Habib University - City Campus
CS 335: Introduction to Deep Learning
Spring 2024
Activity Sheet # _____
Date _____

Name _____ Student ID: _____

Question 1: [0 points]

In this activity, we will be dealing with Masked Language Model i.e. BERT. More specifically, this activity deals with the the NLP task fill-mask.

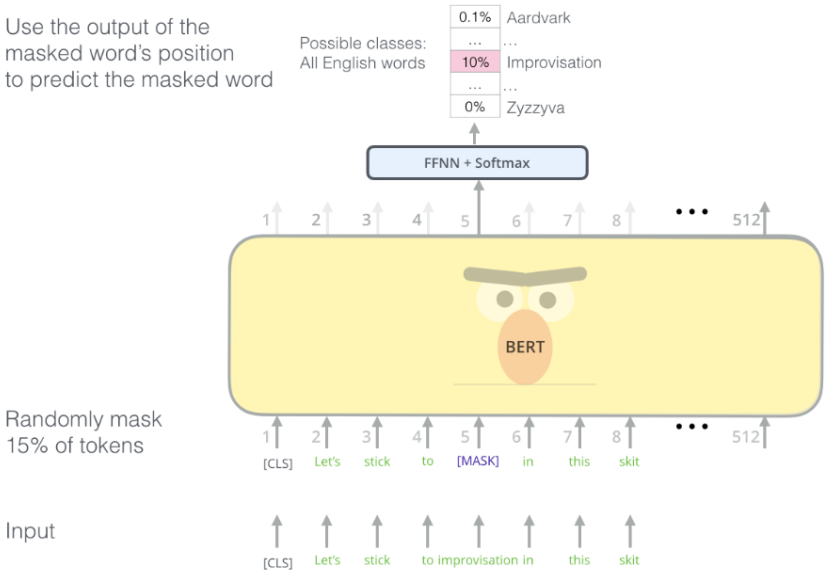

Figure 1: BERT (Bidirectional Encoder Representations from Transformers)

The BERT model in the figure above takes has a maximum sequence length of 512, the dimension of the embedding is equal to 768 and the vocabulary size is 30522. However, for this example, we assume the the following:

- The vocabulary contains the 5 unique words and the vocabulary dictionary is equal to {"Deep": 0, "learning": 1, "is": 2, "fun":3,"boring": 4}
- The maximum input length is 4
- The input sequence is "Deep Learning is fun" and the masked input is equal to "Deep Learning is $\langle MASK \rangle$ "
- The key, query, and value for each input word is as follows:

Word	Keys	Query	Value
Deep	[0.2, 0.5, 0.7]	[0.1, 0.2, 0.3]	[0.3, 0.4, 0.5]
Learning	[0.6, 0.3, 0.9]	[0.4, 0.5, 0.6]	[0.6, 0.7, 0.8]
is	[0.1, 0.8, 0.4]	[0.7, 0.4, 0.1]	[0.2, 0.9, 0.6]
MASK	[0.4, 0.6, 0.2]	[0.7, 0.2, 0.5]	[0.8, 0.5, 0.9]

(a) Compute the attention score for $\langle MASK \rangle$ by the computing the dot product of its query vector with the every other key vector include itself.

(b) Divide by the computed score by ($\sqrt{d_k} = \sqrt{3}$) where d_k is the dimension of the key.

$$score(q_i, k_j) = (\mathbf{q_i} \cdot \mathbf{k_j})/\sqrt{d_k}$$

(c) Compute the SoftMax function on the normalised attention scores

$$\alpha_{(i,j)} = softmax(score(\mathbf{q_i}, \mathbf{k_j}))$$

(d) Using the result of the previous part, computed weighted sum of the value vectors.

$$\mathbf{y_i} = \sum \alpha_{(i,j)} \mathbf{v_j}$$

(e) The output from the masked is then passed onto a FFN. Assume that logit values are equal to $[1.8, 1.7, 1.6, 0.78, 3.2]$. Apply SoftMax function on the value array and predict the value of the masked token.

(f) Compute the cross-entropy loss for the given input and predicted output.

The cross entropy loss function is defined as follows:

$$L(y, \hat{y}) = \sum_{i=1}^{j} y_j \log(\hat{y}_j)$$