Evaluating the Robustness of Defense Mechanisms based on AutoEncoder Reconstructions against Carlini-Wagner Adversarial Attacks

Petru Hlihor[1,2,*], Riccardo Volpi[1,*], and Luigi Malagò[1,*]

[1] Romanian Institute of Science and Technology, Cluj-Napoca, Romania
[2] MPI for Mathematics in the Sciences, Leipzig, Germany

Abstract

Adversarial Examples represent a serious problem affecting the security of machine learning systems. In this paper we focus on a defense mechanism based on reconstructing images before classification using an autoencoder. We experiment on several types of autoencoders and evaluate the impact of strategies such as injecting noise in the input during training and in the latent space at inference time. We test the models on Carlini-Wagner adversarial examples for the stacked system, composed by the autoencoder and the classifier, in the white-box scenario. Denoising autoencoders as well as injecting noise in the dataset before training and in the latent space at test time are effective strategies to improve the robustness of classifiers.

1 Introduction

Adversarial examples are a serious threat for machine learning systems. They can be divided in two main categories, white box-attacks [5, 12, 4, 15] in which the attacker has complete access to the model (topology of the network and its weights) and black-box attacks [16, 7, 20] in which the attacker has only access to the predictions of the network. Middle ground solutions are also present in the literature in which the model is partially hidden. Defense mechanisms can be found in the literature both in the form of adversarial training [5, 13] or of an input transformation at inference time, such as compression, random croppings, and/or reconstructions [6, 8, 19, 14]. Gu and Rigazio [6] proposed the use of an autoencoder to preprocess the images in input to a classifier, with the aim of cleaning input images from possible adversarial perturbations. Lately, Huang et al. studied this defense using Variational AutoEncoders [9]. Reconstructing adversarial examples generated for a classifier with an autoencoder yields performance close to the original one on clean examples [6, 9]. However, while the system is now robust to white-box adversarial examples for the classifier, the defense fails against white-box adversarial examples computed for the composite system formed by the stacked autoencoder+classifier [6]. Athalye et al. [1] showed more generally how obfuscating the gradients through transformations of the input (an example of which is the reconstruction through an autoencoder) does not constitute an effective defense strategy. In order to make claims about the robustness of a preprocessing strategy based on autoencoders, it is necessary to study the worst case scenario, where the attacker has full access to both networks. For this reason, in our study we use the Carlini-Wagner (CW) L^2 attack [4], considered one of the strongest in the literature [2, 3].

In this paper we study the desirable properties for a defense mechanism based on autoencoders. We present a detailed analysis of the robustness of the stacked network when using different types of autoencoders, such as vanilla AutoEncoders and Variational AutoEncoders. We evaluate the impact of denoising and contractive regularization on the latent space of the autoencoder and will show how this affects the robustness of the full system.

*{hlihor,volpi,malago}@rist.ro

https://doi.org/10.7557/18.5173

2 Autoencoders

In this section we present the different types of AutoEncoders considered in our study. Variational Autoencoders [10, 17] are a particular kind of autoencoders with a loss function derived from variational inference principles. The ELBO, which provides a lower bound for the log-likelihood is composed of two terms, the reconstruction loss term and a Kullback-Leibler divergence penalty term. The reconstruction term encourages mapping similar images close to each other in the latent space, while the KL penalty term concentrates all hidden representations in regions where the prior has high probability density. The trade-off between these two terms provides a regularization of the space during training.

Denoising autoencoders [26] are trained to reconstruct corrupted images to the corresponding clean ones, differently from regular autoencoders which are trained to compute the identity map. Denoising is thus a more difficult task, but the reconstructions generated by these models are less blurry and better resemble natural images, which can make them easier to classify.

Contractive Autoencoders [18] penalize sharp changes in the latent representations caused by small changes in the input. The regularization term is the squared L^2-norm of the Jacobian of the function $x \mapsto h$, where x is the input and h is the hidden representation. We propose to adapt the contractive penalty for Variational Autoencoders, so that the loss function becomes

$$-\text{ELBO}(x) + \alpha \left\| \frac{\partial \mu_i}{\partial x_j} \right\|_2^2 + \beta \left\| \frac{\partial \sigma_i^2}{\partial x_j} \right\|_2^2 \qquad (1)$$

where α and β are parameters. For the purposes of this paper, we will set $\beta = 0$, since having sharp changes in the mean is a behavior that we believe is more likely to lead to differences in the reconstructions of two similar images. We experiment with vanilla AutoEncoders (AE), Variational AutoEncoders (VAE), as well as with their denoising (DAE, DVAE) and contractive versions (CAE, CVAE).

