Adversarial Attacks on Variational Autoencoders

Contributions of latex 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 corresponding 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. Although, 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 $\ell_2$ 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 $\ell_2$). We add limiting lines to the plot: no distortion added (and original reconstruction) at the leftmost/gray and topmost/orange lines; the $\ell_2$-distance between the target and the reconstruction of the target by the model at the bottommost/red line; the $\ell_2$-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.
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.

<table>
<thead>
<tr>
<th>Steps</th>
<th>VAE</th>
<th>CVAE</th>
<th>DRAW*</th>
<th>DRAW</th>
<th>DRAW*</th>
<th>DRAW</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>27 ± 2</td>
<td>35 ± 3</td>
<td>27 ± 1</td>
<td>35 ± 3</td>
<td>71 ± 5</td>
<td><strong>91 ± 3</strong></td>
<td>47 ± 3</td>
</tr>
<tr>
<td>SVHN</td>
<td>19 ± 1</td>
<td>18 ± 1</td>
<td>09 ± 1</td>
<td>27 ± 2</td>
<td>74 ± 6</td>
<td><strong>96 ± 2</strong></td>
<td>41 ± 4</td>
</tr>
<tr>
<td>CelebA</td>
<td>31 ± 1</td>
<td>28 ± 1</td>
<td>21 ± 2</td>
<td>36 ± 1</td>
<td>81 ± 4</td>
<td><strong>97 ± 1</strong></td>
<td>49 ± 4</td>
</tr>
<tr>
<td>Average</td>
<td>25 ± 1</td>
<td>27 ± 2</td>
<td>19 ± 2</td>
<td>33 ± 1</td>
<td>75 ± 3</td>
<td><strong>95 ± 1</strong></td>
<td>46 ± 2</td>
</tr>
</tbody>
</table>

Attacks on latent representation

| 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  |

Attacks on output

| 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  |

All attacks

* 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.

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


