yun Tsai[1]   Ju-Chin Chao[1]   Albert Wen[1]   Zhaoyuan Yang[2]
Chengzhi Mao[1]   Tapan Shah[2]   Junfeng Yang[1]
[1]Columbia University   [2]GE Research

{yt2781,jc5859,aw3575,cm3797}@columbia.edu,
{zhaoyuan.yang,tapan.shah}@ge.com, junfeng@cs.columbia.edu

Abstract

Training-time defenses, known as adversarial training, incur high training costs and do not generalize to unseen attacks. Test-time defenses solve these issues but most existing test-time defenses require adapting the model weights, therefore they do not work on frozen models and complicate model memory management. The only test-time defense that does not adapt model weights aims to adapt the input with self-supervision tasks. However, we empirically found these self-supervision tasks are not sensitive enough to detect adversarial attacks accurately. In this paper, we propose DRAM, a novel defense method to detect and repair adversarial samples at test time via Masked autoencoder (MAE). We demonstrate how to use MAE losses to build a Kolmogorov-Smirnov test to detect adversarial samples. Moreover, we use the MAE losses to calculate input reversal vectors that repair adversarial samples resulting from previously unseen attacks. Results on large-scale ImageNet dataset show that, compared to all detection baselines evaluated, DRAM achieves the best detection rate (82% on average) on all eight adversarial attacks evaluated. For attack repair, DRAM improves the robust accuracy by 6%∼41% for standard ResNet50 and 3%∼8% for robust ResNet50 compared with the baselines that use contrastive learning and rotation prediction.

1. Introduction

Despite remarkable advances in prediction accuracy and wide adoption of deep learning (DL) models, they are known to be vulnerable to adversarial attacks in which attackers add carefully crafted, human-imperceptible perturbations to the samples to fool the models [1,3,16,27]. Most prior defenses happen at training time [24,28,46], which incurs high training cost and cannot defend against unseen attacks. Recent work proposed test-time defenses [8,13,41, 48] that adapt the model weights to improve prediction accuracy on out-of-distribution data, but they require access to
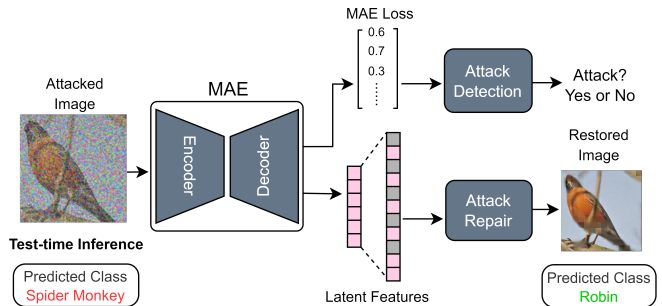

Figure 1. A schematic of DRAM. (1) Attack detection: Leverages the MAE loss to detect adversarial samples via a statistic-based test. (2) Attack Repair: Adapts inputs by adding a reversal vector to the input space and optimizing it with MAE loss, which can potentially repair the adversarial sample.

the weights and cannot work on frozen models. Additionally, they may require multiple versions of the same models to operate simultaneously, complicating model management and increasing memory consumption. Another test-time defense [29, 40] reverses attacks by adapting inputs computed using self-supervised learning (SSL) tasks such as contrastive learning [5] or rotation prediction [14]. However, as our experiments reveal, these tasks are too weak to capture sufficiently detailed representations for effective detection of adversarial samples.

This paper presents DRAM, a test-time defense using masked autoencoder (MAE) [17], one of the strongest SSL tasks, to detect and repair adversarial attacks, including unforeseen ones. MAE is a ViT-based architecture that reconstructs masked patches based on the sparse information of unmasked patches. Prior work [17] showed that MAE is both efficient at reducing redundancy in feature representations and capturing detailed information from holistic image statistics. We leverage MAE to detect adversarial samples generated from potentially unforeseen attacks via a statistic-based test on the MAE loss. Moreover, without knowing the ground truth of the samples, DRAM repairs

the samples by minimizing the MAE loss at test time. We highlight our main contribution as follows:

- We propose a test-time defense method for unknown adversarial attacks using self-supervised objectives and masked autoencoder.