3 The Carlini-Wagner Attack

We generate adversarial examples using the Carlini Wagner L^2 attack (CW) [4], which is considered to be a strong attack in the literature, having broken many defenses [2, 3]. CW is formulated as an optimization problem with a loss function composed of two terms, the first one is the L^2 norm of the perturbation, therefore encouraging finding subtle adversarial examples, and the second one is an adversarial loss which encourages the resulting image to be as harmful as possible. In the literature of adversarial examples [5, 13], the threat model usually considered is where the attacker is allowed to modify the input with a perturbation of a certain maximal L^∞ norm equal to ϵ. This is equivalent to being able to modify each pixel of an image by a maximum amount ϵ. We created CW adversarial examples that obey this constraint. At each iteration of gradient descent, we project the current perturbation on the L^∞ ϵ-box, centered in zero. For VAEs, we consider the worst-case scenario attack, where the attacker has disabled sampling in the latent layer when generating adversarial examples. By enabling sampling, noisy gradients would be obtained resulting in a weaker attack.

4 Defense Mechanisms

The defense strategy studied in this paper is based on preprocessing the input to a classifier with an autoencoder [6, 9]. The idea is that a well trained autoencoder should be able to learn the relevant features of an image and reconstruct it properly, by removing the adversarial perturbation. Moreover, the autoencoder should also be robust enough not to become the target of an adversarial attack itself [21, 11].

4.1 Injection of Noise in the Input

When training VAEs, adding noise to the input is not just a form of data augmentation, but it is also a recommended practice to guarantee numerical stability and to avoid overfitting [24, 25]. Due to the KL regularization in the latent space of a VAE, the approximate posterior is encouraged to be close to the Gaussian prior $\mathcal{N}(0, 1)$. If we do not train with noise, most clean samples will be mapped to latent representations having high probability density with respect to the Gaussian prior, while perturbed images will be mapped outside this region. Training with noise is thus necessary to be able to learn a

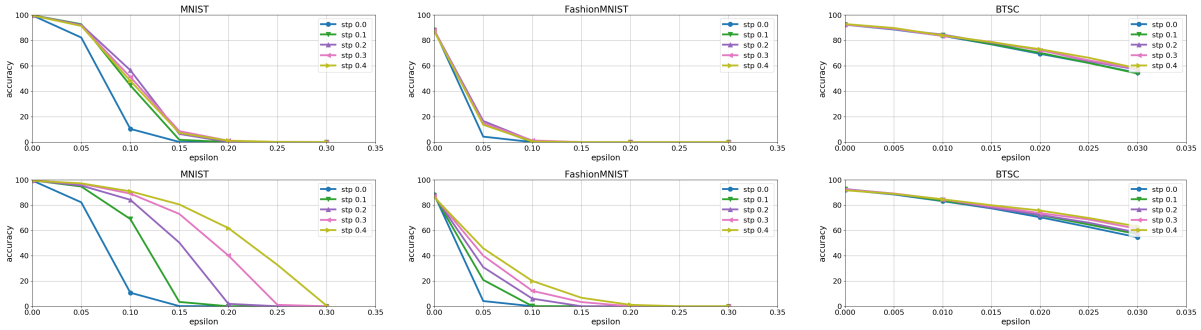

Figure 1: Accuracy of autoencoder+classifier on CW adversarial images vs maximum L^∞ norm. AE (top) and DAE (bottom) for different Gaussian noise standard deviations during training (stp).

