MagNet is NOT Robust to Transfer Attacks

Pei-Hsuan Lu*, Pin-Yu Chen†, Kang-Cheng Chen♠, and Chia-Mu Yu*
*National Chung Hsing University, Taiwan
†AI Foundations Learning Group, IBM Thomas J. Watson Research Center, USA
♠Yuan Ze University, Taiwan

***Abstract*— In recent years, defending adversarial perturbations to natural examples in order to build robust machine learning models trained by deep neural networks (DNNs) has become an emerging research field in the conjunction of deep learning and security. In particular, MagNet consisting of an adversary detector and a data reformer is by far one of the strongest defenses in the *black-box* setting, where the attacker aims to craft transferable adversarial examples from an undefended DNN model to bypass a defense module without knowing its existence. MagNet can successfully defend a variety of attacks in DNNs, including the Carlini and Wagner's transfer attack based on the L_2 distortion metric. However, in this paper, under the black-box transfer attack setting we show that adversarial examples crafted based on the L_1 distortion metric can easily bypass MagNet and fool the target DNN image classifiers on MNIST and CIFAR-10. We also provide theoretical justification on why the considered approach can yield adversarial examples with superior attack transferability and conduct extensive experiments on variants of MagNet to verify its lack of robustness to L_1 distortion based transfer attacks. Notably, our results substantially weaken the existing transfer attack assumption of knowing the deployed defense technique when attacking defended DNNs (i.e., the *gray-box* setting).**

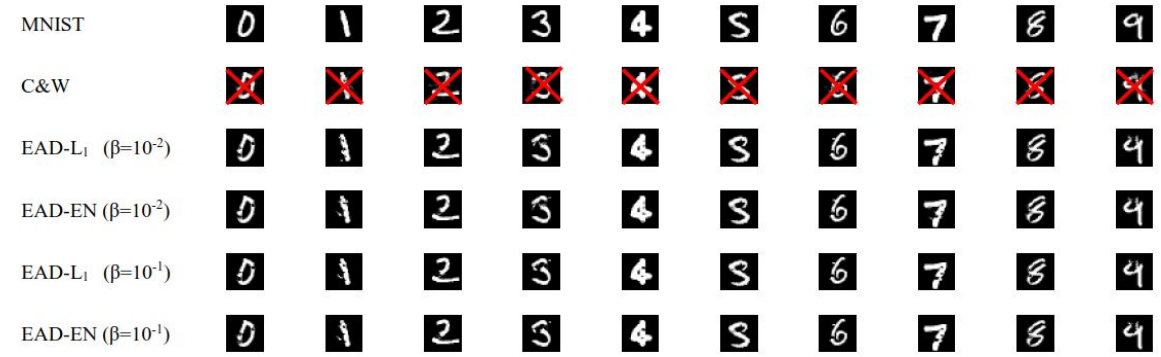
I. Introduction

DNNs are extensively used in many machine learning and artificial intelligence tasks, including computer vision, image classification, speech recognition, natural language processing, automated planning, and game playing, to name a few. However, recent studies have highlighted that well-trained DNNs, albeit achieving superior test accuracy on natural examples, are in fact quite vulnerable to *adversarial examples*. For example, carefully designed adversarial perturbations to natural images can cause state-of-the-art image classifiers trained by DNNs to misclassify, while the adversarial perturbations can be made visually imperceptible [1], [2]. Even worse, in addition to digital spaces, adversarial examples can also be crafted in physical world by means of realizing adversarial perturbations via colorful stickers [3], 3D printing [4], or robust image transformations [5]. Due to the existence and ease of generating adversarial examples from DNNs, the inconsistent decision making between DNN-based machine learning models and human perception as well as its robustness implications to safety-critical applications have given rise to the emerging research field intersecting deep learning and security.

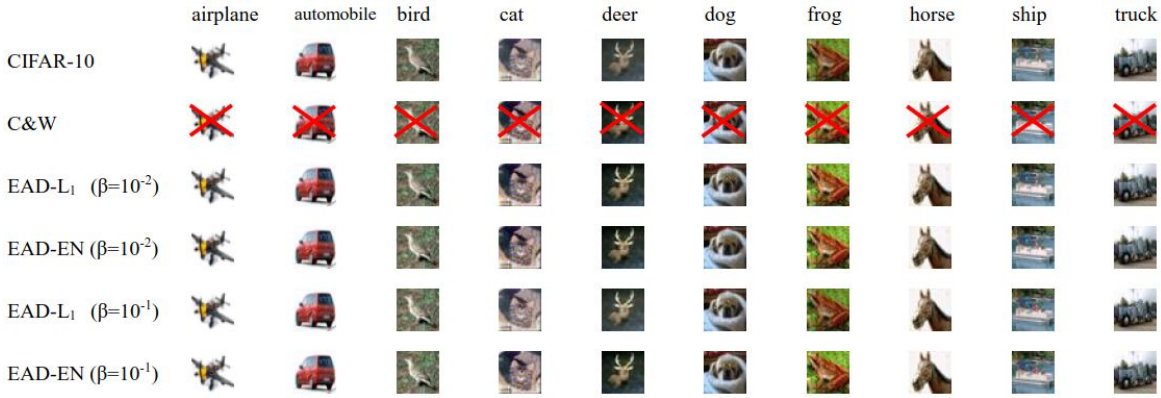
In the context of adversarial examples in DNNs, *attacks* refer to means of crafting visually indistinguishable adversarial perturbations to natural examples, whereas *defenses* refer to methods of mitigating adversarial perturbations for building a robust DNN model. For the task of image classification, *targeted attacks* aim to craft adversarial perturbations to render the prediction of the target DNN model towards a specific label, while *untargeted attacks* aim to find adversarial perturbations that will lead the target DNN model to a different prediction label than that of the natural example. More interestingly, the adversarial perturbations can be crafted even when the model details of the target DNN are totally unknown to an attacker, known as the black-box attack setting [6], [7].

An important and perhaps surprising property of adversarial examples in DNNs is their attack transferability - adversarial examples generated from one DNN can also successfully fool another DNN, which we call transfer attacks [1], [8], [9]. Transfer attacks are widely used for evaluating the performance of attacks and defenses against adversarial examples. In the defender's perspective, transfer attacks from an undefended DNN to a defended DNN serve as the baseline evaluation of the deployed defense techniques. In the attacker's foothold, executing a transfer attack is a preferable and practical option, as one can easily craft transferable adversarial examples from a DNN at hand to attack the target DNN without any prior knowledge of the target model. Although various defense methods such as defensive distillation [10], adversarial training [11] and different types of adversary detectors have been proposed to defend transfer attacks, they have been continuously broken or bypassed by the subsequent attacks. For example, defensive distillation fails to defend the transfer attack proposed by Carlini and Wagner [12], adversarial training in [11] is not effective against the transfer attack proposed by Chen et al. [13], and 10 different adversary detection methods are bypassed by Carlini and Wagner [14].

Notably, the attack framework established by Carlini and Wagner in [12], which we call *C&W attack* for short, is a strong attack that is capable of crafting highly transferable adversarial examples by tuning the *confidence* parameter. However, a recent defense method called MagNet [15], proposed by Meng and Chen, has demonstrated robust defense performance against C&W transfer attack under different confidence levels. In addition, MagNet can also defend transfer attacks using other methods including the fast gradient sign method (FGSM) [2], iterative FGSM [17], and DeepFool [18]. The success of MagNet in defending adversarial examples roots in its two complementary defense modules: (i) a *detector* that compares the statistical difference between an input image and the training data; and (ii) a *reformer* trained by an auto-encoder that regulates an input image to the data manifold of training examples. Generally speaking, the detector module declares



(a) MNIST



(b) CIFAR-10

Fig. 1: Visual illustration of transferable adversarial examples crafted by different attack methods from undefended DNNs to MagNet in the black-box setting [15]. Unsuccessful transfer attacks to MagNet are marked by red cross sign. EAD [16] can yield highly transferable and visually similar adversarial examples, whereas C&W attack [12] fails to bypass MagNet.

an input image as an adversarial example if its statistical distribution is significantly different from the training data. Otherwise, the input image further undergoes the reformer module, and the DNN will use the reformed example for label prediction. We defer the details of MagNet to Section II-A.

In this paper, we demonstrate that MagNet's defense against transfer attacks can be broken by elastic-net attacks to DNNs (EAD) proposed in [16]. The major difference of C&W attack and EAD is the distortion metric when crafting adversarial examples. C&W attack is a pure L_2 distortion based method, whereas EAD is a hybrid attack using both L_1 and L_2 distortion metrics. As will be explained in Section II-B, the use of L_1 distortion metric is able to filter out unnecessary perturbations to insignificant pixels and hence yielding adversarial examples with better attack transferability. Specifically, our experimental results of transfer attacks using EAD to MagNet with the default defense setting show that about 90% of adversarial examples on MNIST and 80% of those on CIFAR-10 can successfully fool MagNet, whereas under the same defense setting C&W attack only attains 10% and 52% attack success rate, respectively. For visual illustration, Figure 1 shows some adversarial examples that successfully break MagNet using EAD. These adversarial examples are still visually similar to the natural examples but will cause MagNet to misclassify. Furthermore, to corroborate that MagNet is indeed not robust to L_1 distortion based transfer attacks using EAD, we also conduct extensive experiments to evaluate the defense performance of MagNet under different defense settings, including tweaking the parameters of the detector module and changing the form of the reconstruction error when training the reformer.