- We illustrate the method of DRAM on both detection and repair: (1) Adversarial detection by statistic-based learning with reconstruction loss. (2) Repair of the adversarial samples by injecting imperceptible perturbations and optimizing with self-supervised loss. (3) We launch an adaptive attack that assumes the attacker has knowledge of our defense method and shows the effectiveness of DRAM on the defense-aware attack.

- DRAM achieves 82% detection rate on eight kinds of adversarial samples average and outperforms other detection baselines. For attack repair, DRAM successfully repairs the adversarial samples and improves robust accuracy by 6%∼41% on eight attacks for ResNet50 under normal training setting and 3%∼8% under robust training setting. Our method outperforms two self-supervision tasks, including contrastive learning and rotation prediction.

2. Related Works

Self-Supervised Learning Self-supervised learning (SSL) is one of the most efficient ways to learn effective representations from images without annotations [2,7,9,20]. Prior works have shown that representations learned from different pretext tasks (e.g., jigsaw puzzles [34], rotation prediction [14], image colorization [25] and deep clustering [22]) can be leveraged on several downstream tasks such as image classification [4], object detection [11] and test-time domain adaptation [38]. Another well-known branch of SSL is contrastive learning, which aims at grouping associated features for a set of transformations for samples and distancing from other samples with dissimilar features in the dataset [6, 18, 35]. Prior works proposed to use SSL method for outlier detection [20, 37, 45], which aims at learning generalizable out-of-detection features and rejecting them during the testing time. However, these discriminative SSL tasks learn to discard a large portion of their input information. Representations learned from these tasks keep only a small amount of information or features from the input space. Consequently, these objective functions are less sensitive to adversarial attacks.

Masked Autoencoder MAE learns to reconstruct the randomly masked patches of a given input. Recent research leverages MAE to learn downstream tasks since it allows the models to learn representations beyond the low-level statistics. It first encodes a small subset of unmasked patches to latent space as embedding. Then, the masked tokens are inserted after encoding. The concatenation of unmasked patches' embedding and masked tokens are processed by the decoder to predict the unmasked patches [17, 43]. Existing works show the success of mask-based pre-training on different modalities, such as improving point cloud models for 3D object detection [33], and model adaptation on sequential data [42]. In [15], the authors proposed to reverse engineer the perturbation and denoise adversarial samples with an autoencoder via class-aware alignment training. [13] leverages MAE on test-time training by reconstructing each test image at test time for corrupted samples. Inspired by this, we leverage MAE on adversarial attack detection and repair.

Defense for Adversarial Attacks Robustness to adversarial attacks is a critical issue in the machine learning field [16]. Large amounts of work have studied adversarial attacks in either whitebox or blackbox (e.g., FGSM, C&W, and ZOO attack [1, 3, 16]). To improve model robustness on these attacks, adversarial training is one of the most efficient methods [24, 28, 31, 46]. Madry et al. [28] propose to train the model by minimizing the worst-case loss in a region around the input. TRADES [46] smoothen the decision boundary by considering the ratio between the loss of clean input and the adversarial input. Mao et al. [30] shows that learning multiple tasks improves adversarial robustness. However, the previously mentioned methods need retraining on the model.

Test-time defense circumvents retraining and largely reduces the computational cost. Most of the test-time adaptation adjusts the model weights with batch samples during the inference time [8, 12, 13, 26, 41, 48, 49]. Test-time transformation ensembling (TTE) [36] augments the image with a fixed set of transformations and ensemble the outputs through averaging. Test-time Training (TTT) [39] trains the model with an auxiliary SSL task (e.g., rotation prediction) and adapts the model weights based on the SSL loss. MEMO [47] adapts every sample individually by minimizing the marginal cross-entropy on the transformation set. Instead of adjusting model weights, test-time reverse [29, 40] aims at adjusting input by adding some noise vectors and optimizing them with self-supervision tasks.

3. Method

In this section, we define the problem and present the details of each component in the pipeline of DRAM, including test-time detection and repair of attacks.

3.1. Problem Definition

We formalize the problem of learning to capture the natural input's manifold for detecting and repairing the mali-

cious inputs. Assuming we have a self-supervised learning (SSL) task $\mathcal{S}$, we denote its objective function as $\mathcal{L}_s$. To capture the manifold's natural inputs for a given classifier $\mathcal{C}$, we learn the SSL task $\mathcal{S}$ using the training set of SSL task $\mathcal{S}$ doesn't involve any attacked samples.

