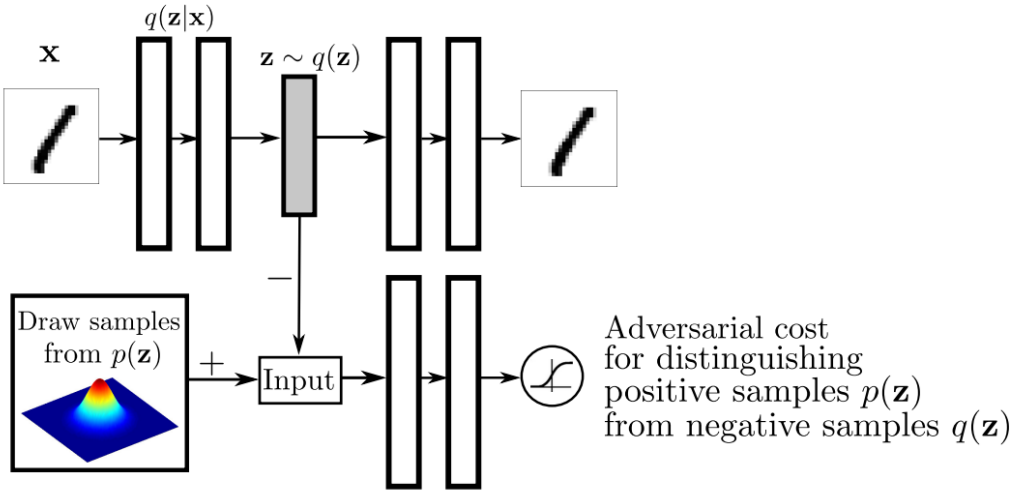
Basics Adversarial Auto Encoder

April 19, 2022

Useful papers - [MSJ+15]

1 Introduction

The architecture of the adversarial auto encoder is presented in the figure 1. This model makes



classical auto encoder a generative model. Encoder implicitly models conditional latent distribution $q(z|x)$. The implicitness is that encoder generates only samples, while in VAE encoder models whole latent conditonal distribution. The most advantages compared to VAE model is that AAE is able to model arbitrary a priori distribution due to lack of KL-divergence. AAE uses following loss functions

$$\mathcal{L}_{decoder}(\theta) = \sum_{i=1}^{n} \left[\int q(z|x_i, \phi) \log p(x_i|z, \theta) dz \right] \rightarrow max_\theta \tag{1}$$

$$\mathcal{L}_{discr}(\eta) = \sum_{i=1}^{n} \left[\int p(z) \log D_\eta(z) dz + \int q(z|x_i, \phi) \log(1 - D_\eta(z)) dz \right] \rightarrow max_\eta \tag{2}$$

$$\mathcal{L}_{encoder}(\phi) = \sum_{i=1}^{n} \left[\int q(z|x_i, \phi) \log p(x_i|z, \theta) dz + \lambda \int q(z|x_i, \phi) \log(D_\eta(z)) dz \right] \rightarrow max_\phi \tag{3}$$

For optimization phase stochastic gradient will be used. Also Monte Carlo estimation will be used. Important to note that

$$\nabla_\theta \mathcal{L}_{decoder}(\theta) \approx \frac{n}{m} \sum_{j=1}^{m} \nabla_\theta \log p(x_j|z_j^*, \theta), \;\; z_j^* \sim q(z_j|x_j, \phi) \tag{4}$$

$$\nabla_\eta \mathcal{L}_{discr}(\eta) \approx \frac{1}{m} \sum_{k=1}^{m} \nabla_\eta \log D_\eta(z_k^*) + \frac{n}{m} \sum_{j=1}^{m} \nabla_\eta \log D_\eta(1 - z_j^*), \;\; z_j^* \sim q(z_j|x_j, \phi); \;\; z_k^* \sim p(z_k) \tag{5}$$

$$\nabla_\phi \mathcal{L}_{encoder}(\phi) \approx \frac{n}{m}\sum_{j=1}^{m} \nabla_\phi \log p(x_j|z_j^*,\theta) + \lambda \frac{n}{m}\sum_{j=1}^{m} \nabla_\phi \log D_\eta(z_j^*),\ z_j^* \sim q(z_j|x_j,\phi) \tag{6}$$

In this model encoder is a generator. Adversarial Auto Encoder trains in two phases:

1. **Reconstruction phase**, updating decoder and encoder to minimize reconstruction loss

$$(\phi^{'},\theta^{'}) = (\phi,\theta) + \frac{1}{m}\sum_{j=1}^{m} \nabla_{\phi,\theta} \log p(x_j|z_j^*,\theta) \tag{7}$$

2. **Regularization phase**, updating discriminator to tell apart true and fake samples and encoder to confuse discriminator

$$\eta^{'} = \eta + \frac{1}{m}\sum_{k=1}^{m} \nabla_\eta \log D_\eta(z_k^*) + \frac{1}{m}\sum_{j=1}^{m} \nabla_\eta \log D_\eta(1-z_j^*),\ z_j^* \sim q(z_j|x_j,\phi);\ z_k^* \sim p(z_k) \tag{8}$$

$$\phi^{'} = \phi + \frac{1}{m}\sum_{j=1}^{m} \nabla_\phi \log D_\eta(z_j^*),\ z_j^* \sim q(z_j|x_j,\phi) \tag{9}$$

References

[MSJ+15] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.