Introduction to Artificial Intelligence HW4 Report

110550088 李杰穎

May 21, 2022

1 Attention Mechanism of BERT

Figure 1: Screenshot of ExBERT. [1]

BERT (Bidirectional Encoder Representation for Transformers) is an NLP model based on transformers. BERT is pre-trained on two tasks, masked language model and next sentence prediction.

Masked language model is a task that language model needs to predict the masked word. For example, in following sentence, "The man went to the [MASK] to buy a [MASK] of milk.". There

[1] From original paper `https://arxiv.org/pdf/1910.05276.pdf`

are two words being masked. BERT needs to predict that those two masked words are "store" and "gallon", respectively.

Next sentence prediction is a task that model will be given two sentences, let's say sentence A and B. It needs to tell us is B the next sentence of A.

In this section, I will use ExBERT, a visualizing tool for different variances of BERT, including `bert-base-cased` and `distilbert-based-uncased`, to understand the attention mechanisms of BERT.

1.1 Using BERT as Masked Language Model

Masked Language Model (MLM) is a task that its input is a sentence with part of words being masked. The goal of language model is to predict the masked words using the context of sentence.

In this section, I will use $\text{BERT}_{\text{base}}$ as language model.

As we can see in Figure 2, if we mask the word "escape" and let BERT predict which word should appear here. In layer 9 of BERT, the words that get attention are "ran", "to" and "din". These words are exactly the words that is relevance with the masked word "escape". This is an example of attention mechanism.

Another example is shown as Figure 3. I mask the subject and see how BERT know what word should be filled in. As a normal human would do. BERT pay its attention on the subject of the first sentence, which is my name "Jay". We can also observe that "He" has the highest chance to appear in the masked position. This also shows that BERT knows Jay is usually the name of a male.

The last example is to show that BERT is able to know that who does a specific pronoun refer to. As we can see in the Figure 4, BERT pays its attention at "Mary" when it predict the masked word, instead of "Jay". This shows that BERT can analysis the grammar structure in

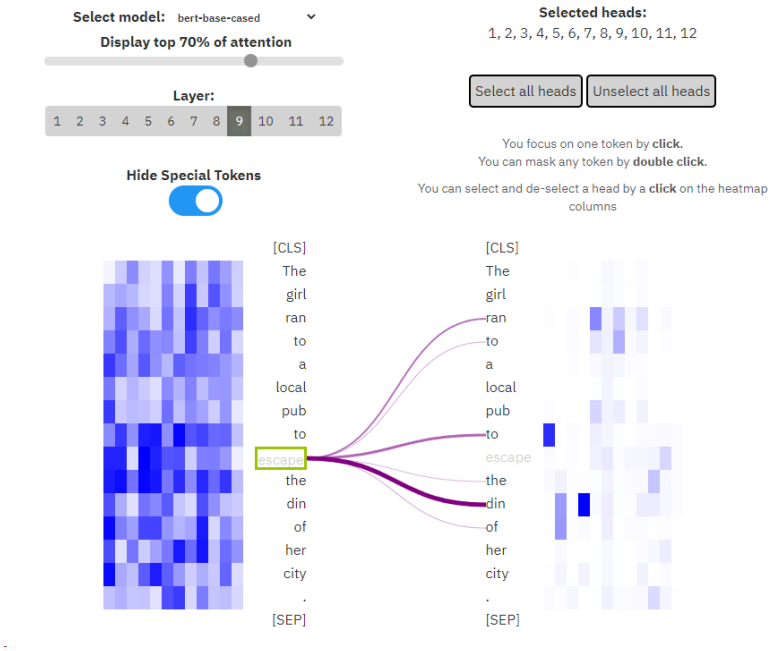

Figure 2: Using the sentence "The girl ran to a local pub to escape the din of her city." and mask the word "escape".

