THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY

Machine Learning

Homework 3

Due Date: See course website

Your answers should be typed, not handwritten. You can submit a Word file or a pdf file. Submissions are to be made via Canvas. Note that penalty applies if your similarity score exceeds 40. To minimize your similarity score, don't copy the questions.

Question 1: Consider the image X and filter F given below. Let X be convolved with F using no padding and a stride of 1 to produce an output Y. Assume the bias is 2 and the activation function is ReLU. What are values of the cells a, b, e and f in the output Y?

$X =$

1	0	-2	3	4	1
2	9	5	6	0	-1
0	-3	1	3	4	4
6	5	2	0	6	8
-5	4	-3	1	3	-2
4	1	2	8	9	7

$F =$

-1	-1	-1
-1	8	-1
-1	-1	-1

$Y =$

a	b	c	d
e	f	g	h
i	j	k	l
m	n	o	p

Question 2: 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 3: What is batch normalization? What is layer normalization? What are their pros and cons?

Question 4 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 5 Consider a self-attention layer with the following input token embeddings:

x_1: 0.0, 1.0
x_2: 1.0, 0.0
x_3: 0.5, 0.5
x_4: 0.8, 0.2
x_5: 0.2, 0.8

Let the key matrix W^K and the value matrix W^V both be the identity matrix.

(a) Suppose the query matrix is $W_1^Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. What are the attention weights when calculating the output taken embeddings $z_1, \ldots, z_5$ of the self-attention layer? What are the output token embeddings.

(b) Suppose the query matrix is $W_2^Q = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. What are the attention weights when calculating the output taken embeddings $z_1, \ldots, z_5$ of the self-attention layer? What are the output token embeddings.

(c) In Part (a), what would be the attention weights if the input taken embeddings were:

x_1: 0.8, 0.2
x_2: 0.2, 0.8
x_3: 0.0, 1.0
x_4: 1.0, 0.0
x_5: 0.5, 0.5

Note how the attention matrix influences the output (a vs b), and how the attention weights changes with respect to input (a vs c).

Question 6 BERT input representation is the sum of token embedding, segment embedding, and positional embedding. Briefly explain why positional embedding is introduced in the architecture.

[More questions will be added later.]