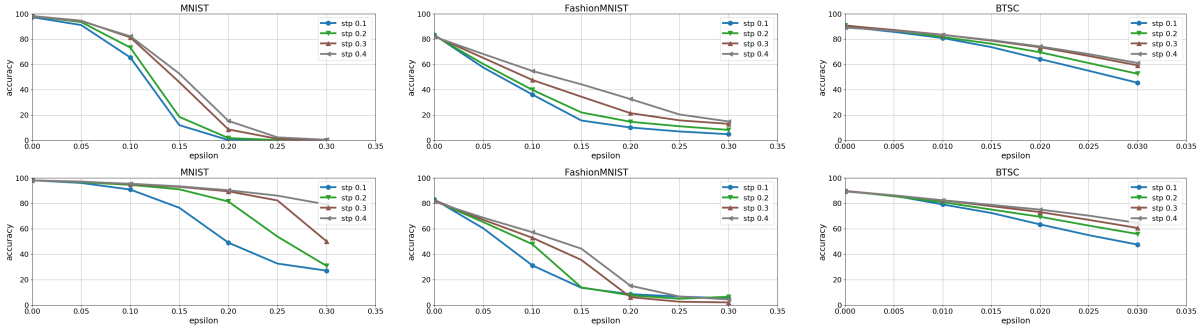

Figure 2: Accuracy of autoencoder+classifier on CW adversarial images vs maximum L^∞ norm. VAE (top) and DVAE (bottom) for different Gaussian noise standard deviations during training (stp). The scaling factor for the sampling in the latent space is fixed to 1.

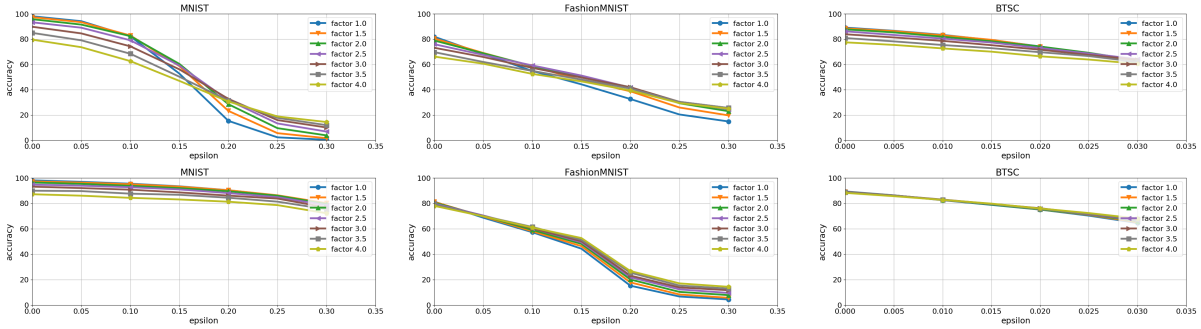

Figure 3: Accuracy of autoencoder+classifier on CW adversarial images with different maximum L^∞ norm. VAE (top) and DVAE (bottom) with stp equal to 0.4 for various scaling factors.

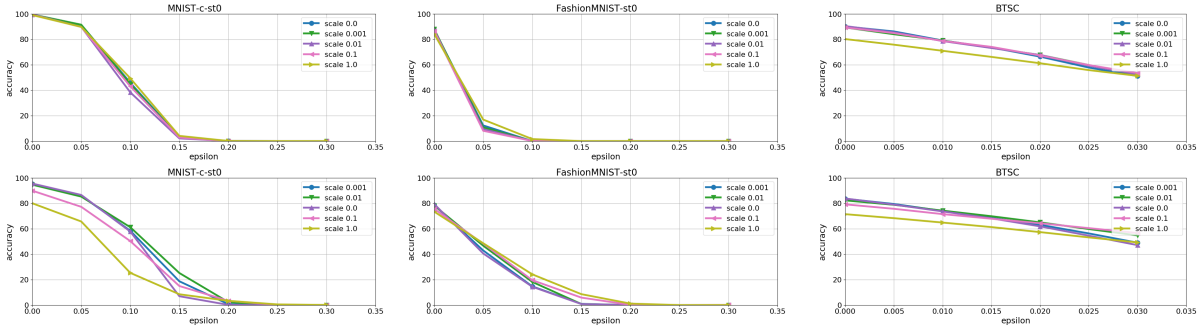

Figure 4: Accuracy of autoencoder+classifier on CW adversarial images with different maximum L^∞ norm. CAE (top) and CVAE (bottom) with stp equal to 0.1 for various penalty coefficients.

latent space in which both clean and perturbed images are mapped to nearby points, and eventually lead to similar reconstructions.

