



Ασκήσεις μελέτης B8-Transformers

Lab 9

Human-Computer Interaction, AUEB
Εαρινό εξάμηνο 2022-2023

Lab Assistant: Sofia Eleftheriou



Άσκηση B8.3

3. (a) We were given a BERT model pre-trained on a generic English corpus and we want to use it to build a Machine Reading Comprehension (MRC) system. The MRC system will be given a question and a paragraph (as shown in the figure) and will aim to predict the spans (sequences of tokens) of the paragraph that answer the question. The first token of each answer span should be classified as B (begin), the other tokens of the answer span as I (inside), and all the other tokens of the paragraph as O (outside). Let h_i be BERT's top-level representation of the i -th token of the paragraph and let $p_i \in [0,1]^3$ be the probability distribution over the three classes (B, I, O) produced by the MRC model for the same token. We add a task-specific dense layer on top of BERT to obtain the p_i distribution for each token of the paragraph from the corresponding h_i . **Write a formula showing how p_i is obtained from h_i , assuming $h_i \in \mathbb{R}^{128}$. Also write down the dimensions of all the matrices and vectors used in the formula.**

BERT – Fine-tuning for MRC

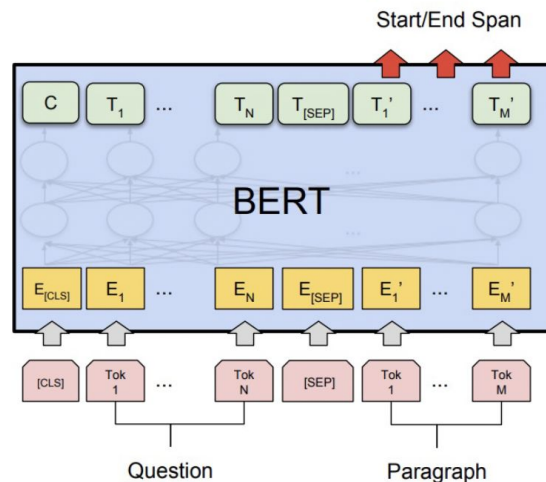


Figure from Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, 2018.



(a) Formula: $p_i = \text{softmax}(Wh_i + b)$, where $p_{i,j} = \frac{[\exp(Wh_i + b)]_j}{\sum_{k=1}^3 [\exp(Wh_i + b)]_k}$ and $j \in \{1, 2, 3\}$.

Dimensions: $W \in \mathbb{R}^{3 \times 128}, b \in \mathbb{R}^3$



(b) Write a detailed formula showing how you would compute the overall loss (L) for an **input training instance** when training the model (jointly fine-tuning BERT and training the task specific dense layer on top). Assume for simplicity that in all the training instances, the question is N_Q tokens long and the paragraph is N_P tokens long. Call t_i the correct (gold) output probability distribution for the i -th token of the paragraph. **Do not assume that every t_i is 1-hot**. For example, t_i may be a gold per-token probability distribution over the classes, based on the opinion of multiple human annotators; we may have three annotators, two of them may have said that the i -th token is a B, and the third annotator may have said it is an O, in which case the gold distribution for the token over B, I, O is 2/3, 0, 1/3.



(b) Loss: $L = - \sum_{i=1}^{N_P} \sum_{j=1}^3 t_{i,j} \log p_{i,j}$



(c) Now assume that every t_i is 1-hot. Show how the formula of the previous sub-question can be simplified. Clearly explain the steps of the simplification.



(c) Now for every token (at position i) of the paragraph, $t_{i,j} = 1$ if the correct class of the token is the j -th one, and $t_{i,j} = 0$ otherwise. Let $r(i)$ be the (index of the) correct class of the i -th token. Then the loss becomes:

$$L = - \sum_{i=1}^{N_p} t_{i,r(i)} \log p_{i,r(i)} = \sum_{i=1}^{N_p} \log p_{i,r(i)}$$

i.e., we maximize the log-likelihood of the correct classes of the paragraph's tokens.