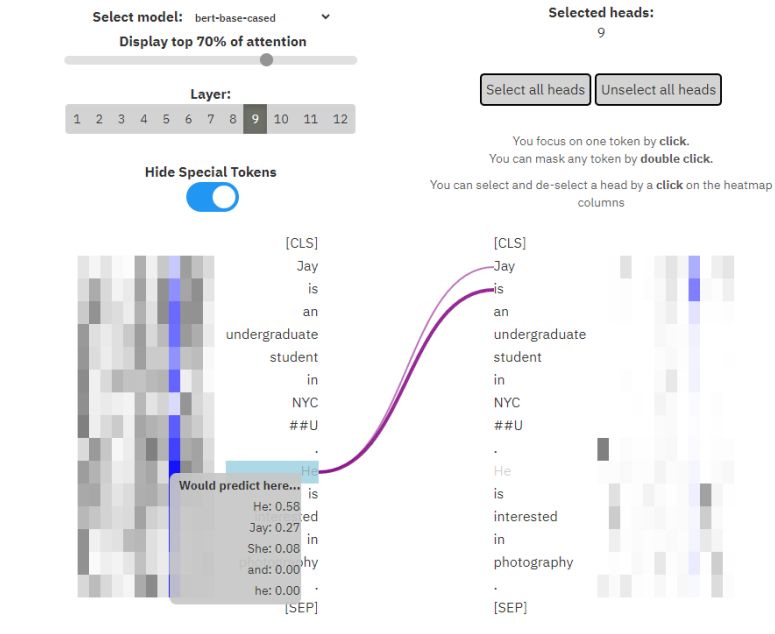

Figure 3: Using the sentence "Jay is an undergraduate student in NYCU. He is interested in photography." and mask the word "He".

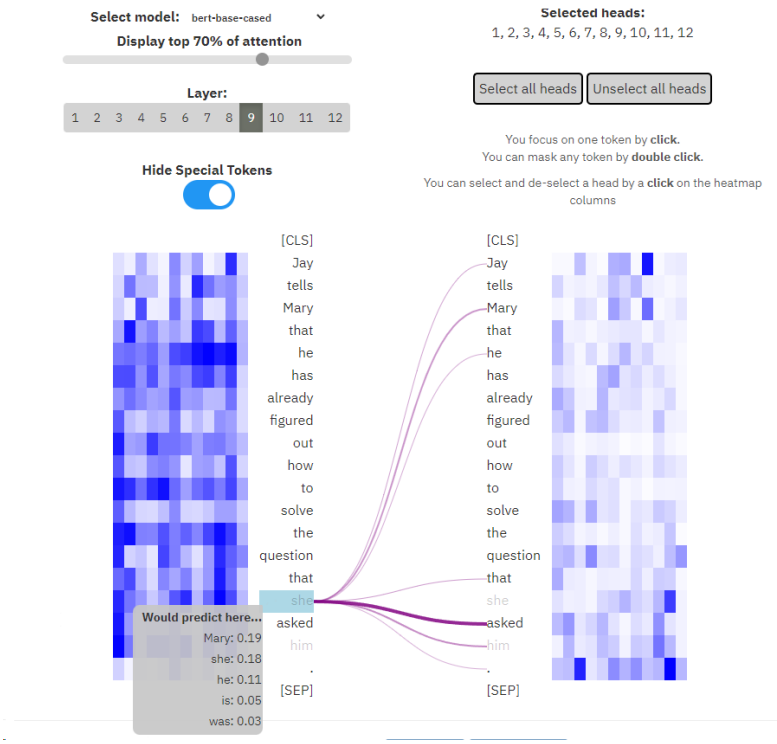

Figure 4: Using the sentence "Jay tells Mary that he has already figured out how to solve the question that she asked him." and mask "she" and "him" in the second sentence.

the sentence and know which word should be paid attention at.

In conclusion, the examples above show that how attention mechanism work in BERT. And indeed, this mechanism helps BERT perform better compared with other methods like n-gram model or ELMo.

2 Comparison of BERT and DistilBERT

DistilBERT is smaller version of BERT. The basic idea of DistilBERT is to use a light-weight model architecture to "mimic" the behavior of the original BERT. The author of DistilBERT claimed that this method can "reduce the size of a BERT by 40%, while retaining 97% of its language understanding capabilities and being 60% faster."[2]

In ExBERT, we can use both `bert-base-cased` and `distilbert-base-uncased`. Thus, we will also use ExBERT to compare these two model.