It is also worth mentioning that this paper is the first work to verify the lack of robustness of MagNet to transferable adversarial examples in the *black-box* setting [15], where the attacker is completely unaware of the deployed defense mechanisms. In contrast, the recent work in [19] also claims to break MagNet but under a much stringent transfer attack assumption: the *gray-box* setting where the attacker knows the deployed defense technique but not the exact parameters. Specifically, in order to craft transferable adversarial examples, Carlini and Wagner modified their attack by leveraging the knowledge that auto-encoder is the primary defense technique used in MagNet [19]. Nonetheless, in the black-box setting MagNet is still effective in defending the original C&W attack in [12]. Consequently, our results substantially weaken the stringent gray-box attack assumption of knowing the deployed defense technique when attacking defended DNNs, which therefore provides novel and effective transfer attacks to DNNs.

II. Background and related work

A. MagNet: Defending Adversarial Examples using Reformer and Detector [15]

The essential component used in both the detector and the reformer of MagNet is the auto-encoder, denoted by $AE(\mathbf{x})$. The auto-encoder $AE(\mathbf{x})$ takes an image $\mathbf{x} \in \mathbb{R}^p$ as an input, compresses its information to a lower dimension, and then reconstructs the image $\mathbf{x}$ back to the original dimension. The default MagNet setting learns a AE by minimizing the mean squared error $\|\mathbf{x} - AE(\mathbf{x})\|_2$ averaged over all training examples. For an input image $\mathbf{x}$, let $F(\mathbf{x}) \in [0,1]^K$ denote a DNN image classifier of K classes, which outputs a K-dimensional probability distribution of class predictions. The defense of MagNet is a serial two-stage process. First, MagNet computes the Jensen-Shannon divergence (JSD) between $F(\mathbf{x})$ and $F(AE(\mathbf{x}))$ with a temperature parameter T, denoted by $JSD(F(\mathbf{x})/T), F(AE(\mathbf{x})/T)$. The input $\mathbf{x}$ is deemed adversarial if its JSD is greater than a certain threshold. Otherwise, $\mathbf{x}$ then undergoes the reformer $AE(\mathbf{x})$ before passing down to the DNN for classification. The reformer is responsible for projecting the input example to the data manifold learned by the auto-encoder such that the DNN is expected to yield correct label prediction after reforming the input example. Overall, MagNet uses the detector to filter out adversarial examples with statistically significant perturbations and relies on the reformer to rectify the adversarial examples with small perturbations (those who are not rejected by the detector) towards correct class prediction. In Section III we will evaluate the defense performance of the default MagNet setting and its robust variants.

B. EAD: Elastic-Net Attacks to DNNs [16]

Let $(\mathbf{x}_0, t_0)$ denote a natural example with a associated class label t_0 and let $(\mathbf{x}, t)$ denote its adversarial example with a target attack label $t \neq t_0$. The L_q norm of the image difference $\boldsymbol{\delta} = \mathbf{x} - \mathbf{x}_0$, defined as $\|\boldsymbol{\delta}\|_q = (\sum_{i=1}^p |\boldsymbol{\delta}_i|^q)^{1/q}$ when $q \geq 1$, is a widely used distortion metric between natural and adversarial examples. For targeted attacks, EAD finds an effective adversarial example by solving the following optimization problem:

$$\begin{aligned} &\text{minimize}_{\mathbf{x}} \quad c \cdot f(\mathbf{x}, t) + \|\mathbf{x} - \mathbf{x}_0\|_2^2 + \beta \|\mathbf{x} - \mathbf{x}_0\|_1 \\ &\text{subject to} \quad \mathbf{x} \in [0,1]^p, \end{aligned} \tag{1}$$

where the box constraint $\mathbf{x} \in [0,1]^p$ ensures every pixel value of $\mathbf{x}$ lies within a valid normalized image space, $c, \beta \geq 0$ are regularization parameters for f and L_1 distortion, respectively, and $f(\mathbf{x}, t)$ is the attack loss function defined as

$$f(\mathbf{x}, t) = \max\{\max_{j \neq t} [\mathbf{Logit}(\mathbf{x})]_j - [\mathbf{Logit}(\mathbf{x})]_t, -\kappa\}, \tag{2}$$

where $\mathbf{Logit}(\mathbf{x}) = [[\mathbf{Logit}(\mathbf{x})]_1, \ldots, [\mathbf{Logit}(\mathbf{x})]_K] \in \mathbb{R}^K$ is the logit of $\mathbf{x}$ (the internal layer representation prior to the softmax layer) in the considered DNN, also known as the *unnormalized probabilities*. The parameter $\kappa \geq 0$ is called the *confidence* that accounts for attack transferability. The hinge-like loss in (2) implies that the attack loss f is minimized when its unnormalized probability of being the target class t is κ lager that that of being the next possible class prediction. Similarly, for untargeted attacks, EAD uses the following attack loss function (dropping the notation t):

$$f(\mathbf{x}) = \max\{[\mathbf{Logit}(\mathbf{x})]_{t_0} - \max_{j \neq t} [\mathbf{Logit}(\mathbf{x})]_j, -\kappa\}. \tag{3}$$

Notable, C&W attack [12] is a special case of EAD when $\beta = 0$, resulting in a pure L_2 distortion based attack. We argue that considering the L_1 distortion (i.e., set $\beta > 0$) is crucial in crafting transferable adversarial examples, which can be explained by the fact that β plays the role of nulling unnecessary perturbations to insignificant pixels and shrinking the perturbation to important pixels, as indicated by C&W attack. Specifically, let $g(\mathbf{x}) = c \cdot f(\mathbf{x}, t) + \|\mathbf{x} - \mathbf{x}_0\|_2^2$ be the C&W attack objective function from (1) by setting $\beta = 0$. When solving (1) via gradient descent, EAD uses the iterative shrinkage-thresholding algorithm (ISTA) [20]:

$$\mathbf{x}^{(k+1)} = S_\beta(\mathbf{x}^{(k)} - \alpha_k \nabla g(\mathbf{x}^{(k)})), \tag{4}$$

where $\mathbf{x}^{(k)}$ is the k-th iterate with $\mathbf{x}^{(0)} = \mathbf{x}_0$, $\nabla g(\mathbf{x}^{(k)})$ denotes the gradient of g at $\mathbf{x}^{(k)}$, α_k denotes the step size, and $S_\beta : \mathbb{R}^p \mapsto \mathbb{R}^p$ is an pixel-wise projected shrinkage-thresholding function defined as

$$[S_\beta(\mathbf{z})]_i = \begin{cases} \min\{\mathbf{z}_i - \beta, 1\}, & \text{if } \mathbf{z}_i - \mathbf{x}_{0i} > \beta; \\ \mathbf{x}_{0i}, & \text{if } |\mathbf{z}_i - \mathbf{x}_{0i}| \leq \beta; \\ \max\{\mathbf{z}_i + \beta, 0\}, & \text{if } \mathbf{z}_i - \mathbf{x}_{0i} < -\beta, \end{cases} \tag{5}$$

for any $i \in \{1, \ldots, p\}$. Therefore, with the use of ISTA, at each iteration EAD retains the original pixel value $[\mathbf{x}_0]_i$ if the level of perturbation, indicated by $|[\mathbf{x}^{(k)} - \mathbf{x}_0 - \alpha_k \nabla g(\mathbf{x}^{(k)})]_i|$, is no greater than β. Otherwise, it shrinks the level of perturbation by β and projects the resulting pixel value to the box $[0,1]$ if $|[\mathbf{x}^{(k)} - \mathbf{x}_0 - \alpha_k \nabla g(\mathbf{x}^{(k)})]_i| > \beta$. Furthermore, since g is the attack objective function of C&W attack, EAD can be interpreted as a sparsity-induced C&W attack, where the ISTA step at each iterate adds zero perturbation to the i-th pixel if its C&W attack gradient $[\nabla g]_i$ is small (i.e., the pixel is deemed insignificant for attack), or reduces the perturbation by β if $[\nabla g]_i$ is large, leading to sharp adversarial examples with better attack transferability.

III. Experiments

In this section, we demonstrate EAD can frustrate MagNet on two popular image classification datasets - MNIST and CIFAR-10. MNIST is a popular handwritten digit dataset. Each image in the dataset represents a number from 0 to 9. For CIFAR-10, there are 10 image categories: airplanes, cars, birds, cats, deer, dogs, frogs, boats, trucks. Due to space limitations, we summarize the experimental results in the form of tables, while all plots regarding the performance of transfer attacks to MagNet are presented in the supplementary material. The visual illustrations of transferable adversarial examples crafted by different attack methods are displayed in Figure 1.

