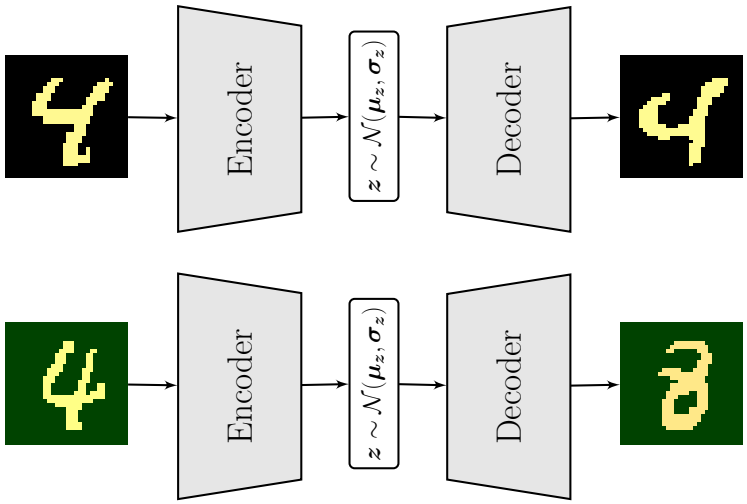
Adversarially Robust Generative Modeling with Deep Variational Autoencoders



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 carefully crafted, minimally disturbed input image which still preserves the meaningful content (i.e., the image is still a "4" with a different background color). In doing so, the encoded representation can be forced to lie on a critical region of the learned (conditional) generative distribution, which in turn may result in an unexpected decoder reconstruction. In this project we will explore different aspects of this phenomenon and investigate novel defense methods that aim to address this problem.

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