Indian Institute of Technology Kharagpur
Class Test 4 2020-21

Date of Examination: 30 Mar. 2021 | **Duration: 45 minutes**
Subject No.: CS60010 | **Subject: Deep Learning**
Department/Center/School: CSE | **Credits: 3** | **Full marks: 40**

---

1. Consider an implementation of the encoder of a self-attention layer as follows:

   - There is a sequence of n words in the input.
   - Each word is represented by a vector of size $d = 100$.
   - Suppose that the query and key are of dimensions $dim_q = dim_k = 36$ and that the value is of dimension $dim_v = 25$.
   - For 1.(b) and 1.(c), consider that $n = 1000$, $d = 100$, $dim_q = dim_k = 40$, $dim_v = 25$

   Suppose you use scaled dot product attention.

   (a) (8 points) i. What is the complexity of forward computation of one layer of self-attention?
   ii. What are the trainable parameters are there in the first layer of self-attention? For each vector or matrix parameter mention their dimension.
   iii. Write the expression for scaled dot product attention for the query position 5 while considering the key position 30.

   (b) (2 points) Now consider that there are K heads.
   i. What is the complexity of forward computation of one layer of self-attention.
   ii. Assuming $K = 4$ heads, how many total parameters are there now?

   (c) (4 points) Now suppose that you add a fully connected feed-forward network after the self-attention layer, which is applied to each position separately and identically. The FFN consists of two linear transformations with a ReLU activation in between.

   $$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2$$

   The dimensionality of input and output is = 100, and the inner-layer has dimensionality $d_{ff} = 200$.
   i. What is the total complexity of the self-attention followed by the FFN layer?
   ii. How many additional parameters are required in the FFN layer?

2. (10 points) Answer the following questions about BERT.

   (a) In BERT m% of the input words are masked. Consider the following sentence.

   > The giant vessel had been wedged diagonally across Egypt's Suez Canal since March 23.

   Randomly pick two word mentions to mask and write out the masked sentence.

   (b) During training how are the masked words represented? Give one example for each case with respect to the sentence above.

(c) During training what is/ are the desired output for the masked positions?

(d) What are the components of the loss function used by BERT?

3. (10 points) (a) Briefly explain in one sentence how XLnet is able to capture bidirectional context while using autoregressive language modelling and without masking.

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

(c) Electra uses Replaced Token Detection during training. With respect to the sentence in 2(a) and the masks chosen by you, give an example of a replaced token.

(d) Electra uses a Generator block and a Discriminator block. Explain the input and output of these two blocks.

4. (6 points) (a) Ignore computational complexity and speed. State one reason why you will prefer a transformer over a RNN.

(b) State one reason to prefer a RNN over a transformer.

(c) You have a Hindi corpus with 80,000 distinct words (V). While doing a NLP task, you are provided two options:

1. You may tokenize each word and choose $|V| = 80,000$ as the vocabulary size.
2. You may obtain 10,000 word pieces and tokenize based on the set of wordpieces.

i. State one advantage of the first method.

ii. State one advantage of the second method.