		MNIST				CIFAR-10			
Attack method	β	κ	ASR	L_1	L_2	κ	ASR	L_1	L_2
C&W(L_2)		15	10	3.553	1.477	20	52	3.675	0.126
EAD (EN rule)	10^{-3}	20	46.2	3.116	2.165	15	69.2	3.024	0.242
	10^{-2}	15	87.8	0.531	2.509	15	74.5	2.73	0.380
	$5 \cdot 10^{-2}$	15	90.1	0.266	2.730	15	77	2.810	0.544
	10^{-1}	15	90.2	0.433	2.803	15	78.6	3.234	0.681
EAD (L_1 rule)	10^{-3}	20	70.2	1.89	2.507	15	60.5	1.1718	0.327
	10^{-2}	15	84.5	0.449	2.701	15	66.7	1.646	0.495
	$5 \cdot 10^{-2}$	15	80.5	0.351	2.876	15	75.9	2.258	0.678
	10^{-1}	15	83.8	0.381	2.922	15	79.8	2.883	0.805

TABLE I: Comparison of different transfer attacks to MagNet (default setting) with different confidence κ on MNIST and CIFAR-10. ASR means attack success rate (%). The distortion metrics are averaged over successful examples.

A. Experiment Setup and Parameter Setting

We follow the black-box transfer attack setting used in MagNet [15] to implement untargeted transfer attacks from an undefended DNN to a defended DNN protected by MagNet, and hence in this setting the attack success rate (ASR) is equivalent to the misclassification rate of adversarial examples. The same DNN structure and training parameters in [15] are used to train the image classifiers on MNIST and CIFAR-10. We focus on the comparison between C&W attack (L_2 distortion based attack) and EAD (hybrid L_1 and L_2 distortion based attack) when evaluating the defense capability of MagNet, where we use the default settings provided by C&W attack[1] and EAD[2]. The best regularization parameter c is obtained via 9 binary search steps (starting from 0.001) and 1000 iterations are used for each attack with the same initial learning rate 0.01. For EAD, we report the attack results using different L_1 regularization parameter β and decision rules (elastic-net (EN) or L_1 distortion) for selecting the final adversarial example. On MNIST, we craft adversarial examples with different confidence level κ picked in the range of [0, 40]. On CIFAR-10, we generate adversarial examples with different confidence level κ picked in the range of [0, 100]. The default MagNet setting[3] and its robust variants are used for defense evaluation. On both MNIST and CIFAR-10, we randomly selected 1000 correctly classified images from the test sets to attack MagNet. All experiments are conducted using an Intel Xeon E5-2620v4 CPU, 125 GB RAM and a NVIDIA TITAN Xp GPU with 12 GB RAM.

B. Evaluation of Black-Box Transfer Attacks to MagNet

Although Meng et al. provide their training parameters and network structure which are used in their best experimental results, we cannot reproduce such effectively defensive results displayed in the MagNet paper [15]. Table I summarizes the statistics of adversarial examples crafted by different attacks under different confidence κ on MNIST and CIFAR-10. It is apparent that considering L_1 distortion when crafting adversarial examples indeed greatly improves attack transferability when compared to merely using L_2 distortion, as discussed in Section II-B. In addition to showing the attack results under the default MagNet setting, we also adjusted the defense parameters used in MagNet to make it more robust, which we call robust MagNet. However, we find that even MagNet's defensive capability can be improved, EAD can still effectively attack robust MagNet. In what follows, Section III-B1 and Section III-B3 discuss the experimental results of EAD attacking MagNet, and Section III-B2 and Section III-B4 discuss the experimental results of EAD attacking robust MagNet.

1) MagNet with default setting on MNIST: Because MNIST is a simple image classification task, Meng et al. [15] use only two detectors based on reconstruction error. Furthermore, the detectors respectively used the L_1 and L_2 norm to approximate the distance between the input and the training examples.

On MNIST, we adjusted the strength of attack by changing the confidence in the range from 0 to 40. Figure 2a shows the performance of C&W attack in the black-box transfer attack setting on MagNet. Figure 4 shows the performance of EAD under different decision rules and L_1 regularization parameter β on MagNet in the same black-box setting. The defense performance of MagNet is measured by the percentage of adversarial examples that are either detected by the MagNet's detector, or classified correctly by the classifier, which we call the classification accuracy of adversarial examples, and it is in contrast to the attack success rate.

When crafting adversarial examples, we can change the attack strength by adjusting the confidence level. The higher the confidence, the stronger the attack strength, and the greater the distortion. Figure 2a shows that the detector rejects more adversarial examples and hence becomes more effective as the confidence increases, which can be explained by the fact that the detector is designed to filter out the input example when it is far away from the data manifold of natural training examples. On the other hand, when the input example is close to the data manifold, the reformer is in charge of rectifying the input via the trained auto-encoder. The detector and reformer hence compliment each other and constitute MagNet. It can be observed in Figure 2a that most of the adversarial examples from C&W attack can be either easily detected or corrected by MagNet under the tested confidence values. Remarkably, the classification accuracy of C&W attack is above 90% at all confidence levels, meaning that C&W attack to MagNet is not effective in the black-box setting. On the other hand, Figure 4 shows the default MagNet fails to defend a majority of

[1]https://github.com/carlini/nn robust attacks.
[2]https://github.com/ysharma1126/EAD-Attack
[3]https://github.com/Trevillie/MagNet

adversarial examples crafted by EAD. For instance, comparing C&W attack with EAD when setting $\beta = 10^{-1}$ (Figure 4h) at the confidence 15, MagNet's classification accuracy reduces significantly from 90% to 9.7%, suggesting that approximately 90% of adversarial examples crafted by EAD with confidence 15 can bypass MagNet in the black-box setting. It it worth noting that since C&W attack only uses L_2 distortion while EAD uses both L_1 and L_2 distortion, this significant increase in attack success rate corroborates the effect of involving the L_1 distortion when crafting transferable adversarial examples.

In Figure 4, different L_1 regularization parameter β in EAD has a great influence on the attack performance. When β is small (e.g., $\beta = 10^{-3}$), we obtain a better performance under the L_1 decision rule than that under the EN rule because in this case L_2 distortion dominates the EN distortion. As β becomes larger, the attack performance of EAD under the EN rule is better than that under the L_1 rule, which can be explained by the potential over-contraction and aggressive thresholding for large β in the ISTA step of EAD. Moreover, we find that EAD under the EN rule can more effectively break MagNet as the L_1 regularization parameter β increases.

Interestingly, there is a dip in Figure 4 when the confidence levels are between 10 and 15 because in this range, the effectiveness of the reformer is diminishing and the detectors are yet ineffective. Obviously, the default MagNet is not strong enough to fill the dip.

2) EAD attack to robust MagNet on MNIST: The default setting of MagNet on MNIST contains only two detectors based on L_1 and L_2 reconstruction errors, respectively. We added two JSD detectors with temperature T of 10 and 40 in order to enhance MagNet's defense capability. In Figure 2b, this MagNet can achieve above 96% classification accuracy on adversarial examples generated by C&W L_2 attack at all confidence levels. Comparing Figure 2a with Figure 2b, we indeed can enhance the defense capability with the robust MagNet by adding JSD detectors. However, EAD can still effectively attack robust MagNet as shown in Figure 6. For example, approximately 40 % of adversarial examples crafted by EAD under EN rule with $\beta = 10^{-1}$ can still bypass MagNet in Figure 6h.

We also changed the number of filters used in a auto-encoder's convolution layer from 3 to 256 (see Table II). Auto-encoders can be more stable and achieve better performance on encoding and decoding by increasing the number of filters within a reasonable range. We find that this actually improves the robustness of MagNet. Comparing Figure 2a with Figure 2c, this change indeed leads to more effective defense against C&W attack, particularly in the confidence level ranging from 5 to 25. However, although we had a even robust MagNet, approximately 70 % of adversarial examples crafted by EAD under the EN rule with $\beta = 10^{-1}$ can still bypass MagNet, as shown in Figure 7h.

Figure 2d and Figure 8 show that MagNet can further improves its defensive ability by jointly changing the number of filter to 256 and adding the JSD detectors. Nonetheless, approximately 50 % of adversarial examples crafted by EAD under the EN rule with $\beta = 10^{-1}$ can still bypass MagNet, implying the lack of robustness in robust MagNet to L_1 distortion based adversarial examples.

To justify the vulnerability of MagNet to L_1 distortion based adversarial examples is not caused by the use of L_2 reconstruction error when training the auto-encoders in MagNet, Figure 10 compares the performance of different auto-encoders in MagNet trained by the mean squared error (L_2 loss) and the mean absolute error (L_1 loss) on the MNIST training set. We find that these auto-encoders used in MagNet can defend C&W L_2 attacks but are both susceptible to EAD.

Table II and Table III summarize the MNIST test accuracy under using MagNet and its robust variants. Table IV displays the best attack success rate of the EN and L_1 decision rules in EAD on MNIST. We conclude that the default and robust MagNet are able to defeat L_2 attacks, while they are still susceptible to L_1 attacks using EAD.

Detector I & Reformer		Detector II	
Conv.Sigmoid	$3 \times 3 \times 256$	Conv.Sigmoid	$3 \times 3 \times 256$
AveragePooling	2×2	Conv.Sigmoid	$3 \times 3 \times 256$
Conv.Sigmoid	$3 \times 3 \times 256$	Conv.Sigmoid	$3 \times 3 \times 1$
Conv.Sigmoid	$3 \times 3 \times 256$		
Upsampling	2×2		
Conv.Sigmoid	$3 \times 3 \times 256$		
Conv.Sigmoid	$3 \times 3 \times 1$		

TABLE II: Defensive architecture of robust MagNet on MNIST, including both encoders and decoders.

	Default [15]	Default+JSD	256	256+JSD
Without MagNet	99.42	99.42	99.42	99.42
With MagNet	99.13	97.75	99.24	97.55

TABLE III: MNIST test accuracy (%).

