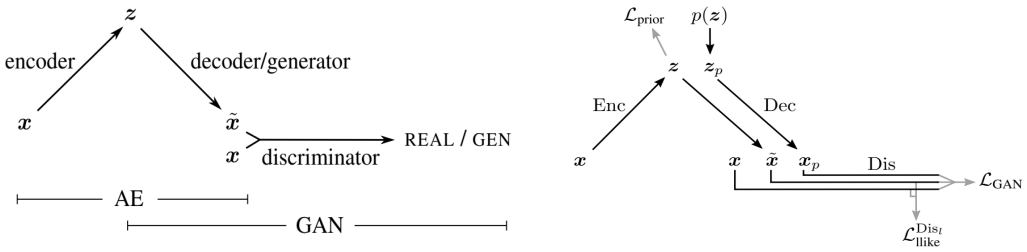
Basics of VAE GAN

May 17, 2022

Useful papers - [LSLW16]

1 Introduction

The architecture of the VAE GAN is presented in the figure 1. The main goal of



this model is to improve standard variational auto encoder by adding discriminative learning. In classical VAE variational lower bound has the following form

$$\begin{aligned}\mathcal{L}(\theta,\phi) = \text{Reconstruction Term} - \text{KL divergence} = \\ = \mathbb{E}_{q(z|x,\phi)} \log p(x|z,\theta) - \mathbf{KL}(q(z|x,\phi)||p(z))\end{aligned} \tag{1}$$

We try to maximize ELBO, i.e., maximize reconstruction term (or minimize reconstruction loss) and minimize Kulback-Leibler divergence. KL divergence is an regularization term, it in charge of making posterior latent distribution more similar to prior distribution. Thanks to it, we can further generate new samples from prior distribution, i.e., KL divergence makes AE as generative model. Reconstruction term is responsible for minimization of "difference" between real and generated objects. This term depends on the choosing of posterior object distribution $p(x|z,\theta)$, usually it Normal $p(x|z,\theta) = \mathcal{N}(x|\mu_\theta(z), \sigma_\theta(z))$ or Bernoulli $p(x|z,\theta) = f_\theta(z)^x \cdot (1 - f_\theta(z))^{1-x}$ (assuming independence $p(x|z,\theta) = \prod_{i=1}^n f_\theta(z_i)^{x_i} \cdot (1 - f_\theta(z_i))^{1-x_i}$) distributions. Interesting remark: minus logarithm of Bernoulli distribution correspond to the binary cross entropy loss. However, it possible to ask (at least, i have such question): How well the chosen distribution will show the "difference" between real and generated objects. We know that mse and bce it is an point wise metrics. Thus, we can generate objects with real high quality but displaced compared to the real objects and it will cause large losses (which is definitely not true). The main idea behind VAE GAN model lies in the idea to entrust the estimation of "difference" between real and generated objects

to neural network (i.e. to discriminator network). In other words, we want to replace $\mathbb{E}_{q(z|x,\phi)} \log p(x|z,\theta)$ to neural network output. It possible to think that neural network will learn non trivial features to estimate the difference. So, such neural network metric might be better than standard approaches like mse or bce. In such a way it possible to combine latent representation of initial object (what encoder do in VAE), generator (or decoder) and discriminative network to better evaluate difference between real and fake objects (what discriminator do in GAN).

Besides replacing reconstruction term to discriminative network, authors of paper do a lot of interesting features. They do not drop completely reconstruction term, but modify it in a following way:

$$\mathbb{E}_{q(z|x,\phi)} \log p(x|z,\theta) \to \mathbb{E}_{q(z|x,\phi)} \log p(D_l(x)|z,\theta)) = \mathcal{L}_{rec} \tag{2}$$

where $D_l(x)$ l-th output of discriminative network. Idea behind this is that discriminator should give out same features for similar objects. In other words, we want to push generator and encoder to create objects with same features as real objects. Final loss has the following form:

$$\begin{gathered}
\mathcal{L} = -\mathcal{L}_{rec} + \mathcal{L}_{prior} + \mathcal{L}_{GAN} \\
\mathcal{L}_{rec} = \mathbb{E}_{q(z|x,\phi)} \log p(D_l(x)|z,\theta) \\
\mathcal{L}_{prior} = \mathbf{KL}(q(z|x,\phi)||p(z)) \\
\mathcal{L}_{GAN} = \mathbb{E}_{p_{data}(x)} \log D_l(x) - \mathbb{E}_{q(z|x,\phi)} \log D_l(G(z)) - \mathbb{E}_{p(z)} \log D_l(G(z)) \\
p(z) = \mathcal{N}(0, I),\ p(D_l(x)|z,\theta) = \mathcal{N}(D_l(x)|\mu_\theta(z), \sigma^2_\theta(z))
\end{gathered} \tag{3}$$

Learning algorithm looks like

$$\begin{aligned}
d^{k+1} &= d^k - \nabla_d(-\mathcal{L}_{GAN}) \\
\phi^{k+1} &= \phi^k - \nabla_\phi(\mathcal{L}_{prior} - \mathcal{L}_{rec}) \\
\theta^{k+1} &= \theta^k - \nabla_\theta(\mathcal{L}_{GAN} - \mathcal{L}_{rec})
\end{aligned} \tag{4}$$

$\theta, \phi, d -$ parameters of generator, encoder and discriminator, respectively

References

[LSLW16] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pages 1558–1566. PMLR, 2016.