During the inference time, given an adversarial input u_{adv} which is created by adding an adversarial perturbation δ_{adv} to a clean sample $u \in \mathcal{Z}$, resulting in an output $\hat{y} = C(u_{\text{adv}})$ that does not match the ground truth label y. To mitigate these adversarial attacks, we propose using a reverse vector δ_{rev} to minimize the difference between the adapted image and the original input, effectively generating an adversarial perturbation for correction. Our goal is to find an optimal reverse vector δ_{rev} with a small ϵ range that can minimize the self-supervised loss $\mathcal{L}_s$ and restore the correct prediction, such that $C(u) = y = C(u + \delta_{\text{adv}} + \delta_{\text{rev}})$.

3.2. Learning with Masked Autoencoder (MAE)

Self-supervised learning (SSL) helps capture the natural inputs manifold, yet standard self-supervision pretext tasks such as rotation prediction, re-colorization, and contrastive learning are not sensitive enough to adversarial attacks. The MAE is one of the strongest self-supervision learning tasks, which can repair the randomly masked-out input patches by minimizing the MAE loss of predicted patches. Motivated by this, we show how to learn the reverse vector for repairing the attacked images with MAE.

Given a masked autoencoder (MAE), we define the encoder, parameterized by θ, as $f_\theta : \mathbb{R}^n \to \mathbb{R}^m$. Let $x \in \mathcal{X}$ be an input data in the image space, and $b \in \{0, 1\}^n$ be the binary matrix that is sampled following a probability distribution. The encoder takes a masked sample $b \odot x$ as input and outputs a latent variable $z = f_\theta(b \odot x)$. We define a decoder, parameterized by ϕ, as $g_\phi : \mathbb{R}^m \to \mathbb{R}^n$, which takes a latent variable z and reconstruct the mask regions of the input image as $\hat{x}_{\text{mask}} = g_\phi(z) = (g_\phi \circ f_\theta)(b \odot x)$.

The objective function of MAE is to compute the mean square error between the masked patches of input x and their corresponding predicted patches, which can be defined as Eq. 1:

$$\mathcal{L}_{\text{mse}}(\cdot) = \|(1 - b) \odot x - (g_\phi \circ f_\theta)(b \odot x)\|_2. \quad (1)$$

We have discovered that the properties of the MAE can effectively capture the underlying structure of natural inputs, allowing us to achieve our objective of detecting and correcting adversarial inputs. This key insight forms an integral part of our proposed DRAM method, designed to improve the robustness and reliability of machine learning systems by detecting and repairing adversarial inputs.

3.3. Test-time Attack Detection

After obtaining a well-trained MAE (f_θ, g_ϕ), the detection of anomalous adversarial attacks is straightforward by using the MAE loss $\mathcal{L}_{mse}(\cdot)$. The rationale lies in our training for f_θ and g_ϕ being on the clean sample sets $\mathcal{S}$, which can be viewed as the manifold learning on in-distribution data. Thus, the magnitude of MAE loss enlarges while encountering adversarial samples u from $\mathcal{Z}$, thus triggering the detection.

Here, we utilize MAE loss $\mathcal{L}_{mse}$ with a statistic-based method for detection in DRAM, the Kolmogorov-Smirnov test (KS test) [32]. We hypothesize that loss values associated with non-attacked or clean images will be distinctly lower on average compared to those associated with attached images. Moreover, in Figure 2, the sampled distributions of the loss values formed the histogram which delineates the difference in distribution between non-attacked and attacked images. Combining SSL loss with the KS test for adversarial detection is better than other detection methods that calculate the distance metric for logit outputs. In Section 4.2, we illustrate the details of other baselines and demonstrate the detection results.