Decision rule	β	Default [15]	Default+JSD	256	256+JSD
EAD (EN rule)	10^{-3}	46.2	7.5	31.2	1.9
	10^{-2}	87.8	34	90.1	39.5
	$5 \cdot 10^{-2}$	90.1	51.6	93.6	60
	10^{-1}	90.2	55.6	94.3	65.1
EAD (L_1 rule)	10^{-3}	70.2	18.9	72.9	14.1
	10^{-2}	84.5	38.8	92.6	49.5
	$5 \cdot 10^{-2}$	80.5	48.8	90.3	62.6
	10^{-1}	83.8	51	92.1	66.3

TABLE IV: Best attack success rate (%) of the elastic-net (EN) and L_1 decision rules in EAD on MNIST.

3) CIFAR-10: Training and securing a classifier on CIFAR-10 is more challenging than that on MNIST. In Magnet, Meng et al. use two types of detectors based on L_1 and L_2 reconstruction errors as well as two JSD detectors with the temperature T of 10 and 40. Specifically, the JSD detectors are shown to be effective in detecting adversarial examples with large reconstruction errors.

We adjusted the strength of attack by changing the confidence in the range from 0 to 100. Figure 3a and Figure 5 show the performance of C&W L_2 attack and EAD on MagNet in the black-box setting, respectively.

On CIFAR-10, approximately 70% of the adversarial examples crafted by EAD will not be detected or corrected by

MagNet at confidence 10. Moreover, there is still a dip in the classification accuracy at the confidence levels ranging from 10 to 15. Despite using effective detectors on CIFAR-10, MagNet still provides an apparent attack opportunity to EAD in the confidence range of [10,15].

4) EAD attack to robust MagNet on CIFAR-10: We change the number of filters used in a auto-encoder's convolution layer from 3 to 256 (see Table V) and summarize the resulting CIFAR-10 test accuracy in Table VI. Comparing Figure 3a with Figure 3b, this robust MagNet aids in more effective defense against C&W L_2 attack at all confidence levels than the default MagNet on CIFAR-10. However, we report that EAD under the L_1 decision rule can still attain high attack success rate as β increases in both defense settings, as displayed in Table VII. Similar to the default MagNet on CIFAR-10, in Figure 9 there is still a dip in the classification accuracy of the robust MagNet at the confidence range [10,15], suggesting the lack of robustness to EAD.

Detectors & Reformer	
Conv.Sigmoid	$3 \times 3 \times 256$
Conv.Sigmoid	$3 \times 3 \times 256$
Conv.Sigmoid	$3 \times 3 \times 3$

TABLE V: Defensive architecture of robust MagNet on CIFAR-10, including both encoders and decoders.

	Default	256
Without MagNet	86.91	86.91
With MagNet	83.33	83.4

TABLE VI: CIFAR-10 test accuracy (%).

Decision rule	β	Default	256
EAD (EN rule)	10^{-3}	69.2	55.6
	10^{-2}	74.5	72
	$5 \cdot 10^{-2}$	77	86.3
	10^{-1}	78.6	91.5
EAD (L_1 rule)	10^{-3}	60.5	49.2
	10^{-2}	66.7	71.8
	$5 \cdot 10^{-2}$	75.9	90.9
	10^{-1}	79.8	93.7

TABLE VII: Best attack success rate (%) of the elastic-net (EN) and L_1 decision rules in EAD on CIFAR-10.

Figure 11 shows the defense performance of MagNet with different reconstruction errors in training the auto-encoders on the CIFAR-10 training set. Here we replaced the mean squared error with the mean absolute error and found that these auto-encoders can defend L_2 attacks but not EAD on CIFAR-10.

IV. Conclusion

We summarize the main results of this paper as follows:

- On MNIST, the default MagNet using the detectors merely based on reconstruction errors is not robust.
- The setting of auto-encoders has a great influence on MagNet's defensive ability.
- Despite its success in defending L_2 distortion based adversarial examples on MNIST and CIFAR-10, MagNet and its robust variants are ineffective against L_1 distortion based adversarial examples crafted by EAD. Furthermore, even though we implemented improved detectors in MagNet, EAD can still easily craft transferable adversarial examples that bypass MagNet in the black-box setting, which substantially weakens the attack assumption of knowing the deployed defense technique when attacking defended DNNs in the existing literature.

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," *arXiv preprint arXiv:1412.6572*, 2014.

[3] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and D. Song, "Robust physical-world attacks on machine learning models," *arXiv preprint arXiv:1707.08945*, 2017.

[4] A. Athalye and I. Sutskever, "Synthesizing robust adversarial examples," *arXiv preprint arXiv:1707.07397*, 2017.

[5] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.

[6] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *ACM Asia Conference on Computer and Communications Security*, 2017, pp. 506–519.

[7] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. 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*, 2017, pp. 15–26.

[8] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," *arXiv preprint arXiv:1611.02770*, 2016.

[9] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *arXiv preprint arXiv:1605.07277*, 2016.

[10] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *IEEE Symposium on Security and Privacy (SP)*, 2016, pp. 582–597.

[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[12] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57.

[13] Y. Sharma and P.-Y. Chen, "Attacking the madry defense model with l_1-based adversarial examples," *arXiv preprint arXiv:1710.10733*, 2017.

[14] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," *arXiv preprint arXiv:1705.07263*, 2017.

[15] D. Meng and H. Chen, "Magnet: a two-pronged defense against adversarial examples," *ACM CCS*, 2017.

[16] P.-Y. Chen, Y. Sharma, H. Zhang, J. Yi, and C.-J. Hsieh, "Ead: Elastic-net attacks to deep neural networks via adversarial examples," *arXiv preprint arXiv:1709.04114*, 2017.

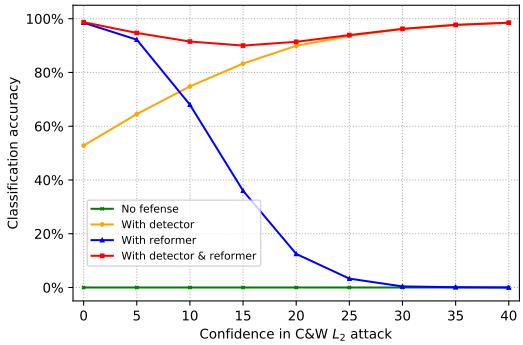
[17] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *ICLR'17; arXiv preprint arXiv:1611.01236*, 2016.

[18] 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.

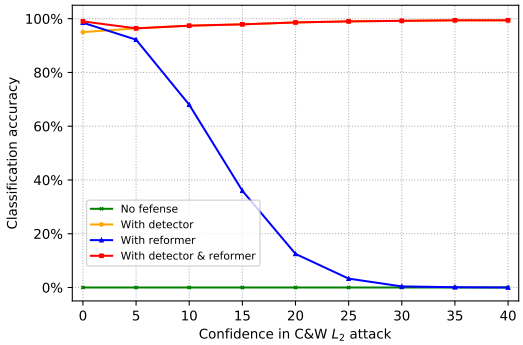
[19] N. Carlini and D. Wagner, "Magnet and" efficient defenses against adversarial attacks" are not robust to adversarial examples," *arXiv preprint arXiv:1711.08478*, 2017.

[20] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

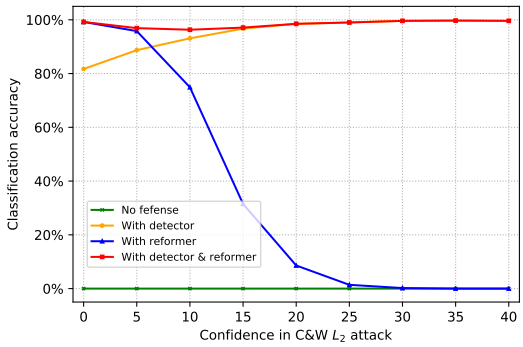
Supplementary Material: Defense Performance Plots of MagNet



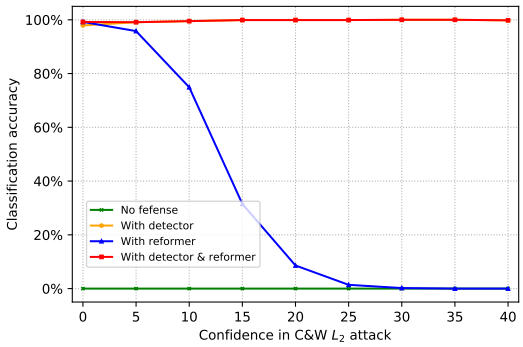
(a) default



(b) JSD

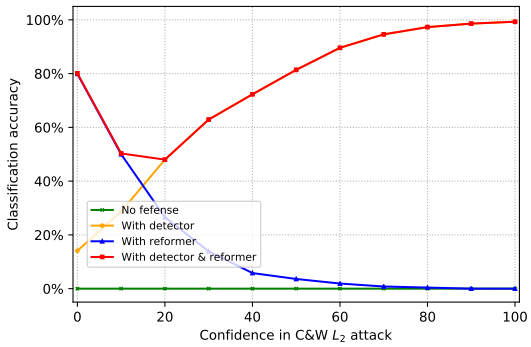


(c) 256

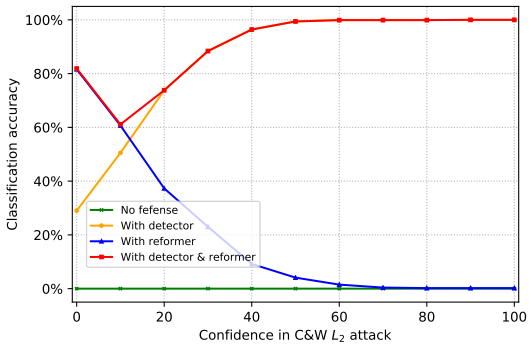


(d) 256+JSD

Fig. 2: C&W L_2 attack to MagNet under different auto-encoder structure on MNIST dataset with varying confidence.

