Adversarial Attacks on Variational Autoencoders

Anonymous Author(s)
Affiliation
Address
email

Contributions of latinx individuals: all three authors identify as latinx.

Contributions of the presenter: the main idea on how to attack variational autoencoders, how to measure the attack resistance trade-off and its correspoding visualization, partial writing of the final text, editing.

Adversarial attacks derail models by crafting malicious inputs. Image classifiers mislabel those inputs — visually indistinguishable from ordinary ones — with high confidence.

In comparison to the extensive literature on adversarial attacks for classifiers [1, 2, 3, 4, 5, 6], attacks for autoencoders are mostly unexplored, possibly because those attacks are hard both to perform and to assess [7, 8]. Still, as autoencoders are advanced as powerful schemes for compressing information [9], attacks on them are potentially at least as dangerous as attacks on classifiers.

Evaluating generative models is hard [10], there are no clear-cut success criteria for autoencoder reconstruction, and therefore, neither for the attack. We bypass that difficulty by analyzing how inputs and outputs differ across varying compromises between distorting the input and approaching the target. Altough, autoencoders admit many variations: sparse [11], denoising [12], variational [13, 14], Wasserstein [15], symmetric [16], etc, we are particularly interested on Variational Autoencoders (VAEs) since they behave as both autoencoders and as generative models, which brought them the community's attention.

Following up on [7], we propose a scheme to attack different VAEs, as well as a quantitative evaluation framework for the attacks that bypass the need for a success criterion. We compare three kinds of autoencoders: simple variational autoencoders (with fully-connected layers), convolutional variational autoencoders, and DRAW — a recently proposed recurrent variational autoencoder with an attention mechanism [17]. We show that the latter is more resistant to the attacks, and that its recurrent and attention mechanism both contribute to the resistance. We run all — statistically validated — experiments in three datasets (MNIST, SVHN, and CelebA) and show that our quantitative assessment correlates well with a qualitative perception of the attacks.

Tabacof et al. [7] introduced attacks on autoencoders, showing that they are possible and much harder than attacks on classifiers. They proposed the graphs we call Distortion–Distortion plots here and evaluated attack success by visual inspection of those graphs. Right after, Kos et al. [8] followed up with a work that attacked both the latent representation and the output of VAE–GAN autoencoders.

We explore here two types of attacks on VAEs: 1) input attack where the optimization goal is to find the distortion which minimizes the ℓ_2 distance between the target and the VAE's reconstruction image; and 2) latent attack where the goal is to minimize the Kullback–Leibler divergence between the target's and the VAE's resulting latent variables. In both methods, a regularization term is added to the optimization goal in order to keep the distortion norm small.

However, there is no sharp criterion to define whether the attack succeeded. We address this shortcoming with the **AUDDC** (Area under Distortion–Distortion Curve). For a given original and target pair, we compute different results, with different approximation compromises. The Distortion–Distortion plots show, for each attempt, how much we distorted the original and how much we approached the target (both measured by ℓ_2). We add limiting lines to the plot: no distortion added (and original reconstruction) at the leftmost/gray and topmost/orange lines; the ℓ_2-distance between the target and the reconstruction of the target by the model at the bottommost/red line; the ℓ_2-distance between the original and target image. Those limits represent, respectively, the starting point, the intrinsic limitation of the model, and the maximum "sensible" distortion (which allows going from the original to the target directly). We normalize the graph so that the distance between those lines is 1. The AUDDC is the area under the curve given by the linear interpolation of the points. The closer this area is to 1, the more resistant the model was to the attack (and the less successful the attack was) (Figure 1).

We employed three datasets: MNIST [18], SVHN [19], and CelebA [20]. We evaluated four models: VAE with only fully-connected layers (VAE); VAE with (de)convolutional layers (CVAE); the recurrent autoencoder DRAW [17] without and with its attention mechanism and different number of recurrent steps. For each pair model–dataset, we run 20 attacks with different pairs of image–target. For the quantitative analysis, we averaged the AUDDC

Submitted to Latinx in AI Research Workshop, 2018, Montréal, Canada.

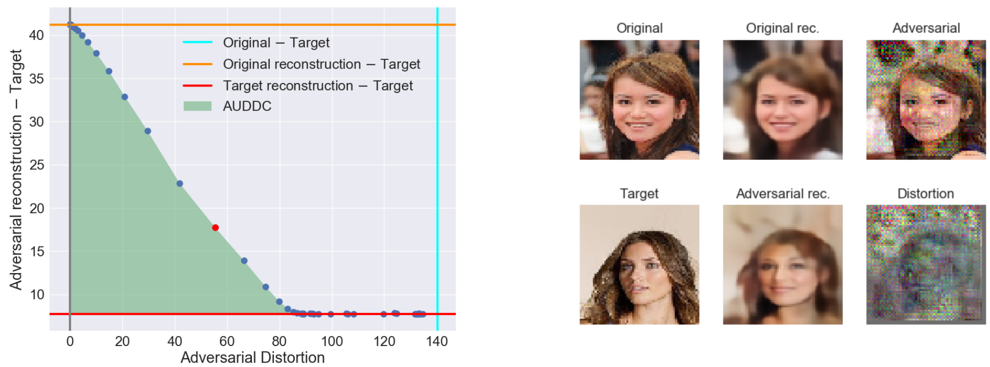

Figure 1: Left: the proposed metric: Area Under the Distortion–Distortion Curve (AUDDC). Right: visualization of a single point (red dot) of the left plot.