3.4. Test-time Attack Repair

The second part of the DRAM method involves repairing adversarial attack inputs from malicious to benign. Unlike the existing method [15], our method is fully unsupervised and can be done parallel to detection. To repair the potential adversarial samples u_{adv}, we inject a reverse vector δ_{rev} into u_{adv} such that $\hat{u} = u_{adv} + \delta_{rev}$ minimizes the MAE loss $\mathcal{L}_{mse}$. We initialize the δ_{rev} under the uniform distribution following a specific ϵ range. After optimization, the adapted sample $\hat{u} = u_{adv} + \delta_{rev}$ is fed to the classifier C for prediction, yielding classification scores $\hat{y} = C(\hat{u})$. The objective function for optimizing δ_{rev} is defined as Eq. 2:

$$\delta_{rev} = \arg \min_{\|\delta_{rev}\|_\infty \leq \epsilon} \mathcal{L}_{mse}((1 - b) \odot \hat{u}, (g_\phi \circ f_\theta)(b \odot \hat{u})), \quad (2)$$

We aim to find an optimal vector δ_{rev} for generating a adapted sample $\hat{u} = u_{adv} + \delta_{rev}$ such that the gap between masked patches $(1 - b) \odot \hat{u}$ and predicted patches $(g_\phi \circ f_\theta)(b \odot \hat{u})$ of the adapted sample can be reduced. Specifically, we update δ_{rev} using projected gradient descent, subject to ℓ_∞ with a specific ϵ range. The update of δ_{rev} is described in Eq. 3.

$$\delta_{rev}^{t+1} = \text{clip}(\delta_{rev}^t + \alpha \cdot \text{sign}(\nabla_X \mathcal{L}_{\text{mse}}), -\epsilon, \epsilon). \quad (3)$$

Here, t is the timestamp, α is the step size, and ϵ is the perturbation range. The update of δ_{rev} is performed by taking the gradient of the MAE loss $\mathcal{L}_{mse}$ with respect to the input and clipping the update of δ_{rev} to the ϵ range.

3.5. Defending against an Adaptive Attack

We show that DRAM is able to defend against an adaptive attack in which the attackers have knowledge about our defense method, including the MAE architecture, attack detection, and repair. To launch an adaptive attack that can fool both MAE and classifier $\mathcal{C}$, an attacker can generate adversarial examples by jointly optimizing the MAE loss and cross-entropy loss for the classifier. We change the optimization problem of δ to the following Eq. 4 for the adaptive attack.

$$\max \mathcal{L}_{ce}(u, y), \text{s.t.} \quad \mathcal{L}_{mse}((1-b)\odot u, (g_\phi \circ f_\theta)(b\odot u)) \leq \epsilon \tag{4}$$

The attacker should maximize the cross-entropy loss between the ground truth and output probability score of u from backbone model $\mathcal{C}$, while reducing the mean-square error loss of MAE for u. As Eq. 5 formulates, we incorporate the Lagrange multiplier to solve the constrained optimization problem from Eq. 4.

$$\mathcal{L}_{da}(\cdot) = \mathcal{L}_{ce}(u, y) - \lambda \mathcal{L}_{mse}((1 - b) \odot u, (g_\phi \circ f_\theta)(b \odot u)), \tag{5}$$

where the λ is the parameter to control the adversarial budget. If the λ increases, the optimization gradually shifts the budget for attacking the classification loss $\mathcal{L}_{ce}$ to the MAE loss $\mathcal{L}_{mse}$. Optimizing with $\mathcal{L}_{da}$, the optimal δ can be found to generate the defense-aware adversarial samples. We empirically found the trade-off between $\mathcal{L}_{mse}$ and $\mathcal{L}_{ce}$, where the joint optimization on these two objective functions leads to a degradation of the attack success rate for the defense-aware adversarial samples. Our discovery is consistent with the theoretical analysis presented by Mao et al. [29]. Therefore, our defense method is robust against adaptive attacks.

4. Experiment

4.1. Experimental Settings

Dataset Detail We train and test DRAM using the ImageNet dataset [10], a 1K-class large-scale vision dataset. Notice that the training process of DRAM uses only the original ImageNet train set with $\sim$1M images. To evaluate DRAM, we generate the adversarial samples on the ImageNet test set [10] with 50K images. Here, we consider eight attacks and generate them from two kinds of models, the standard [19] and robust [27] ResNet50.

Model Detail The MAE model is based on the visual transformer (ViT) arcitecture [23], which consists of the same encoder module as ViT that splits images into patch size 14×14 and encodes their corresponding projected sequences. The encoder module of MAE consists of alternating layers of multiheaded self-attention (MSA) and MLP blocks. The batch normalization layer (BN) is applied before every block, and residual connections after every block. The encoded features are further processed by the decoder module for reconstruction.