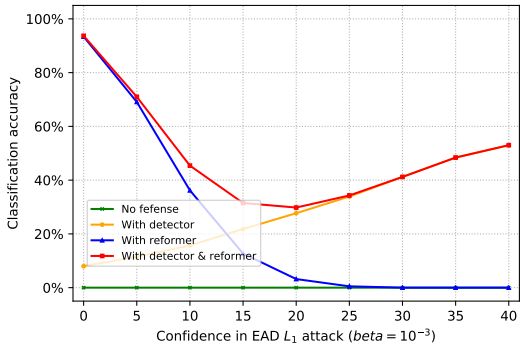


(a) default

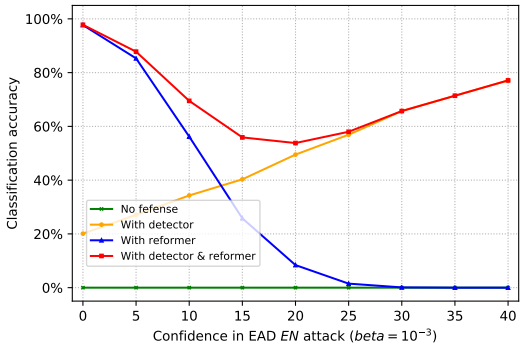


(b) 256

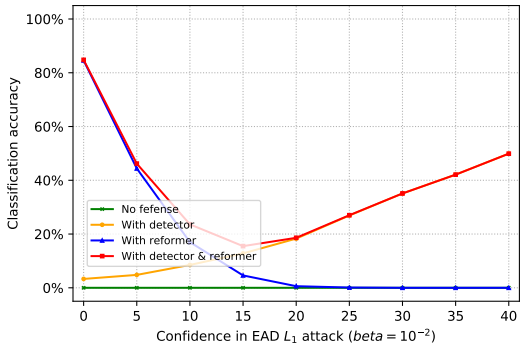
Fig. 3: C&W L_2 attack to MagNet under different auto-encoder structure on CIFAR-10 dataset with varying confidence.



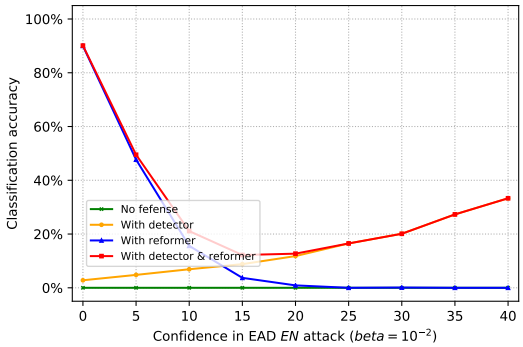
(a) L_1 decision rule $\beta = 10^{-3}$



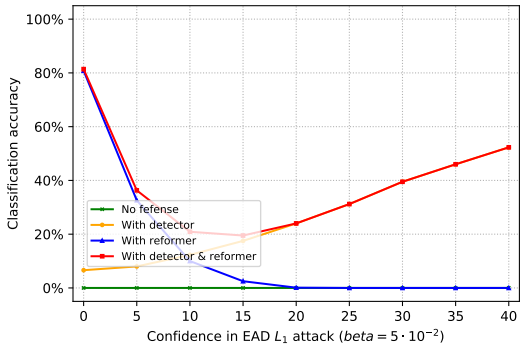
(b) EN decision rule $\beta = 10^{-3}$



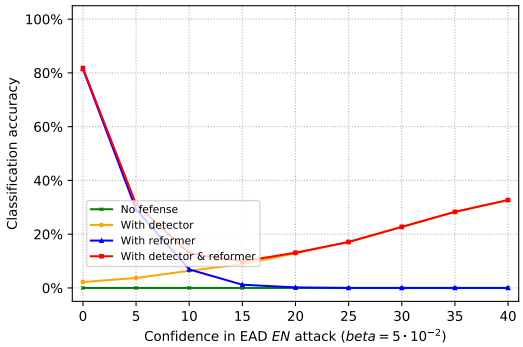
(c) L_1 decision rule $\beta = 10^{-2}$



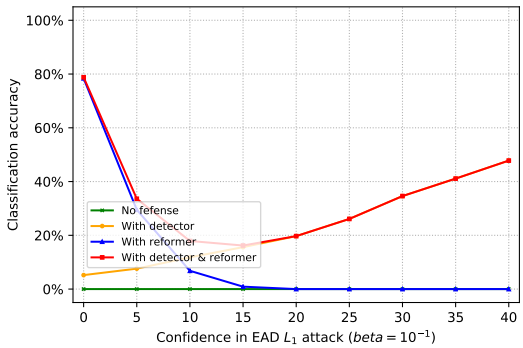
(d) EN decision rule $\beta = 10^{-2}$



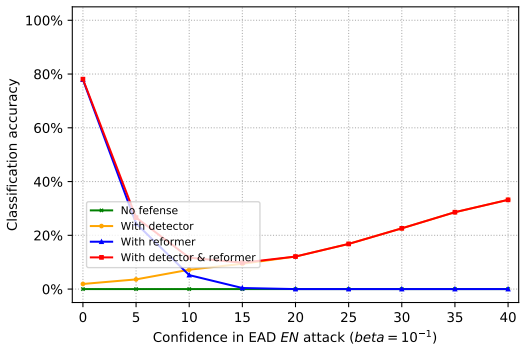
(e) L_1 decision rule $\beta = 5 \cdot 10^{-2}$



(f) EN decision rule $\beta = 5 \cdot 10^{-2}$

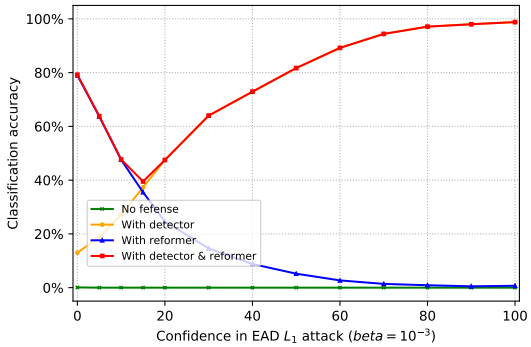


(g) L_1 decision rule $\beta = 10^{-1}$

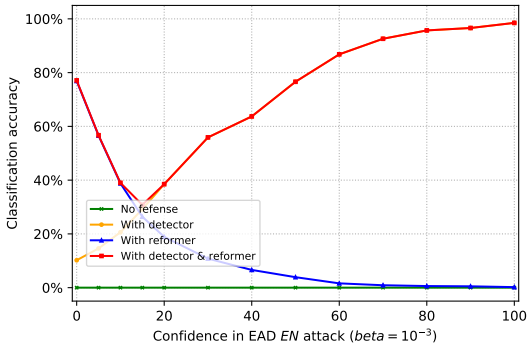


(h) EN decision rule $\beta = 10^{-1}$

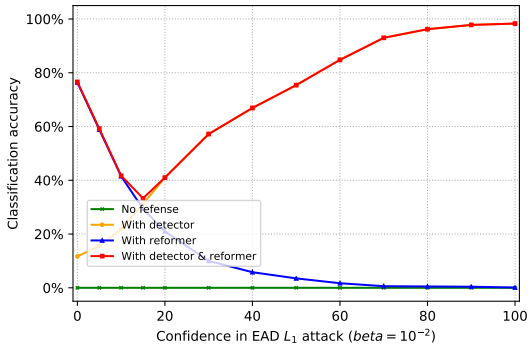
Fig. 4: EAD attacks on default MagNet under different β and different decision rules on MNIST with varying confidence.



