Indian Institute of Technology Kharagpur
Class Test 03 2021-22

Date of Examination: 07 Apr. 2022 **Duration:** 1 Hour
Subject No.: CS60010 **Subject:** Deep Learning
Department/Center/School: Computer Science **Credits:** 3 **Full marks:** 20

Instructions

i. This question paper contains 3 pages and 4 questions. All questions are compulsory. Marks are indicated in parentheses. This question paper has been cross checked.
ii. Please write your name, roll number, subject name and code, date and time of examination on the answer script before attempting any solution.
iii. **Organize your work**, in a reasonably neat and coherent way. Work scattered all across the answer script without a clear ordering will receive very little marks.
iv. **Mysterious or unsupported answers will not receive full marks**. A correct answer, unsupported by calculations, explanation, will receive no marks; an incorrect answer supported by substantially correct calculations and explanations may receive partial marks.
v. In the online mode of the quiz, you need to upload your answer scripts as **pdf file**. You can scan your worked out example or you can use latex to produce the pdf.

1. (a) (2 points) Typically, we use the softmax function to model the probability of one of several classes given the input. Instead, consider a binary classification problem and softmax function applied to it with output labels $j \in \{0, 1\}$ to model the probability of the binary label $y = j$ given the input x:

$$P(y = j|x) = \frac{e^{w_j^T x}}{\sum_{j'=0,1} e^{w_{j'}^T x}}$$

Where w_j is the parameter vector of the j^{th} class. Show that the above problem can be expressed as a logistic regression classifier with parameter some parameter w and give the expression for w.

(b) (1.5 points) We have seen the evolution of RCNN architectures from RCNN to Fast-RCNN to Faster-RCNN. The evolution happened for region proposal mechanisms as well classifier used to classify the proposed regions. Complete the cells of the following table indicating what type of region proposal mechanism and classifiers are used for each of the three RCNN architectures. Your options for region proposal are 'External' or 'Region Proposal Network (RPN)'. For classifier, the options are 'External' or 'CNN'.

Architecture	Region Proposal	Classifier
RCNN		
Fast RCNN		
Faster RCNN		

(c) (0.5 points) Suppose a neural network is lazy and just produces the same constant output whatever training data we give it. What can you say about its bias and variance?

(d) (0.5 points) A perceptron has two inputs $\mathbf{x}_1$ and $\mathbf{x}_2$ with weights $\mathbf{w}_1$ and $\mathbf{w}_2$, and a bias of b_0. The activation function of the perceptron is $h(\mathbf{x})$. Write down the expression of the output of the perceptron.

(e) (0.5 points) If the hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$ separates all the training points $(\mathbf{x}_i, y_i)$, where $y_i = \{+1, -1\}$, then which of the following is true.
i) $||\mathbf{w}|| = 0$, ii) $b = 0$, iii) $\mathbf{w}^T\mathbf{x}_i + b \geq 0$ for all i, iv) $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 0$ for all i

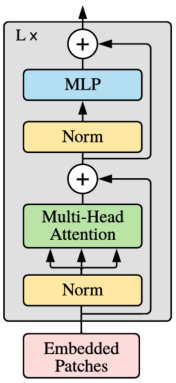
2. (a) (4 points) Lets consider an image with 4 cats where a detector provides 8 detections. The 8 detections are ranked according to the detection scores and are provided in the table below in decreasing order of the scores. The second column of the table provides if the corresponding detection is a true positive or a false positive. Fill in the precision and recall values in all the rows. Both values with precision upto two decimal places are required.

Rank	Correct	Precision	Recall
1	True Positive		
2	True Positive		
3	False Positive		
4	False Positive		
5	False Positive		
6	True Positive		
7	False Positive		
8	True Positive		

(b) (1 point) You are applying batch normalization to a fully connected (dense) layer with an input size of 10 and an output size of 20. How many training parameters does this layer have, including batch normalization parameters?

3. Consider an implementation of the encoder of a transformer as follows:

1. The input is a sequence of $n = 75$ words
2. The size of the word embedding is $d = 100$
3. The query and key dimensions are $dim_q = 36$ and $dim_k = 36$ respectively, while the value dimension $dim_v = 25$



(a) (2 points) Find out the number of multiplication operations for one forward computation of self-attention. Note that the input to self-attention layer is the combined embedding after positional encoding and output are the key, query and value vectors as well as weighted sum of the value vectors. You don't need to consider softmax operation. For this computation, assume a single head of self-attention layer.

(b) (1 point) What is the number of trainable parameters in the self-attention layer?

(c) (0.5 points) Now consider multi-headed self-attention with $k = 4$ heads. Assume that the different heads are exactly of the same architecture as described above for the single head. What are the number of parameters in the *'LayerNorm'* layer following this.

(d) (1.5 points) Now there is a fully connected feed-forward network (FFN) as an MLP. The FFN consists of two linear layers with a ReLU activation in between.

$$FFN(x) = W_2[max(0,\ W_1x + b_1)] + b_2$$

The dimensionality of the output after the first linear layer is $d_1 = 200$ and the dimensionality of the output after the second linear layer is 100.

1. Find out the total number of multiplication operations for these two linear layers.
2. What is the number of trainable parameters in FFN?
3. What is the total number of trainable parameters in one block of the encoder containing multi-headed self-attention, LayerNorm, FFN and again LayerNorm?

4. (a) (2 points) Suppose we have a cost function

$$J(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{\theta}^T\mathbf{x}^{(i)} + b\mathbf{y}^{(i)}\right) + \frac{1}{2}\boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta}$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ is the parameter vector $\mathbf{x}^{(i)} \in \mathbb{R}^d$, $y^{(i)} \in \mathbb{R}$, $\{\mathbf{x}^{(i)}, y^{(i)}\}$ are N training data points, $\mathbf{A} \in \mathbb{R}^{d\times d}$ is a symmetric matrix and $b \in \mathbb{R}$. We want to find parameters $\boldsymbol{\theta}$ using gradient descent. Find the vector of partial gradients of the cost function.

(b) (1 point) Give the closed-form solution of θ from the above expression you found.

(c) (1 point) Which class of models - BERT or GPT is more suitable for the task of Machine Translation? Explain.

(d) (1 point) Word2Vec model learns multiple representations of the same word during optimization. Describe how a single representation is obtained during inference.