MIT 6.806/6.864: Advanced Natural Language Processing **Spring 2021**

Homework 4

Due Date: Mon. April 26, 2021, 11:59 PM (mid-night) EDT (*Canvas submission*)

Question Answering

Introduction: In this assignment, you will explore the use of a transformer-based deep learning model called DistilBERT for a Question-Answering task. After analysing some of the basic properties of the pre-trained model, you will fine-tune it on a public QA dataset called SQuAD and evaluate it on a span-based answer extraction task. Finally, you will submit a report describing both experiments and answers to the questions in this hand-out. We have provided the scaffolding for this assignment in the following Colab notebook.

General Report Guidelines: Homework assignments should be submitted in the form of a research report on Canvas. Please upload a single PDF of your report concatenated with printouts of your code (e.g., the Jupyter Notebooks with your implementation). The report section of your submitted PDF should consist of a maximum of four single-spaced pages (6.806) or six single-spaced pages (6.864) typeset with LATEX. Reports should have one section for each part of the assignment below. Each section should describe the details of your code implementation and include whatever analysis and figures are necessary to answer the corresponding set of questions.

Please answer the following questions after you finishing your implementation on Colab.

Question 1 [Experimental]

We used contextualized word embedding calculated by `DistilBERT`. Instead of using representations from the final layer of `DistilBERT`, try using representations from its (non-contextual) *word embedding* layer. How do your results change? Why?

Question 2 [Experimental]

Report accuracy for the 3 decoding strategies you implemented in Task 5 of the notebook. Explain your results.

- Select $i = \arg\max_i \ S^i_{start}$, then select $j = \arg\max_j \ S^j_{end}$. $(i \leq j)$
- Select $j = \arg\max_j \ S^j_{end}$, then select $i = \arg\max_i \ S^i_{start}$ $(i \leq j)$
- Select (i, j) by $i, j = \arg\max_{i,j} \ S^i_{start} + S^j_{end}$ $(i \leq j)$

Question 3 [Theoretical]

Suppose you are given an input like:

- **Context**: *The film has two actors: Tom and Jerry.*
- **Question**: *Name one actor.*

This question is genuinely ambiguous: there are two names you could return. Accordingly, suppose your model outputs the following distribution (close to, but not quite, uniform):

	...two	*actors:*	*Tom*	*and*	*Jerry*	.
$P(start)$	0	0	0.51	0	0.49	0
$P(end)$	0	0	0.49	0	0.51	0

Suppose also that a prediction is considered correct if it consists of *either* the exact string *Tom* or the exact string *Jerry*.

a. Pick one of the decoding strategies you implemented in `logits_to_ans_loc()`. Using this strategy when decoding with the scores above, what's the probability of getting the answer correct?

b. If you sample start and end positions independently from $P(start)$ and $P(end)$, what's the probability of getting the answer correct?

c. Describe (but don't implement) an alternative version of the question answering model and decoding strategy that will return *Tom* with roughly 50% probability and *Jerry* with roughly 50% probability. You can modify any part of the QA model you implemented for HW4 (modeling, decoding, etc).

(6.864) Question 4 [Experimental]

Implement changes that improve the performance of your question-answering model, and describe what you did. for example: design up with a different decoding algorithm (think about your answer to the previous question) or change something about the way questions and answers are encoded (feel free to refer to the question answering lecture).