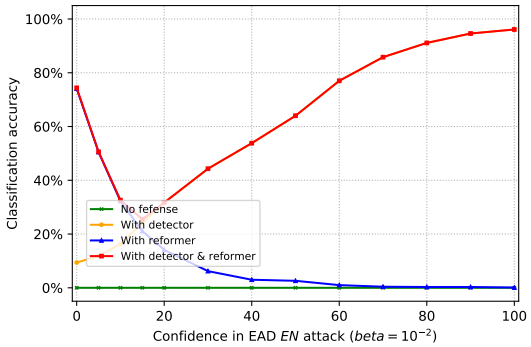
(a) L_1 decision rule $\beta = 10^{-3}$



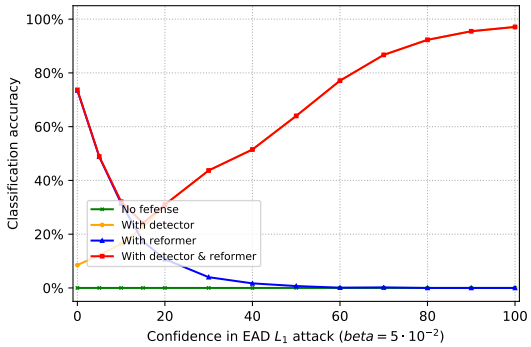
(b) EN decision rule $\beta = 10^{-3}$



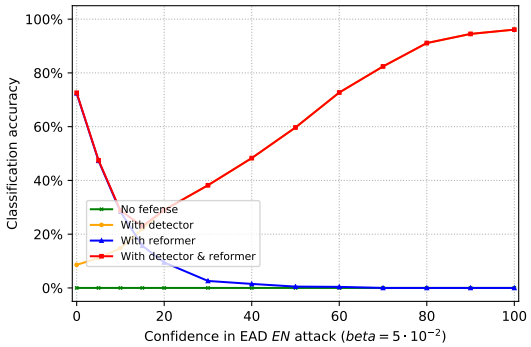
(c) L_1 decision rule $\beta = 10^{-2}$



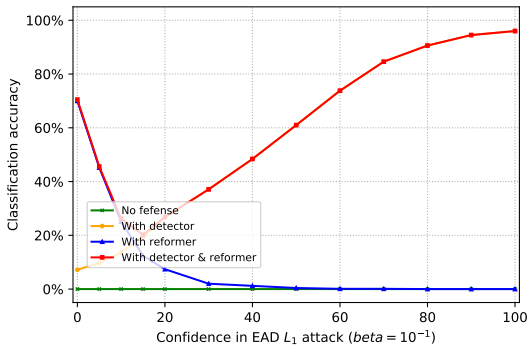
(d) EN decision rule $\beta = 10^{-2}$



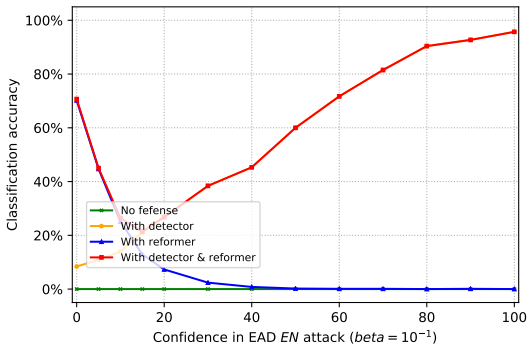
(e) L_1 decision rule $\beta = 5 \cdot 10^{-2}$



(f) EN decision rule $\beta = 5 \cdot 10^{-2}$

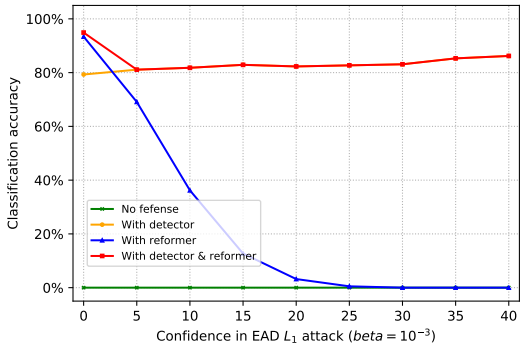


(g) L_1 decision rule $\beta = 10^{-1}$

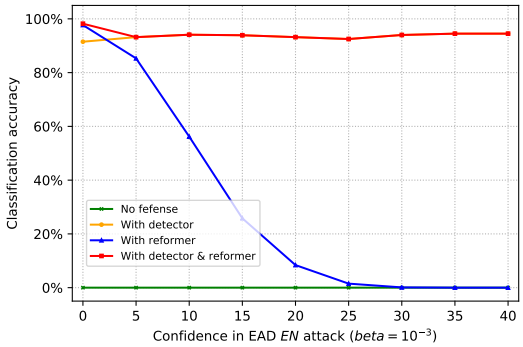


(h) EN decision rule $\beta = 10^{-1}$

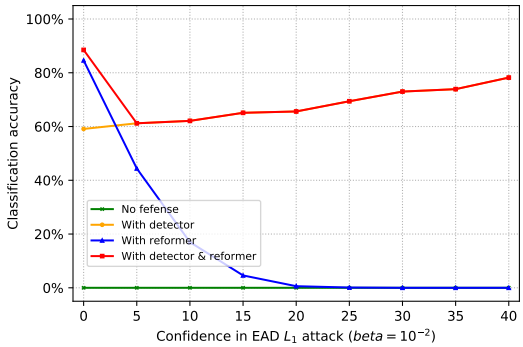
Fig. 5: EAD attacks on default MagNet under different β and different decision rules on CIFAR-10 with varying confidence.



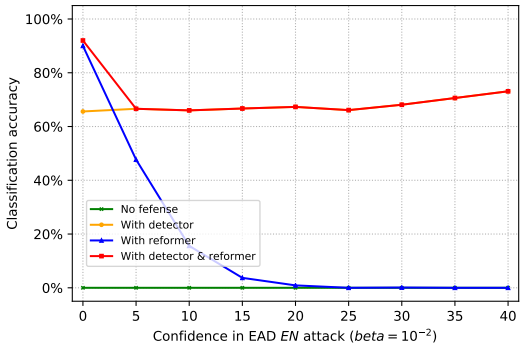
(a) L_1 decision rule $\beta = 10^{-3}$



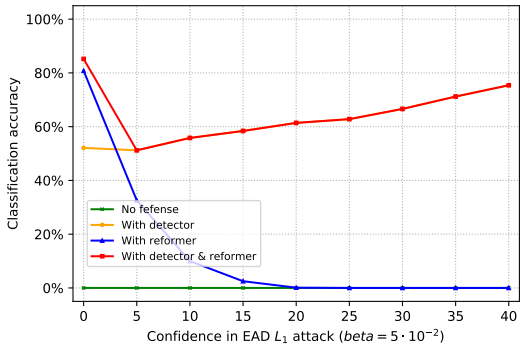
(b) EN decision rule $\beta = 10^{-3}$



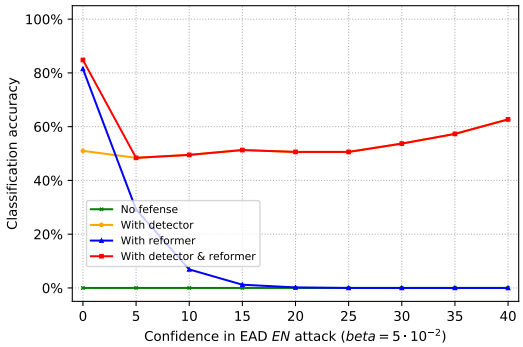
(c) L_1 decision rule $\beta = 10^{-2}$



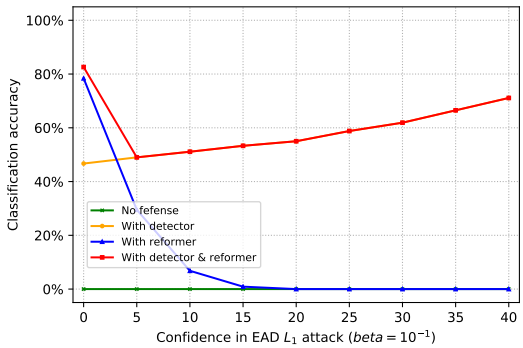
(d) EN decision rule $\beta = 10^{-2}$



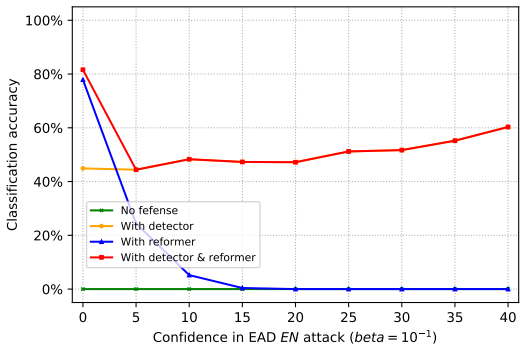
(e) L_1 decision rule $\beta = 5 \cdot 10^{-2}$



(f) EN decision rule $\beta = 5 \cdot 10^{-2}$

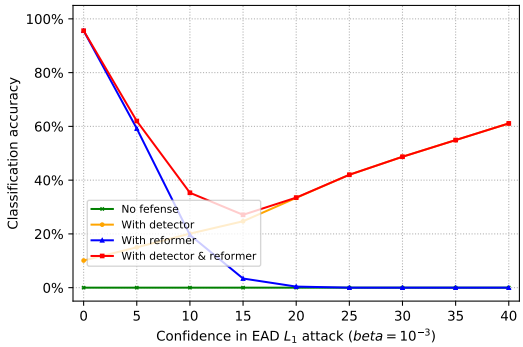


(g) L_1 decision rule $\beta = 10^{-1}$

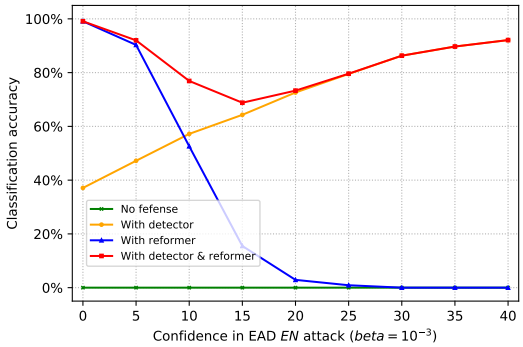


(h) EN decision rule $\beta = 10^{-1}$

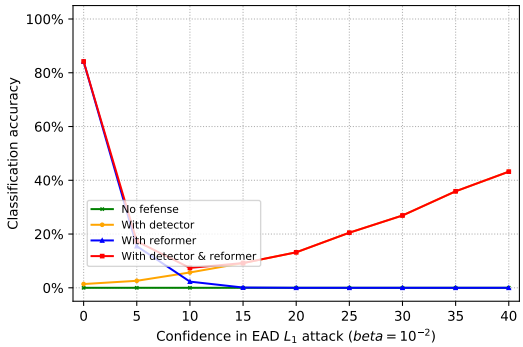
Fig. 6: EAD attacks on robust MagNet under different β and different decision rules on MNIST with varying confidence. Two JSD detectors are added into MagNet.