4.2. Test-time Attack Detection

We compare the detection part of DRAM with the following four kinds of baselines:

• **Feature Squeezing (FS)**: The baseline [44] utilizes multiple data transformations such as median smoothing, bit quantization, or non-local mean to extract the squeezing features of input samples. The ℓ_1 distance is then computed between the transformed version and the original version of inputs for detection. If the ℓ_1 distance is larger than the threshold, the samples are rejected as adversarial.

• **Noise-based Detection (ND)**: This method [21] detects whether an input is adversarial or not via perturbing the input with random noise. It samples the ϵ in a normal distribution with specific radius r and computes the ℓ_1 distance as $\| h(x) - h(x + \epsilon) \|_1$ for given input x. The input x is detected as adversarial if the ℓ_1 distance is large.

• **Targeted Detection (TD)**: The detection method [21] aims to evaluate the PGD attack algorithm on every input x and record the number of steps K to fool the sample x from the current predicted class to targeted class. If the number of steps K is larger than the threshold, we detect the sample x as adversarial.

• **Detection with Self-Supervised Loss (SSL)**: The detection method combines SSL loss with the 2-sample KS-test. It measures the distributions of the observations from the clean and attacked samples. The samples are rejected as adversarial if the p-value exceeds a threshold. We compare DRAM with two SSL tasks, including contrastive learning (SimCLR [5]) and rotation prediction [14].

Table 1 shows the true positive rate (TPR) of our method against eight kinds of adversarial attacks on ImageNet data as the detection rate. Our method combines the $\mathcal{L}_{mse}$ loss with a two-sample KS test and set the p-value threshold based on a fixed false positive rate (FPR) of 0.2. Here, we evaluate the detectors for every baseline using the same FPR. Compared with feature squeezing [44], noise-based detection [21], targeted detection [21], and detection with the two SSL tasks, our detection rate hovers around 82%, which outperforming all other baselines. The superior detection results of DRAM suggest that the MAE loss is more sensitive to variance in distributions. Compared to the other two self-supervision tasks, DRAM outperforms rotation prediction by up to 20% and contrastive learning by up to 60%. Moreover, the two-sample KS-test has the additional benefit of being able to better control the TPR-

Detector	FPR	FGSM-ℓ_∞ 4/255	FGSM-ℓ_∞ 8/255	PGD-ℓ_∞ 8/255	PGD-ℓ_∞ 16/255	PGD-ℓ_2 1	CW-ℓ_2 1	BIM-ℓ_∞ 8/255	AutoAttack -
FS [44]	0.2	0.001	0.000	0.0015	0.000	0.0235	0.022	0.000	0.000
ND [21]	0.2	0.6844	0.6736	0.7024	0.6752	0.3535	0.6296	0.697	0.712
TD [21]	0.2	0.089	0.067	0.107	0.125	0.066	0.051	0.1165	0.1045
SSL-*SimCLR*	0.2	0.2959	0.4081	0.6479	0.8163	0.4286	0.3980	0.6582	0.4948
SSL-*Rotation*	0.2	0.6582	0.8367	0.9031	0.9639	0.5510	0.3520	0.8934	0.7653
DRAM (ours)	0.2	**0.8504**	**0.9680**	**0.9129**	**0.9837**	**0.5900**	**0.5739**	**0.9024**	**0.8201**

Table 1. Detection results on eight kinds of adversarial attacks. We compared with five baselines, including Feature Squeezing (FS), Noise-based Detection (ND), Targeted Detection (TD), SSL detection using contrastive loss (SSL-*SimCLR*), and loss of rotation prediction (SSL-*Rotation*). We show the true positive rate (TPR) for every attack with the fixed False positive rate as 0.2 for all methods by varying the thresholds. The numbers in bold represent the best performance.