4.2 Increasing the Variance in the Latent Layer

The idea of including stochastic layers in inference in order to protect from adversarial examples has been explored in the literature [28, 22]. The encoder of a VAE has a stochastic layer which learns the parameters of a multivariate Gaussian distribution with diagonal covariance matrix from which the latent representation is sampled. The diagonal covariance matrix of the multivariate Gaussian can be scaled by a certain factor which is a hyperparameter of the defense. Increasing the entries of the covariance matrix by a factor corresponds to adding Gaussian noise to the latent representation after sampling, perturbing the internal representation of a VAE. By doing this, we hope to find an advantageous trade-off between quality of reconstruction and robustness of the classifier to an attack, by perturbing it enough to escape the adversarial region.

5 Results

We made experiments on the MNIST, FashionMNIST [27], and Belgian Traffic Signs for Classification (BTSC) [23] datasets. For each dataset, we trained classifiers with 4 convolutional layers, followed by a fully connected layer, and a softmax layer. We trained AEs and VAEs whose encoders and decoders have 2 convolutional layers with 32 channels each, 3x3 filters, latent size 128 and ReLU activation. In the training of AEs and VAEs, we injected Gaussian noise with standard deviation in the range [0.1, 0.4], also referred to as the stochastic parameter (stp). For each combination of stacked autoencoder+classifier, we created sets of adversarial examples, one for each value of ϵ in the range $[0.05, 0.3]$ for the MNIST and FashionMNIST datasets and in $[0.005, 0.03]$ for the BTSC dataset. We ran the CW attack over the first 1,000 images in the test sets of MNIST and FashionMNIST and over 1,000 randomly selected images from the test set of BTSC, for 100 iterations, with learning rate 0.1, and parameters $\gamma = 1, k = 0$. In Figs. 1-4 we present the accuracy as a function of the maximum L^∞ norm (ϵ) allowed for the attack. Training with denoising significantly improves the accuracy in some cases (DAE on MNIST and FashionMNIST, and DVAE on MNIST), while in others does not seem to negatively influence the results. Autoencoders trained with high stochastic parameter (stp) make the stacked networks more robust, both for AEs and for VAEs. The contribution of the stochastic parameter (stp) seems more pronounced for VAE, DVAE, and DAE, especially on MNIST and FashionMNIST (Fig. 1, 2). Rescaling the covariance of the approximate posterior (Fig. 3) improves the robustness of the stacked network for large perturbations. A trade-off exists though, since large scaling factors degrade performance on clean examples. A contractive penalty can be slightly beneficial, for example in the case of CVAE on MNIST and FashionMNIST (Fig. 4). However, in the other cases, this did not bring a significant improvement or it decreased the accuracy of the system.

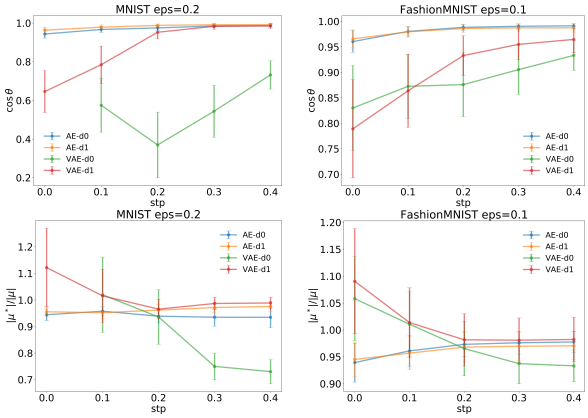

Figure 5: Plots of quantities of interest in the latent space for the different models vs standard deviation of the noise used during training (stp): (top) cosine of the angle between μ^* and μ, (bottom) ratio between the norm of μ^* and the norm of μ, depending on the magnitude of the perturbation (eps). Plots are for a fixed ϵ, specified in the picture; d0 and d1 denote no denoising and denoising, respectively.

In Fig. 5 we show the impact of adversarial perturbations on the latent space analyzing two quantities of interest: the cosine between μ (mean of clean) and μ^* (mean of adversarial), and the ratio of the norms $\frac{\|\mu^*\|_2}{\|\mu\|_2}$. We choose ϵ as the value of interest for each dataset in which performance starts to degrade. The diameter of the set (i.e., the maximum distance between points) of the encoded μ is about 3 times bigger for AE than for VAE. This diameter is growing for DAE while stays about constant for DVAE, since the latent representation of VAE is regularized by the prior. The perturbations in the latent space have a similar magnitude instead, hence the cosine of the angle and the ratio of the norms are more stable for AEs than for VAEs (Fig. 5). DVAE attenuates this by reducing the displacements in the latent space and seem to provide a more stable representation.