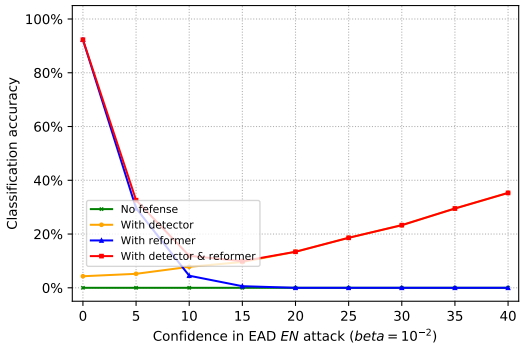
(a) L_1 decision rule $\beta = 10^{-3}$



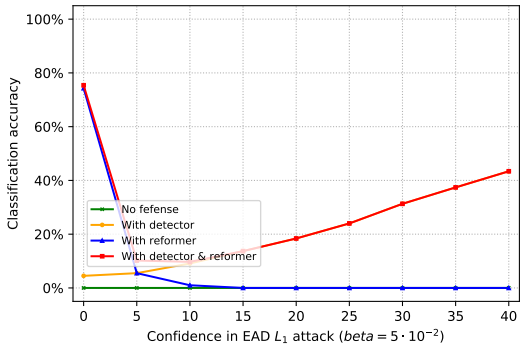
(b) EN decision rule $\beta = 10^{-3}$



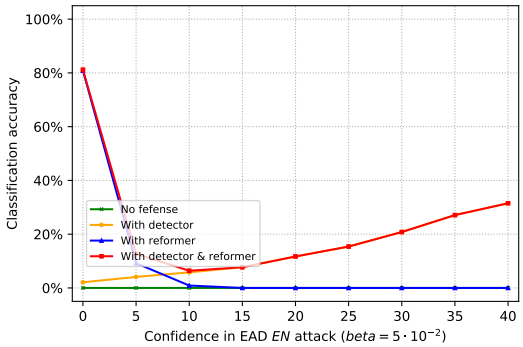
(c) L_1 decision rule $\beta = 10^{-2}$



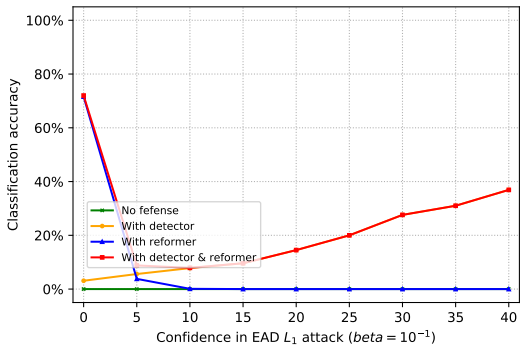
(d) EN decision rule $\beta = 10^{-2}$



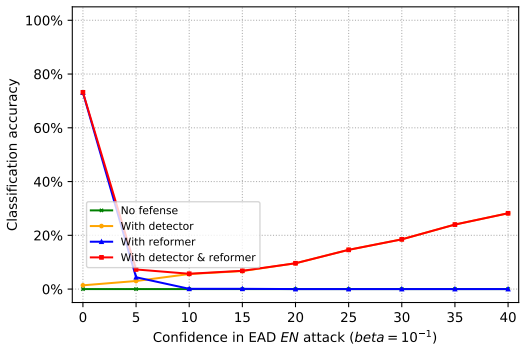
(e) L_1 decision rule $\beta = 5 \cdot 10^{-2}$



(f) EN decision rule $\beta = 5 \cdot 10^{-2}$

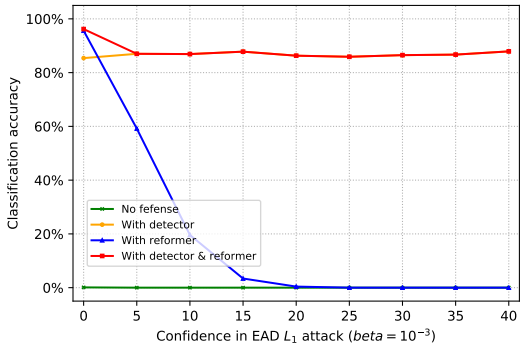


(g) L_1 decision rule $\beta = 10^{-1}$

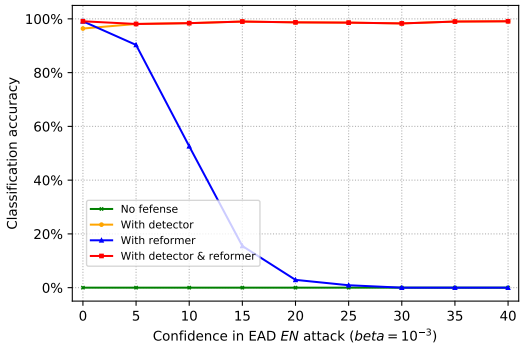


(h) EN decision rule $\beta = 10^{-1}$

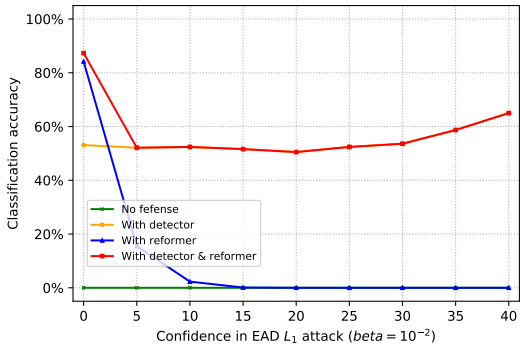
Fig. 7: EAD attacks on robust MagNet under different β and different decision rules on MNIST dataset with varying confidence. The number of filters in a auto-encoder's convolution layer is increased to 256.



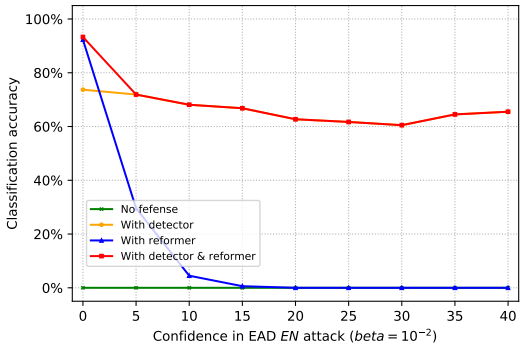
(a) L_1 decision rule $\beta = 10^{-3}$



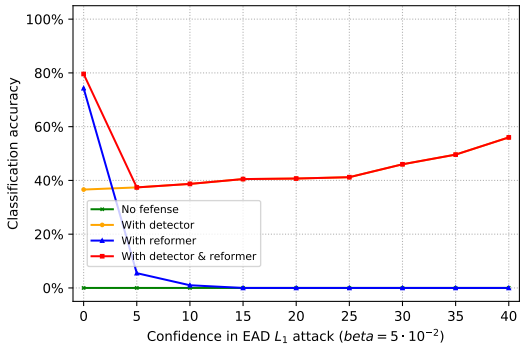
(b) EN decision rule $\beta = 10^{-3}$



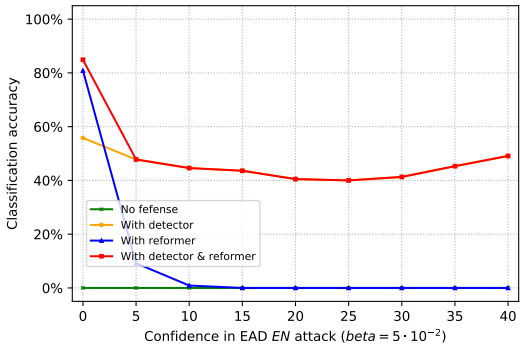
(c) L_1 decision rule $\beta = 10^{-2}$



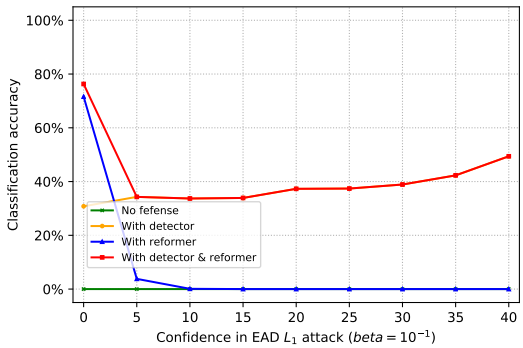
(d) EN decision rule $\beta = 10^{-2}$



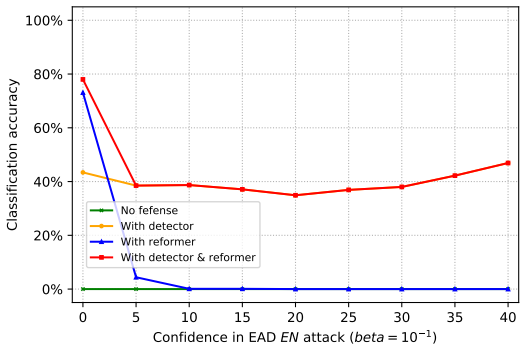
(e) L_1 decision rule $\beta = 5 \cdot 10^{-2}$



