IE643: Deep Learning - Theory and Practice **July-Dec 2024**

Quiz 2: (19^{th} November 2024) Time: 70 minutes

Roll No: *Name:*

1 Instructions

Answer all questions. You are free to refer to your hand-written (and/or paper-printed) class notes and scribes that you carry wi you. However you are not allowed to access books, e-books, electronic devices and internet. No discussion with your classmates i: allowed. Write your answers clearly and state the assumptions you have used. No clarifications will be provided. You can score a maximum of 45 marks in this quiz.

2 Questions

1. Consider the space $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ (where $d_x \in \mathbb{Z}_{++}$, the set of positive integers) from where the data samples arise. Let $\mathbb{P}_r$ denote the true data distribution over $\mathcal{X}$. Let $\mathcal{Z} \subseteq \mathbb{R}^{d_z}$ (where $d_z \in \mathbb{Z}_{++}$) denote a latent space and let $\mathbb{P}_z$ denote the prior distribution over $\mathcal{Z}$. Let $D_\phi : \mathbb{R}^{d_x} \to [0,1]$ denote a discriminator parametrized by ϕ.. Let $G_\theta : \mathcal{Z} \to \mathcal{X}$ denote a generatoı parametrized by θ. Let $\mathbb{P}_\theta$ denote the induced generated distribution from samples of the form $x = G_\theta(z), z \sim \mathbb{P}_z$. Conside a GAN with the discriminator D_ϕ and generator G_θ. Note that the GAN objective can be written as $\min_\theta \max_\phi L(\theta, \phi) = \mathbb{E}_{x\sim\mathbb{P}_r} \log(D(x)) + \mathbb{E}_{x\sim\mathbb{P}_\theta} \log(1 - D(x))$. Let $P_r(x)$ and $P_\theta(x)$ denote the density functions corresponding to $\mathbb{P}_r$ and $\mathbb{P}_\theta$ resp tively. The following questions are with respect to this GAN.

   (a) **[4 marks]** Consider the situation given in Figure 1 obtained by training the GAN described above. Illustrate your understanding of this figure. Explain your observations about the associated trained GAN.

   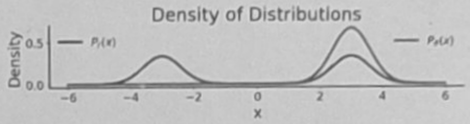

   Figure 1: An example

   (b) **[3 marks]** Consider a region $\Omega \subset \mathcal{X}$ such that the samples from Ω are covered by $\mathbb{P}_r$, but not covered by the distribution $\mathbb{P}_\theta$. Using the densities P_r and P_θ, explain a method which can be used to test if x is sampled from Ω or not. Justify your method with proper explanation.

   (c) **[3 marks]** Suppose the optimal discriminator D_{ϕ^*} is provided as $D_{\phi^*}(x) = \frac{P_r(x)}{P_r(x)+P(\theta(x))}$. Using D_{ϕ^*}, discuss a methoc which can be used to test if x is sampled from Ω (described in part 1b) or not. Justify your method with proper expla nation.

   (d) **[4 marks]** Suppose that a train set T was used to train the GAN. Based on the test condition in part 1c, can you pro vide a quantitative metric to quantify how well the GAN can generate samples in T? Include proper justifications.

2. Consider a VAE where the encoder is a linear mapping $W \in \mathbb{R}^{d_z \times d_x}$ transforming $x \in \mathbb{R}^{d_x}$ to latent space as $z \in \mathbb{R}^{d_z}$ (assuı d_x, d_z to be positive integers). Based on the reparametrization trick, the latent z is produced by further adding noise term $\zeta \in \mathcal{N}(0, \Sigma)$, using $z = Wx + \zeta$, with $\Sigma \in \mathbb{R}^{d_z \times d_z}$ being the encoder variance. Hence the recognition model can be given by $q_\phi(z|x) = \mathcal{N}(Wx, \Sigma)$. Let the decoder be a linear map that parametrizes the distribution $p_\theta(x|z) = \mathcal{N}(Uz, \eta_d^2 I)$ with $U \in \mathbb{R}^{d_x \times d_z}$ and $\eta_d > 0$. Let the prior $p(z)$ be $\mathcal{N}(0, \eta_e^2 I)$ with $\eta_e > 0$. Suppose that for two d dimensional multi-variate Gaussians $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$, the KL divergence is given by:

$$\frac{1}{2}\left[\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} + \text{Trace}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^\top \Sigma_2^{-1}(\mu_2 - \mu_1) - d\right]$$

   where $\det(A)$ and $\text{Trace}(A)$ denote respectively the determinant and trace of matrix A. Answer the following.

   (a) **[5 marks]** Using the above setting, derive the expression of the ELBO for VAE with proper details.

   (b) **[5 marks]** Describe with all relevant details, a training procedure for the VAE described above, with a suitable loss function.

3. Consider a simple U-Net based auto-encoder which is to be trained for reconstruction of videos. Consider the data set wher each video $X \in \mathbb{R}^{f\times h\times w\times c}$, where f denotes the number of frames and each frame is of height h, width w and contains c channels. Fix $f = 10$, $h = w = 32$ and $c = 3$. Answer the following.

   (a) **[3 marks]** In the encoder block of the U-Net, describe how the first convolution operation is performed with a suitabl kernel directly on the video. Describe the dimensions of the associated kernel and convolved feature map.

(b) **[3 marks]** Fix a suitable number $N > 2$ of encoder blocks in the U-Net. Recall that each encoder block contains a ser of M convolution operations followed by a pooling operation. Assume specific dimensions of kernels to be used at each stage of the encoder and describe the corresponding intermediate convolved feature map shapes and the pooled featur map dimensions at each stage of the encoder.

(c) **[5 marks]** Describe with all relevant details, how an upconvolution is performed in a decoder block of the U-Net.

(d) **[3 marks]** Describe the complete structure of the decoder, give details about the dimensions of the intermediate featu maps and describe the dimensions associated with skip connections at every stage of the decoder.

(e) **[2 marks]** Describe a suitable loss function to train the auto-encoder.

4. **[5 marks] [open-ended]** Consider training vision transformer (VIT) on high resolution images (e.g. of size 4096 × 4096 × If a small patch size, say 8×8×3 is used, you might notice that the self-attention matrix size increases tremendously. Howe using a larger patch size, say 128 × 128 × 3 might not be able to capture local context effectively. Instead of just choosing different patch sizes, you will aim to describe a modified VIT architecture, by justifying how your modified architecture can effectively help in training with high resolution images. Precisely indicate how much efficiency in terms of computations, yo modified architecture attains when compared with the usual VIT architecture designs.

END OF QUESTIONS