SEMANTICS PRESERVING ADVERSARIAL ATTACKS

Anonymous authors
Paper under double-blind review

Reviewer 1: Thanks for the useful and constructive feedback. Your insights have helped us greatly improve the quality of our paper. We have changed the manual evaluation process of our generated text as well as how we evaluated the adversarial text generated using Zhao et al. (2018b)'s approach. Before, we considered any attack to be successful therefore the targeted classifier assigns a label to a perturbed sample that is different from the label it assigns to the clean input. Although this definition holds for images, as the reviewer pointed out, however, it is weak for text. We have therefore removed the attack success rate we added for text. We have also updated the text examples we initially gave. Furthermore, we have carried out a new manual evaluation on samples that are adversarial both for our method and Zhao et al. (2018b).

Reviewer 2: Thanks for engaging with us. We have fixed the typos and put the algorithm in one page. We have also changed the terminology "adversarial success rate" to "attack success rate". We also understand your point of view. In light of the changes we brought to the manuscript, we hope you will favourably rate our paper.

Reviewer 3: We have updated the section "Manifold Learning via SVGD" a bit to highlight why a Bayesian framework was needed. In substance, we want to generate semantics preserving adversarial examples. In order to achieve this goal, we need to characterize as faithfully as possible the semantics of our inputs. That is, we need to learn the manifold that captures the semantics of the inputs. We use a VAE for manifold learning. Typically, VAEs learn an encoding function that maps the data manifold to an isotropic Gaussian, and a decoding function that transforms a latent code back to a sample in the input space. The reason for choosing the isotropic Gaussian is to render the ELBO tractable; hence easy to optimize. However, such a constraint imposed on the data manifold leads to learning poorly the semantics of the data, as reported by Jimenez Rezende & Mohamed (2015). Given that we want to generate semantics preserving adversarial examples, it is important that we characterize the semantics of the inputs well and with minimal assumptions about the distribution of the data. In that regard, similar to Pu et al. (2017), we propose to use SVGD. With SVGD, we don't need to impose a parametric/functional form for the distribution we want to learn. As SVGD is an MCMC method, we would, however, need to capture the uncertainty that results from the approximations. Kim et al. (2018) have shown that SVGD can be cast as a Bayesian approach for parameter estimation and uncertainty quantification. Hence, the motivation for using a Bayesian framework.

We are *not* using another recognition network for manifold alignment. To clarify, we have the encoder E' that perturbs the latent codes generated by the encoder E. Like E, E' is parameterized by M model instances. For large M, maintaining M models is computationally intensive. Thus, we introduced the recognition network f'_η to sample such models. The manifold alignment is a regularization technique that constrains E' to follow a distribution similar to E.

With respect to Equation 6, it is a wide practice to use empirical risk minimization to approximate distributions. While such an approximation can sometimes be problematic, we have shown that the distribution that E' follows is similar to the distribution of E. We showed in the Appendix in Fig 5 that these distributions overlap quite well for most of the datasets we considered. Also, using toy data, we showed that the manifolds of the benign examples and the adversarial ones overlap quite well. Finally, the quality of the adversarial images is quite high compared to the baselines we considered.

With respect to using a reconstruction loss, as a reminder, this loss is part-and-parcel of the training of VAEs. Although Song et al. (2018) and Zhao et al. (2018b) did not constrain the distance between generated adversarial and corresponding clean inputs, both use GANs which approximate the true data distribution. As a result, we would expect the adversarial examples they generate to be as natural, as realistically looking, as the clean inputs. However, this is not the case. Also, note that Song et al. (2018) use AC-GAN with gradient penalty (an L_2 loss) to constrain their discriminator to lie within 1-Lipschitz functions. We don't therefore think there is any competitive advantage that the reconstruction loss brings to us that the gradient penalty doesn't confer to them.

Regarding the use of L_∞ as reconstruction loss, now we attack the certified defences using this loss. We have run new experiments using this loss, and we have updated the paper accordingly.