Generative Models and Fooling Neural Nets

CMPUT328 (Visual Recognition)

Nilanjan Ray

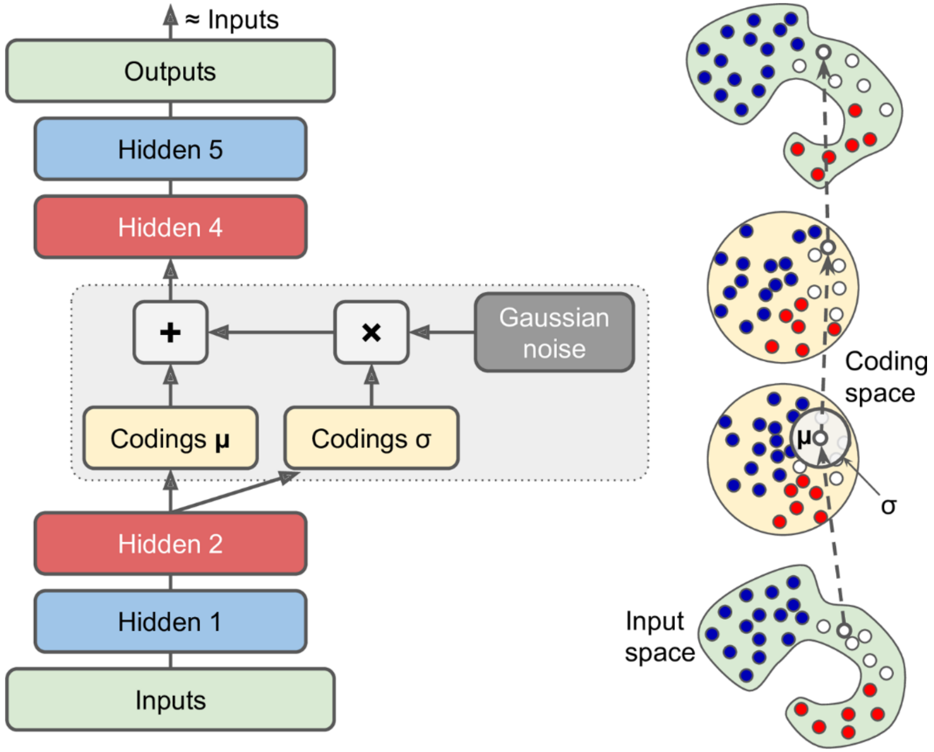
Agenda

- We will study a special type of autoencoder called “Variational Autoencoder” that are able to generate images.

- We will study Generative Adversarial Nets (GANs), which forms the state-of-the-art generative models for images today!

- We will also look at how to fool a neural network – called adversarial attacks.

Variational AE



Cost function has two components:

Reconstruction + constraint for μ and σ

The constraint: the encoded distribution should look like a zero-mean, unit variance Gaussian.

The advantage is that you can generate data (images) that look like the training images.

https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf

Generative Adversarial Nets: Resources

- http://cs231n.stanford.edu/

- https://lilianweng.github.io/lil-log/2017/08/20/from-GAN-to-WGAN.html

- https://pytorch.org/tutorials/beginner/dcgan_faces_tutorial.html

- https://www.kaggle.com/jessicali9530/celeba-dataset

Fooling LeNet

- Simple tricks can easily fool a neural net!
- We will modify some pixels in a digit image (say 7) to make LeNet to think it is another digit (say 3).
- We will refine our solutions.
- Finally, we will have an image with random numbers (close to it) to fool LeNet think that it is a digit with very high probability.