ML2 Assignment 1

Pradip Das (CS2115)

September 27, 2022

1. (a) Denoising Autoencoder:

   Denoising autoencoder is a modification on the standard autoencoders that reduces the risk of learning the identity function. If the autoencoder is too big, then it can just learn the data, so the output equals the input, and does not perform any useful representation learning or dimensionality reduction. Denoising autoencoders solve this problem.

   (b) Pseudocode for Denoising Autoencoder training :

   $\mathbf{x} = [x_1, x_2, ...x_n] \in R^{n \times m}$ is the input matrix, in which $x_i \in [0,1]^m$ $(1 \leq i \leq m)$ is a single input data

   e is the amount of epochs to be iterated

   b is the amount of batches

   l is the learning rate

   c is the corruption level

   $\theta = \{W, \mathbf{b}, \mathbf{b_n}\}$ where $W \in R^{n \times d}$, $\mathbf{b} \in R^d$, $\mathbf{b_n} \in R^d$, θ is the parameters of a Denoising Autoencoders

for *0 to e* **do**
  for *0 to b* **do**
    $\tilde{\mathbf{x}}$ = getCorruptedInput($\mathbf{x}$), in which c is the corrupted level
    $\mathbf{h}$ = sigmoid($\tilde{\mathbf{x}} * W + \mathbf{b}$)
    $\hat{\mathbf{x}}$ = sigmoid($\mathbf{h} * W^T + \mathbf{b_n}$)
    $\mathbf{L}(x, \hat{\mathbf{x}}) = -\sum_{i=0}^{d} [\mathbf{x}_i log \hat{\mathbf{x}}_i + (1 - \mathbf{x}_i) log(1 - \mathbf{x}_i)]$
    cost = mean($\mathbf{L}(x, \hat{\mathbf{x}})$)
    $\mathbf{g}$ = compute the gradients of the cost with respect to θ
    for θ_i, g_i *in* $(\theta, \boldsymbol{g})$ **do**
      $\theta_i = \theta_i - l * g_i$
    end
  end
end

2. (a) Yes, the use of BCE loss lead to the unstable training of GAN by being saturated.
   Yes, LSGANs is the improved stability of learning process.

   (b) Objective function of LSGAN:

   $\underline{1^{st} part}$

   Let the GAN consisting of Generator **G** and a discriminator **D**. Let us take the original data

distribution as p_{data}, that generated by $\mathbf{G}$ as p_g and that of the noise as p_z. Then the LS loss objective function is given by,

$$min_D = \frac{1}{2}E_{x\sim p_{data}}(\mathbf{D}(x) - b)^2 + \frac{1}{2}E_{z\sim p_z}(\mathbf{D}(\mathbf{G}(z)) - a)^2 \tag{1}$$

$$min_G = \frac{1}{2}E_{x\sim p_z}(\mathbf{D}(\mathbf{G}(z)) - c)^2 \tag{2}$$

where a, b and c respectively denotes the generated data label, the real data label and the label of the data which the generator wants the discriminator to believe.

$\underline{2^{nd} part}$

Let,

$$\begin{aligned} f(\mathbf{D}) &= \tfrac{1}{2}E_{x\sim p_{data}}(\mathbf{D}(x) - b)^2 + \tfrac{1}{2}E_{x\sim p_G}(\mathbf{D}(x) - a)^2 \\ &= \tfrac{1}{2}\int_x [p_{data}(x)(\mathbf{D}(x) - b)^2 + p_z(x)(\mathbf{D}(x) - a)^2]dx \end{aligned} \tag{3}$$

Now maximizing $f(\mathbf{D})$ is equivalent to maximizing $p_{data}(\mathbf{D}(x) - b)^2 + p_g(\mathbf{D}(x) - a)^2$. So let,

$$L(\mathbf{D}) = p_{data}(x)(\mathbf{D}(x) - b)^2 + p_g(x)(\mathbf{D}(x) - a)^2 \tag{4}$$

Now,

$\frac{\partial L}{\partial D} = 2p_{data}(x)(\mathbf{D}(x) - b) + 2p_g(x)(\mathbf{D}(x) - a) = 0$

$$\implies \boxed{\mathbf{D}^*(x) = \frac{b\ p_{data}(x)\ +\ a\ p_g(x)}{p_{data}(x)\ +\ p_g(x)}}$$

Then put the value of $\mathbf{D}^*(x)$ in $\frac{1}{2}E_{x\sim p_{data}}(\mathbf{D}(x) - c)^2 + \frac{1}{2}E_{x\sim p_g}(\mathbf{D}(x) - c)^2$ we have,

$$\begin{aligned} 2C &= \int_x [p_{data}(x)(\tfrac{b\ p_{data}(x)\ +\ a\ p_g(x)}{p_{data}(x)\ +\ p_g(x)} - c)^2 + p_g(x)(\tfrac{b\ p_{data}(x)\ +\ a\ p_g(x)}{p_{data}(x)\ +\ p_g(x)} - c)^2]dx \\ &= \int_x \tfrac{((b-c)\ (p_{data}(x)+p_g(x))-(b-a)p_g(x))^2}{p_{data}(x)\ +\ p_g(x)}dx \end{aligned}$$

If we take $b - c = 1$ and $b - a = 2$, then

$$\boxed{2C = \chi^2_{Pearson}(p_{data} + p_g||2p_g)}$$

3. $\underline{1^{st} part}$

Inception Score: The Inception Score is based on a heuristic that realistic samples should be able to be classified when passed through a pre-trained network, such as Inception on ImageNet. Besides high predictability (low entropy), the Inception Score also evaluates a GAN based on how diverse the generated samples are (e.g. high variance or entropy over the distribution of generated samples). This means that there should not be any dominating classes.

$$IS = \exp\left(E_{x\sim p_g} D_{KL}(p(y|x)||p(y))\right) \tag{5}$$

where p_g be the distribution of generated data.

Fréchet Inception Distance: FID estimates realism by measuring the distance between the generated distribution of images and the true distribution. FID embeds a set of generated samples into a feature space given by a specific layer of Inception Net. This embedding layer is viewed as as a continuous multivariate Gaussian, then the mean and covariance are estimated for both the generated data and the real data. The Fréchet distance between these two Gaussians is then used to quantify the quality of generated samples. A lower FID corresponds to more similar real and generated samples.

$$FID(\mathbb{X}_r, \mathbb{X}_g) \;=\; ||\mu_r - \mu_g|| + Tr(\Sigma_r + \Sigma_g + 2(\Sigma_r^{\frac{1}{2}} \cdot \Sigma_g \cdot \Sigma_r^{\frac{1}{2}})^{\frac{1}{2}}) \tag{6}$$

where $\mathbb{X}_r \sim \mathcal{N}(\mu_r, \Sigma_r)$ and $\mathbb{X}_g \sim \mathcal{N}(\mu_g, \Sigma_g)$.

$\underline{2^{nd} part}$

Disadvantages include that optimizing IS was found to lead to adversarial examples, it is sensitive to small network weight changes, and does not constitute a proper distance. FID was found to be more reliable than IS, while pertaining to a number of disadvantages. FID quantifies performance in terms of affinity of the data and model distributions. The two distributions are estimated by fitting Gaussian distributions on the respective Inception feature embeddings of the data. Subsequently, a Frechet distance measures the divergence between the two distributions, and a lower score means that the synthetic data are closer to the true, original data.