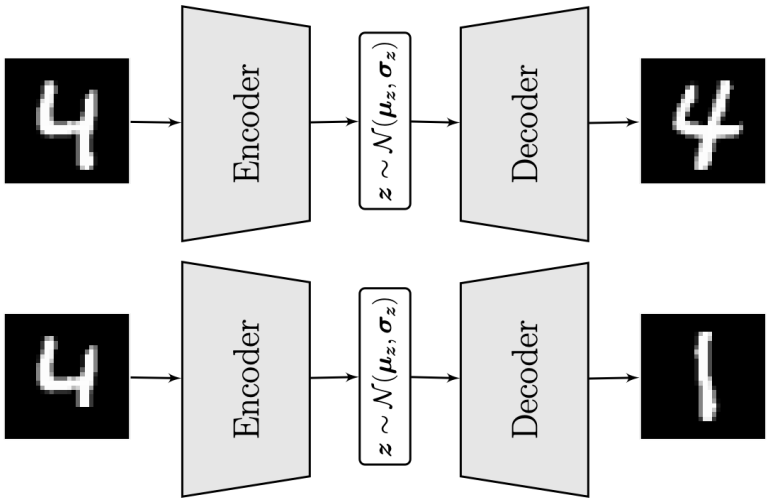
Robust Generative Modeling with Variational Autoencoders



Adversarial vulnerability of deep neural networks (DNNs) is a well-studied phenomenon, where minimally perturbed input examples can unreasonably alter the usual inference process of the network. Similarly for (conditional) variational autoencoders (VAE) -as illustrated above- an encoder can be *adversarially* fooled by presenting a crafted, minimally disturbed input image which still preserves the meaningful content (i.e., the image still looks like a 4 even though the tail is shorter), but would result in the encoded representation to now lie on a mode of the learned generative distribution which would result in a different reconstruction. In this project we want explore this phenomenon on larger scale datasets and develop novel approaches towards addressing this problem. Particular applications will include robustness against adversarially crafted examples or robustness in terms of generalization to learning from biased datasets.

Goals & Tasks

- Review of the state-of-the-art on adversarial robustness of VAEs.
- Implementation of deep VAEs for generative image modeling.
- Exploring and extending defenses against adversarial attacks on VAEs.

Qualifications

- Interest in deep learning and generative modeling with DNNs.
- Experience with Python based deep learning frameworks such as TensorFlow or PyTorch are beneficial.
- Registered to one of the following:
  - ✓ **Bachelor Thesis**
  - ✓ **Seminar Project**
  - ✓ **Master Thesis**

Contact

Ozan Özdenizci
ozan.ozdenizci@igi.tugraz.at