CS 4372
ASSIGNMENT 4

Names of students in your group: Natalie Pedigo (nbp220000), Kylie Quinney (krq210000)

Number of free late days used: 0

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

The Use of Transformers in Chat Bots

The input text (The Return of Sherlock Holmes, in our case) is tokenized and then fed into the transformer model (AlbertForQuestionAnswering). Each token is then encoded into embeddings, so that the model can understand the position and context of each token. These features of a token are what allows AlbertForQuestionAnswering to be a powerful model for question-answering. Transformers can process the input to find contextual links between the question and relevant parts of the text, and can use its fine-tuned understanding to predict what the answer is likely to be. The most probable answer is outputted with a confidence score.

The AlbertForQuestionAnswering Model

AlbertForQuestionAnswering has a transformer encoder architecture. It has multiple layers of transformer encoders that use self-attention to understand relationships between tokens. AlbertForQuestionAnswering also implements parameter sharing, which allows the model to reduce memory usage. This is done by implementing the same parameters for all layers, instead of different parameters for each layer, thus giving us a more compact model. The overall structure of AlbertForQuestionAnswering includes input embeddings and stacked transformer encoded layers, which produces probable outputs. A core component of the AlbertForQuestionAnswering model, and other models as well, is self-attention.

Hyper-parameter Tuning

Iteration	hyper-params	Training loss	Validation loss	Score (on book q)	Answers to Qs	
1	NA	NA	NA	Q1: 0.7764191627502441 Q2: 0.01717839017510414 Q3: 0.01087028719484806	Q1: 1894 Q2: Sherlock Holmes Q3: THE ADVENTURE OF THE EMPTY HOUSE	Using ALBERT untouched
2	learning rate: 2e-5 train batch size: 16 eval batch size: 16 training epochs: 5 weight decay: 0.01	epoch 1: NA epoch 2: 2.681100 epoch 3: 2.681100 epoch 4: 1.135000 epoch 5: 1.135000	epoch 1: 2.385232 epoch 2: 1.704719 epoch 3: 1.559099 epoch 4: 1.550689 epoch 5: 1.564489	Q1: 0.9822187423706055 Q2: 0.589024007320404 Q3: 0.8222362995147705	Q1: 1894 Q2: Sherlock Holmes Q3: Sherlock Holmes	
3	learning rate: 2e-5 train batch size: 16 eval batch size: 16 training epochs: 4 weight decay: 0.01	epoch 1: NA epoch 2: 2.585500 epoch 3: 2.585500 epoch 4: 1.114100	epoch 1: 2.092022 epoch 2: 1.664638 epoch 3: 1.570670 epoch 4: 1.595282	Q1: 0.9458447694778442 Q2: 0.720957338809967 Q3: 0.7049626708030701	Q1: 1894 Q2: young Adair Q3: the Park Lane Mystery	
4	learning rate: 2e-5 train batch size: 16 eval batch size: 16 training epochs: 4 weight decay: 0.10	epoch 1: NA epoch 2: 2.599900 epoch 3: 2.599900 epoch 4: 1.144200	epoch 1: 2.163788 epoch 2: 1.670609 epoch 3: 1.580346 epoch 4: 1.580813	Q1: 0.8227441310882568 Q2: 0.6674026250839233 Q3: 0.44598376750946045	Q1: 1894 Q2: Lady Maynooth Q3: the Park Lane Mystery	
5	learning rate: 2e-5 train batch size: 16 eval batch size: 16 training epochs: 4 weight decay: 0.001	epoch 1: NA epoch 2: 2.596300 epoch 3: 2.596300 epoch 4: 1.133100	epoch 1: 2.216497 epoch 2: 1.658523 epoch 3: 1.584609 epoch 4: 1.593405	Q1: 0.8897749185562134 Q2: 0.4277809262275696 Q3: 0.8366639614105225	Q1: 1894 Q2: Lady Maynooth Q3: the Park Lane Mystery	
6	learning rate: 2e-3 train batch size: 16 eval batch size: 16 training epochs: 4 weight decay: 0.01	epoch 1: NA epoch 2: 5.967700 epoch 3: 5.967700 epoch 4: 5.961100	epoch 1: 5.950644 epoch 2: 5.950644 epoch 3: 5.950644 epoch 4: 5.950644	Q1: 3.857876072288491e-05 Q2: 3.585643207770772e-05 Q3: 3.585643207770772e-05	Q1: the circumstances of the Park Lane Mystery, which were further Q2: a Q3: wound which must have caused instantaneous death. Such were the circumstances of the	
7	learning rate: 2e-7 train batch size: 16 eval batch size: 16 training epochs: 4 weight decay: 0.01	epoch 1: NA epoch 2: 5.862200 epoch 3: 5.862200 epoch 4: 5.775200	epoch 1: 5.856169 epoch 2: 5.801681 epoch 3: 5.766145 epoch 4: 5.753845	Q1: 0.0001003671932267025 Q2: 9.018548007588834e-05 Q3: 9.10957096493803e-05	Q1: Street end of Q2: Street end of Q3: Street end of	

Results

The hyper-parameters used in the second and third iteration seem to produce the best results from our trials. We believe the third iteration to be the most accurate in the answers it produced. The third iteration also showed a consistently high confidence score.

Result Analysis

Using the AlbertForQuestionAnswering model untouched, the only high scoring answer was to the first question, while the next two had confidence scores of only 0.01. This showed us that we needed to tune the AlbertForQuestionAnswering model to our data in order to get correct answers and higher confidence scores. The second and third iterations showed much improvement, however the second iteration had some inconsistencies. Although it scored very high for the first question (0.98), it was not able to assess the next question as well. We see the third iteration gave us consistently higher confidence. It is possible, however, that the model consistently scored worse on the second question because not enough data was given. If we chose a wider sample of the text to use to train the model, there is a chance that all questions could have scored higher.

We conclude that the third iteration's hyper parameters were the most useful in obtaining accurate results from our model.

Assumptions

Instead of using the entire book, we used only a portion to save time and space. Therefore, the model considers only a portion of the text to answer questions.