	FGSM-ℓ_∞ 8/255	FGSM-ℓ_∞ 16/255	PGD-ℓ_∞ 8/255	PGD-ℓ_∞ 16/255	PGD-ℓ_2 1	C&W-ℓ_2 1	BIM-ℓ_∞ 8/255	AutoAttack -
Standard	6.39	8.55	0.0	0.0	0.65	0.84	0.01	0.10
+ SimCLR [4]	9.46	8.82	1.89	0.12	15.59	17.62	1.74	3.31
+ Rotation [14]	11.40	9.85	1.47	0.98	0.64	1.02	3.89	1.49
+ DRAM (ours)	**20.54**	**14.14**	**21.34**	**13.12**	**39.57**	**41.79**	**20.51**	**22.4**
MADRY [27]	24.95	13.9	20.38	8.83	35.11	40.82	20.26	16.08
+ SimCLR [4]	24.19	13.31	20.33	8.75	35.80	43.37	20.31	18.84
+ Rotation [14]	22.57	12.35	19.94	8.94	34.38	41.00	20.04	22.69
+ DRAM (ours)	**25.43**	**14.05**	**23.21**	**11.01**	**38.71**	**43.75**	**23.14**	**24.84**

Table 2. The repair results on eight kinds of attack. We compare DRAM with the other two SSL baselines, including contrastive learning task (SimCLR [5]) and rotation prediction [14]. We show the robust accuracy (%) on two Resnet50 [19] models pre-trained with Standard training and ℓ_∞ robust training (MADRY [27]). The numbers in bold represent the best performance.

FPR tradeoff compared to all other baselines by varying the p-value threshold for class prediction.

4.3. Attack Repair Results

We compare our repair method with two kinds of SSL baselines (1) Contrastive learning learns the representations that map the feature of the transformations of the same image into a nearby place. We follow the SimCLR [5] to implement the contrastive learning task. We create positive/negative pairs for images by applying some transformations such as random cropping/flipping and random rotation. (2) The rotation prediction [14] learns to predict the rotation degree of upcoming input samples. Here, we train the rotation prediction task as one of our SSL baselines with 4 classes ($0°$, $90°$, $180°$, $270°$).

As Section 3 mentioned, we optimize the perturbations for a given input by using the PGD algorithm based on the MAE loss $\mathcal{L}_{\text{mse}}$ extracted from MAE. Table 2 shows we evaluate the robust accuracy on the target classifiers $\mathcal{C}$ from which our adversarial samples were generated. DRAM successfully repairs the adversarial images and improves the robust accuracy without requiring the ground truth labels at the inference time. The target classifiers are pre-trained with two different methods, including standard training and robust training. As Table 2 shows, compared with the two baselines using contrastive and rotation losses, DRAM im-

proves the robust accuracy after repair by $6\% \sim 40\%$ for the attack samples generated from the standard ResNet50 model, whereas contrastive learning (SimCLR) and rotation prediction improves only by $1\%\sim7\%$. For the Robust ResNet50 model, DRAM improves the robust accuracy by $1\%\sim8\%$ for every attack, outperforming the two baselines.