6 Conclusions

We studied the efficacy of a defense based on reconstructing images with different types of autoencoders and explored the role of some hyperparameters. Denoising models are more robust than their corresponding non-denoising versions. Increasing the magnitude of noise used to corrupt images in training leads to more robust stacked networks. The scaling factor in the latent space introduces a trade-off between the accuracy on clean images and the robustness on adversarial examples. The contractive penalty can improve the performance in some cases, while in others it is not beneficial. The diameter of the set of the encodings is much larger in AEs than in VAEs (due to the KL regularization term). The relative adversarial perturbations in the latent space are smaller in AEs than in VAEs. While we could not clearly correlate this fact with a neat increase in the accuracies for AEs, we plan to investigate better this aspect to leverage the stability of the latent space to our advantage. Possible limitations of this study include having used relatively easy-to-learn datasets, instead of more complex ones such as CIFAR-10 or ImageNet, not having tested residual networks as classifiers and not having used other well-known strong attacks, such as BIM [12] or DeepFool [15]. These will be all topics for future study.

7 Acknowledgements

The authors are supported by the DeepRiemann project, co-funded by the European Regional Development Fund and the Romanian Government through the Competitiveness Operational Programme 2014-2020, Action 1.1.4, project ID P_37_714, contract no. 136/27.09.2016.

References

[1] A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *ICML*, 2018.

[2] N. Carlini and D. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *AISEC*, 2017.

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

[4] N. Carlini and D. A. Wagner. Towards evaluating the robustness of neural networks. *IEEE S&P*, 2016.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *ICLR*, 2015.

[6] S. Gu and L. Rigazio. Towards deep neural network architectures robust to adversarial examples. *ICLR*, 2015.

[7] C. Guo, J. R. Gardner, Y. You, A. G. Wilson, and K. Q. Weinberger. Simple black-box adversarial attacks. *ICML*, 2019.

[8] C. Guo, M. Rana, M. Cisse, and L. van der Maaten. Countering adversarial images using input transformations. *ICLR*, 2018.

[9] U. Hwang, J. Park, H. Jang, S. Yoon, and N. I. Cho. Puvae: A variational autoencoder to purify adversarial examples. *IEEE Access*, 2019.

[10] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *ICLR*, 2014.

[11] J. Kos, I. Fischer, and D. Song. Adversarial examples for generative models. *IEEE Symposium on Security and Privacy Workshops*, 2018.

[12] A. Kurakin, I. J. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *ICLR*, 2017.

[13] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *ICLR*, 2018.

[14] D. Meng and H. Chen. Magnet: a two-pronged defense against adversarial examples. In *CCS*, pages 135–147, 2017.

[15] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. *CVPR*, 2016.

[16] N. Papernot, P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against deep learning systems using adversarial examples. *ASIACCS*, 2017.

[17] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *ICML*, 2014.

[18] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *ICML*, pages 833–840, 2011.

[19] P. Samangouei, M. Kabkab, and R. Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *ICLR*, 2018.

[20] J. Su, D. V. Vargas, and S. Kouichi. One pixel attack for fooling deep neural networks. *TEVC*, 2019.

[21] P. Tabacof, J. Tavares, and E. Valle. Adversarial images for variational autoencoders. *NIPS*, 2016.

[22] O. Taran, S. Rezaeifar, T. Holotyak, and S. Voloshynovskiy. Defending against adversarial attacks by randomized diversification. In *CVPR*, pages 11226–11233, 2019.

[23] R. Timofte and L. Van Gool. Sparse representation based projections. In *BMVC*, 2011.

[24] B. Uria, I. Murray, and H. Larochelle. Rnade: The real-valued neural autoregressive density-estimator. In *NIPS*, pages 2175–2183, 2013.

[25] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. *ICML*, 2016.

[26] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, 2008.

[27] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv:1708.07747*, 2017.

[28] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille. Mitigating adversarial effects through randomization. *ICLR*, 2018.