Table 1: Average ± 95%-confidence interval of AUDDC (times 100) for all models and datasets. Higher values indicate higher resistance to the attacks.

Steps	VAE —	CVAE —	DRAW* 1	DRAW 1	DRAW* 16	DRAW 16	Average
			Attacks on latent representation				
MNIST	27 ± 2	35 ± 3	27 ± 1	35 ± 3	71 ± 5	**91 ± 3**	47 ± 3
SVHN	19 ± 1	18 ± 1	09 ± 1	27 ± 2	74 ± 6	**96 ± 2**	41 ± 4
CelebA	31 ± 1	28 ± 1	21 ± 2	36 ± 1	81 ± 4	**97 ± 1**	49 ± 4
Average	25 ± 1	27 ± 2	19 ± 2	33 ± 1	75 ± 3	**95 ± 1**	46 ± 2
			Attacks on output				
MNIST	35 ± 2	56 ± 3	38 ± 2	48 ± 4	29 ± 3	**69 ± 4**	46 ± 2
SVHN	19 ± 1	19 ± 2	13 ± 1	27 ± 2	21 ± 2	**34 ± 2**	22 ± 1
CelebA	27 ± 1	24 ± 1	31 ± 3	35 ± 1	29 ± 2	**40 ± 1**	31 ± 1
Average	27 ± 1	33 ± 3	27 ± 2	37 ± 2	26 ± 1	**47 ± 3**	33 ± 1
			All attacks				
MNIST	31 ± 2	45 ± 3	32 ± 2	42 ± 3	50 ± 5	**80 ± 3**	47 ± 2
SVHN	19 ± 1	19 ± 1	11 ± 1	27 ± 1	47 ± 7	**65 ± 7**	31 ± 2
CelebA	29 ± 1	26 ± 1	26 ± 2	36 ± 1	55 ± 6	**68 ± 7**	40 ± 2
Average	26 ± 1	30 ± 2	23 ± 1	35 ± 1	51 ± 4	**71 ± 3**	39 ± 1

* Attention mechanism disabled.

for the chosen factors. To check which factors lead to significant influence, we used a multi-way ANOVA, with second-order interactions, and post-hoc Tukey honest significant differences which found significant differences (all p-values < 0.015) for all pairs of levels of all factors shown on the Table 1.

Attacking auto-encoders is relatively difficult if compared to attacking classifiers, where the distortions can be invisible to the human eye. Interestingly, DRAW, in particular, was much more resistant to our attacks. No attack succeed in reconstructing the target image well without incurring in immediately visible distortions to the input. Still, not all attempts are equal: some models are significantly more resistant than others. The AUDDC metric allows to quantify that resistance, bypassing the need to establish a clear-cut criterion of success for the attacks, and it correlates well with the qualitative results.

The code to reproduce all experiments will be made available after review.

References

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. arXiv:1312.6199.

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. arXiv:1412.6572.

[3] Pedro Tabacof and Eduardo Valle. Exploring the space of adversarial images. In *2016 International Joint Conference on Neural Networks, IJCNN 2016, Vancouver, BC, Canada, July 24-29, 2016*, pages 426–433, 2016.

[4] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations (ICLR) Workshop*, 2017. arXiv:1607.02533.

[5] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, 2016.

[6] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57, 2017.

[7] Pedro Tabacof, Julia Tavares, and Eduardo Valle. Adversarial images for variational autoencoders. *CoRR*, abs/1612.00155, 2016.

[8] Jernej Kos, Ian Fischer, and Dawn Song. Adversarial examples for generative models. *CoRR*, abs/1702.06832, 2017.

[9] Karol Gregor, Frederic Besse, Danilo Jimenez Rezende, Ivo Danihelka, and Daan Wierstra. Towards conceptual compression. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3549–3557, 2016.

[10] Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016. arXiv:1511.01844.

[11] Andrew Ng. Sparse autoencoder. *CS294A Lecture notes*, 72:1–19, 2011.

[12] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, 11:3371–3408, 2010.

[13] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. arXiv:1312.6114.

[14] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1278–1286, 2014.

[15] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

[16] Yuchen Pu, Weiyao Wang, Ricardo Henao, Liqun Chen, Zhe Gan, Chunyuan Li, and Lawrence Carin. Adversarial symmetric variational autoencoder. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4330–4339. Curran Associates, Inc., 2017.

[17] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. DRAW: A recurrent neural network for image generation. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1462–1471, 2015.

[18] Yann LeCun, Corinna Cortes, and Christopher JC Burges. The MNIST database of handwritten digits, 1998.

[19] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[20] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 3730–3738, Washington, DC, USA, 2015. IEEE Computer Society.