4.4. Defense-aware Attack (DAA) Results

As Section 3.5 mentioned, we launch the adaptive attack on models and demonstrate the effectiveness of DRAM on this stronger attack which assumes the attacker knows our defense method and the MAE model. Figure 3 shows the result of the defense-aware attack on two classifiers, including standard ResNet50 and robust ResNet50 [27]. While increasing the self-supervised attack strength λ for DAA from 0 to 6, the attacks are gradually weakened, and the robust accuracy increases. If the adversary attempts to attack the MAE, the attack will be less effective on the backbone classifiers. On the other hand, if the adversary attempts to attack the classifier by generating samples with high attack success rates, the MAE loss can't be minimized together. While the dotted line shown in Figure 3 indicates the result of DAA, the solid line shows the repair result of DRAM on top of DAA, which indicates that our method has optimal results when λ is set as 0 for both the standard and robust model. Therefore, it is difficult for adversaries to attack both MAE
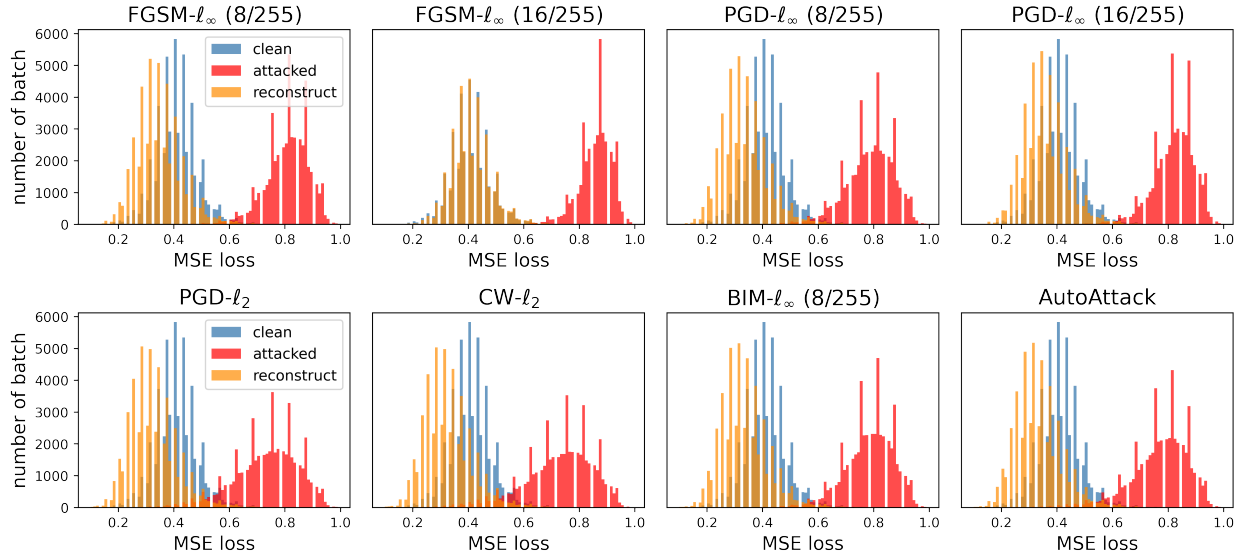
Figure 2. Histogram plot of self-supervised losses $\mathcal{L}_{mse}$ for eight adversarial attacks. We show the loss distribution of original images with the color blue, adversarial images as red, and adapted images as green. The figure shows that adversarial images cause a large shift in MAE loss (from blue to red). After using DRAM to repair, the loss distribution of adapted images shifts back and gets closer to the original images (red to yellow).
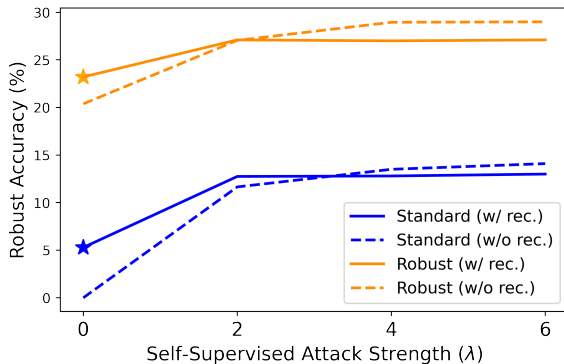

Figure 3. Defense-aware Attack: We show the trade-Off on the robust accuracy and SSL attack strength. When increasing the SSL attack strength λ, the robust accuracy increases, which means the attack is less effective. When the λ is set as zero, DRAM has the best performance on robust accuracy.

and classifiers at the same time.

4.5. Loss Analysis on Attack Detection

To better understand the ability of repair for MAE, we visualize the histogram on the self-supervised loss distribution for adversarial examples. We compare the loss variance between clean samples and different kinds of attack samples by collecting the loss on every batch during the testing-time inference. As Figure 2 shows, the loss variance on clean samples (colored with blue region) and the other eight attack types' samples (red curve) are large before repair. The orange curve shows the loss distribution of samples after re-

pair. Here, DRAM uses MAE loss to repair the samples and demonstrates a huge loss shifting after repair (colored with yellow region), which is closer to the loss distribution of clean samples. This result inspires us to leverage the MAE loss to detect the adversarial sample with KS-test.

5. Conclusion

We have presented a novel test-time defense called DRAM against (previously unknown) adversarial attacks. By using the self-supervision task MAE, DRAM effectively detects and repairs the adversarial samples during the inference time. Results show that DRAM achieves an average 82% detection score on eight previously unseen attacks. Our method for attack repair improved the robust accuracy of ResNet50 by up to 40% and 8% under standard and robust training after repair. In addition, we show DRAM is possible to defend against a defense-aware attack and delineate the trade-off between the attack budgets of the self-supervision task and the classification task. Overall, DRAM improves the robust accuracy across multi-type of adversarial attacks at the inference time, which provides new insight into the frontier of test-time defense for unseen attacks.