(f) EN decision rule $\beta = 5 \cdot 10^{-2}$

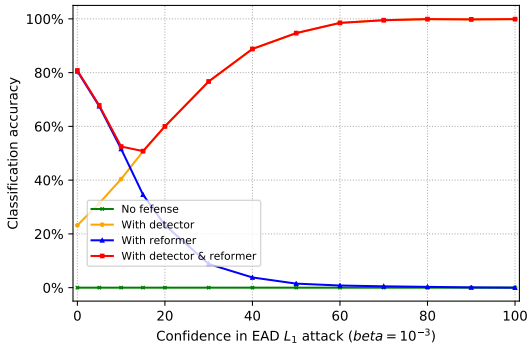


(g) L_1 decision rule $\beta = 10^{-1}$

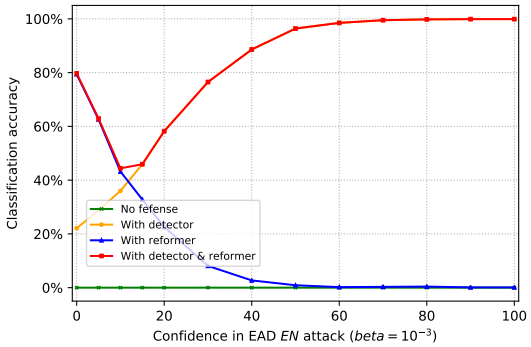


(h) EN decision rule $\beta = 10^{-1}$

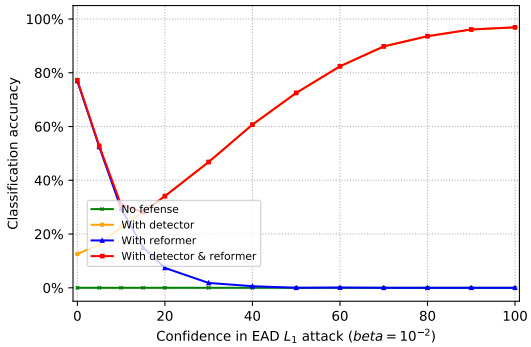
Fig. 8: EAD attacks on robust MagNet under different β and different decision rules on MNIST with varying confidence. The number of filters in a auto-encoder's convolution layer is increased to 256 and two JSD detectors are added into MagNet.



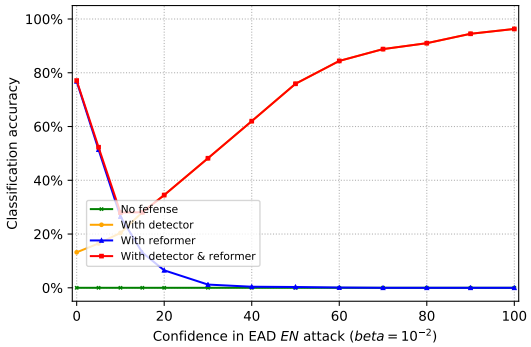
(a) L_1 decision rule $\beta = 10^{-3}$



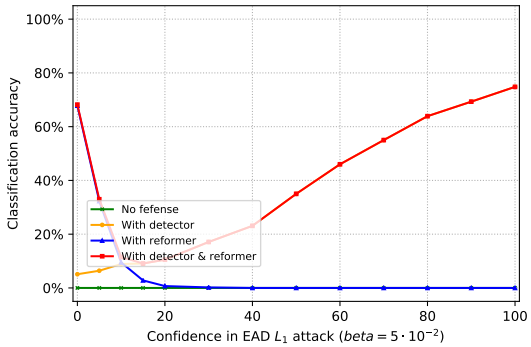
(b) EN decision rule $\beta = 10^{-3}$



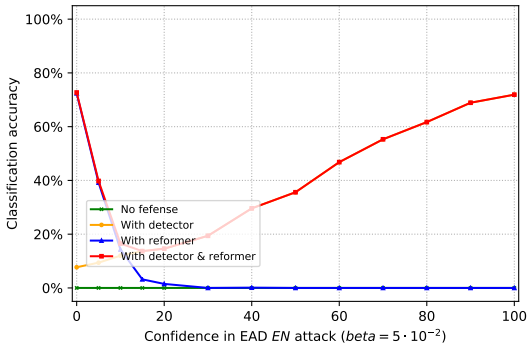
(c) L_1 decision rule $\beta = 10^{-2}$



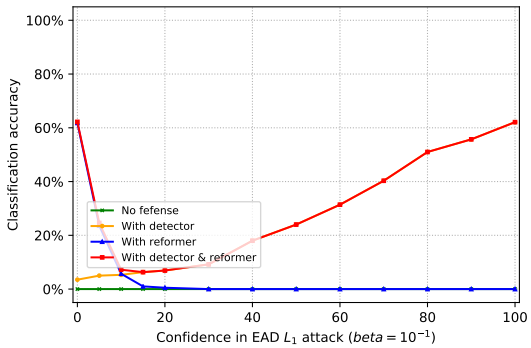
(d) EN decision rule $\beta = 10^{-2}$



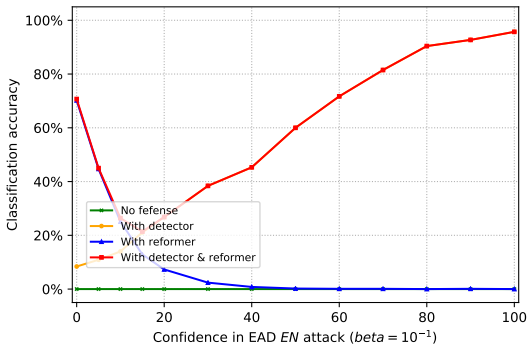
(e) L_1 decision rule $\beta = 5 \cdot 10^{-2}$



(f) EN decision rule $\beta = 5 \cdot 10^{-2}$

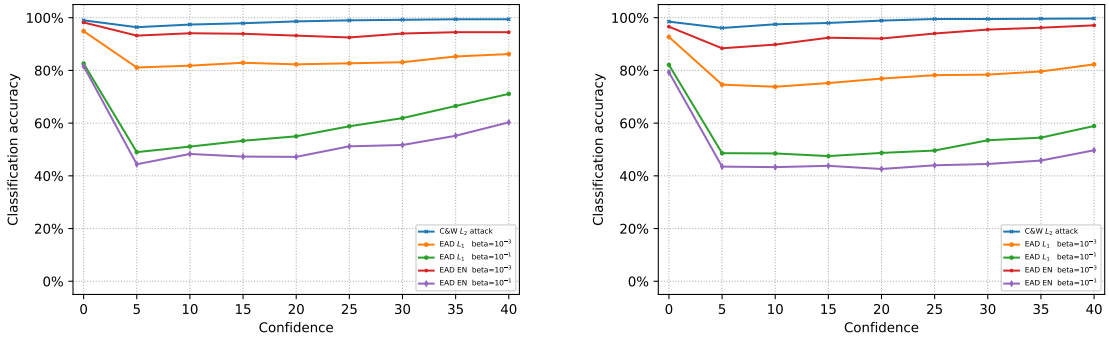


(g) L_1 decision rule $\beta = 10^{-1}$



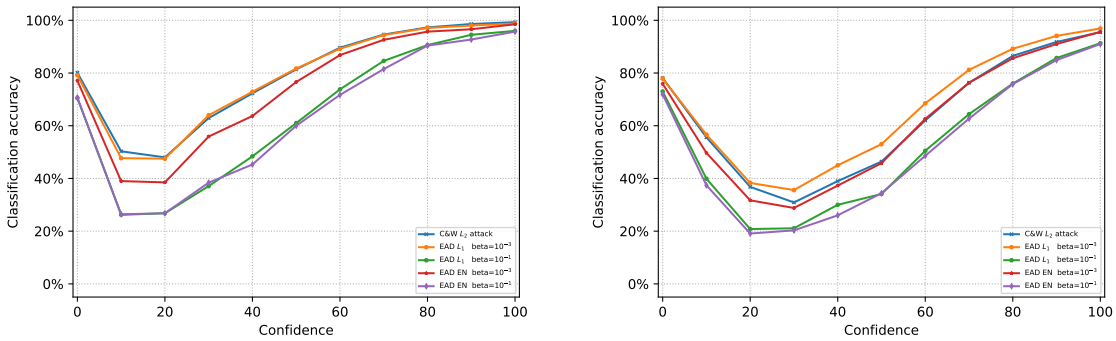
(h) EN decision rule $\beta = 10^{-1}$

Fig. 9: EAD attacks on robust MagNet under different β and different decision rules on CIFAR-10 with varying confidence. The number of filters in a auto-encoder's convolution layer is increased to 256.



(a) mean squared error (b) mean absolute error

Fig. 10: Comparison of training the auto-encoders with different loss functions on the MNIST training set.



(a) mean squared error (b) mean absolute error

Fig. 11: Comparison of training the auto-encoders with different loss functions on the CIFAR-10 training set.