THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY

Machine Learning

Homework 3

Due Date: See course website

Submission is to be made via Canvas by 11:00pm on the due date.

Question 1: The input of a convolutional layer has shape $27 \times 27 \times 256$ (width, height, depth). The layer uses 384 3×3 filters applied at stride 1 with no zero padding. What is the shape of the output of the layer? How many parameters are there? How many float multiplication operations it will take to compute the net inputs of the all the output units?

Question 2: What is batch normalization? What is layer normalization? What are their pros and cons?

Question 3 In an LSMT cell, $\mathbf{h}^{(t)}$ is computed from $\mathbf{h}^{(t-1)}$ and $\mathbf{x}^{(t)}$ using the following formulae:

$$\begin{aligned}
\mathbf{f}_t &= \sigma(\mathbf{W}_f\mathbf{x}^{(t)} + \mathbf{U}_f\mathbf{h}^{(t-1)} + \mathbf{b}_f) \\
\mathbf{i}_t &= \sigma(\mathbf{W}_i\mathbf{x}^{(t)} + \mathbf{U}_i\mathbf{h}^{(t-1)} + \mathbf{b}_i) \\
\mathbf{o}_t &= \sigma(\mathbf{W}_q\mathbf{x}^{(t)} + \mathbf{U}_q\mathbf{h}^{(t-1)} + \mathbf{b}_q) \\
\mathbf{c}_t &= \mathbf{f}_t \otimes \mathbf{c}_{t-1} + \mathbf{i}_t \otimes \tanh(\mathbf{U}\mathbf{x}^{(t)} + \mathbf{W}\mathbf{h}^{(t-1)} + \mathbf{b}) \\
\mathbf{h}^{(t)} &= \mathbf{o}_t \otimes \tanh(\mathbf{c}_t)
\end{aligned}$$

(a) Intuitively, what are the functions of the forget gate $\mathbf{f}_t$ and the input gate $\mathbf{i}_t$ do? Answer briefly.

(b) Why do we use the sigmoid function for $\mathbf{f}_t$ and $\mathbf{i}_t$, but tanh for the memory cell $\mathbf{c}_t$ and the output $\mathbf{h}^{(t)}$? Answer briefly.

Question 4 BERT input representation is the sum of token embedding, segment embedding, and positional embedding. Briefly explain why positional embedding is introduced in the architecture.

Question 5 (optional): Let $p(\mathbf{x}, \mathbf{z})$ and $q(\mathbf{z})$ be two probability distributions. Show that

$$\log p(\mathbf{x}) \geq E_{\mathbf{z}\sim q(\mathbf{z})} \log p(\mathbf{x}|\mathbf{z}) - \mathcal{D}_{KL}(q(\mathbf{z})||p(\mathbf{z}))$$

where $\mathcal{D}_{KL}(q(\mathbf{z})|p(\mathbf{z})) = E_{\mathbf{z}\sim q(\mathbf{z})} \log q(\mathbf{z}) - E_{\mathbf{z}\sim q(\mathbf{z})} \log p(\mathbf{z})$.

Note that the RHS of the inequality is known as the **variational lower bound** of $\log p(\mathbf{x})$, or the **evidence lower bound (ELBO)**. Another way to write the inequality is as follows:

$$\log p(\mathbf{x}) \geq E_{\mathbf{z}\sim q(\mathbf{z})} \log p(\mathbf{x}, \mathbf{z}) - E_{\mathbf{z}\sim q(\mathbf{z})} \log q(\mathbf{z}) = E_{\mathbf{z}\sim q(\mathbf{z})} \log p(\mathbf{x}, \mathbf{z}) + H(q).$$

Reflect on how the variational lower bound is used variational autoencoder. You can do this by pointing out what are the two distributions.

Question 6: Suppose two random variables x and z are related by the following equation:

$$z = x^2 + 2x + \frac{x^2}{2}\epsilon,$$

where ϵ is a random variable that follows the normal distribution $\mathcal{N}(0, 1)$, i.e.,

$$p(\epsilon) = \frac{1}{\sqrt{2\pi}} e^{\frac{-\epsilon^2}{2}}.$$

What is the density function of the conditional distribution $p(z|x)$?

(Note that the purpose of this problem is to help you understand reparameterization.)

Question 7: What are the main differences between variational autoencoder (VAE) and generative adversarial network (GAN) in terms their functionalities and the ways they operate?

Question 8: What are the objective functions for the discriminator and the generator in GAN? What are the objective functions for the critic and generator in WGAN?