6. Acknowledgement

References

[1] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57, 2017.

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020.

[3] 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 *ACM Workshop on Artificial Intelligence and Security*, pages 15–26, 2017.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Everest Hinton. A simple framework for contrastive learning of visual representations. 2020.

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[8] Francesco Croce, Sven Gowal, Thomas Brunner, Evan Shelhamer, Matthias Hein, and Taylan Cemgil. Evaluating the adversarial robustness of adaptive test-time defenses. *arXiv preprint arXiv:2202.13711*, 2022.

[9] Virginia R de Sa. Learning classification with unlabeled data. *Advances in neural information processing systems*, pages 112–112, 1994.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[11] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[12] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, 32, 2019.

[13] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei A Efros. Test-time training with masked autoencoders. *arXiv preprint arXiv:2209.07522*, 2022.

[14] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[15] Yifan Gong, Yuguang Yao, Yize Li, Yimeng Zhang, Xiaoming Liu, Xue Lin, and Sijia Liu. Reverse engineering of imperceptible adversarial image perturbations. *arXiv preprint arXiv:2203.14145*, 2022.

[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021.

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[20] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in Neural Information Processing Systems*, 32, 2019.

[21] Shengyuan Hu, Tao Yu, Chuan Guo, Wei-Lun Chao, and Kilian Q Weinberger. A new defense against adversarial images: Turning a weakness into a strength. *Advances in Neural Information Processing Systems*, 32, 2019.

[22] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9865–9874, 2019.

[23] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. 2021.

[24] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

[25] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European conference on computer vision*, pages 577–593. Springer, 2016.

[26] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

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

[28] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[29] Chengzhi Mao, Mia Chiquier, Hao Wang, Junfeng Yang, and Carl Vondrick. Adversarial attacks are reversible with natural supervision. *arXiv preprint arXiv:2103.14222*, 2021.

[30] Chengzhi Mao, Amogh Gupta, Vikram Nitin, Baishakhi Ray, Shuran Song, Junfeng Yang, and Carl Vondrick. Multitask learning strengthens adversarial robustness. In *European Conference on Computer Vision*, pages 158–174. Springer, 2020.

[31] Chengzhi Mao, Ziyuan Zhong, Junfeng Yang, Carl Vondrick, and Baishakhi Ray. Metric learning for adversarial robustness. *Advances in Neural Information Processing Systems*, 32, 2019.

[32] Frank J. Massey. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

[33] Chen Min, Dawei Zhao, Liang Xiao, Yiming Nie, and Bin Dai. Voxel-mae: Masked autoencoders for pre-training large-scale point clouds. *arXiv preprint arXiv:2206.09900*, 2022.

[34] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[35] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020.

[36] Juan C Pérez, Motasem Alfarra, Guillaume Jeanneret, Laura Rueda, Ali Thabet, Bernard Ghanem, and Pablo Arbeláez. Enhancing adversarial robustness via test-time transformation ensembling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 81–91, 2021.

[37] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. *arXiv preprint arXiv:2103.12051*, 2021.

[38] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9229–9248. PMLR, 13–18 Jul 2020.

[39] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020.

[40] Yun-Yun Tsai, Chengzhi Mao, Yow-Kuan Lin, and Junfeng Yang. Self-supervised convolutional visual prompts. *arXiv preprint arXiv:2303.00198*, 2023.

[41] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

[42] Kexin Wang, Nils Reimers, and Iryna Gurevych. Tsdae: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning. *arXiv preprint arXiv:2104.06979*, 2021.

[43] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022.

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

[45] Zhiyuan Zeng, Keqing He, Yuanmeng Yan, Hong Xu, and Weiran Xu. Adversarial self-supervised learning for out-of-domain detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5631–5639, 2021.

[46] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[47] M. Zhang, S. Levine, and C. Finn. MEMO: Test time robustness via adaptation and augmentation. 2021.

[48] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13025–13032, 2020.

[49] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021.