To compare DistilBERT with BERT. I use the same three sentences as in previous section. Below are the experiments results. Noticed that these results are all from layer 3 of DistilBERT. This is because after some observations, I think this is layer that DistilBERT pay attention to the context of sentence. Other layers are paying attention to either the previous word or the next word or the [CLS] tag.

As we can see in Figure 5, DistilBERT is able to predict correctly, and the words it pays attention to is similar to those BERT pays attention to. This shows that DistilBERT learn some "knowledges" from the original BERT.

But if we look at more complicate examples, like sentences in Figure 6 and Figure 9, DistilBERT can't predict the masked word as well as original BERT can do.

Let's first look at Figure 6. DistilBERT can predict that the masked word can be the name "Jay"

[2]From abstract of original paper. `https://arxiv.org/abs/1910.01108`

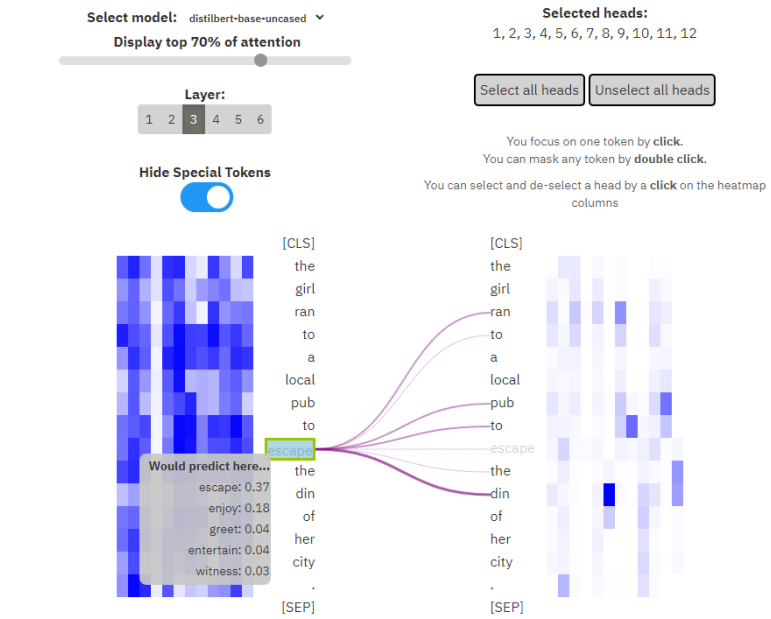

Figure 5: Using the sentence "The girl ran to a local pub to escape the din of her city." and mask the word "escape".

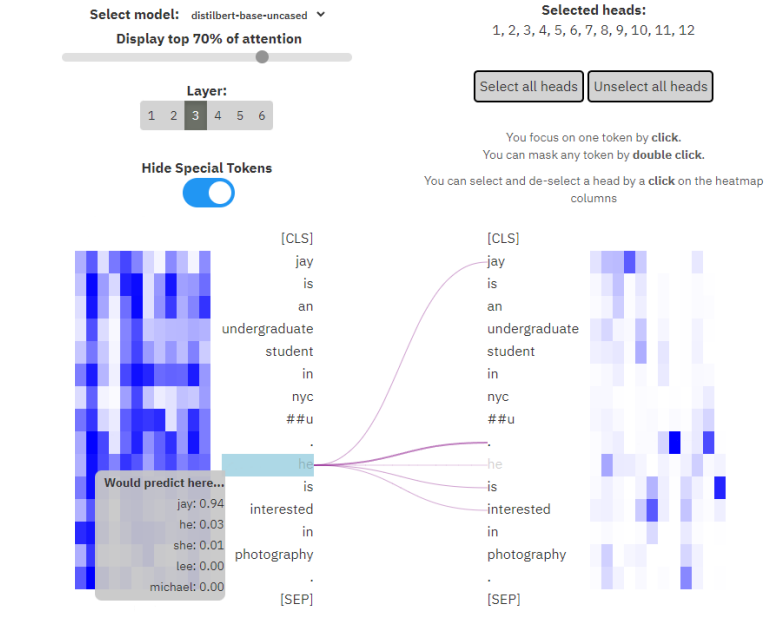

Figure 6: Using the sentence "Jay is an undergraduate student in NYCU. He is interested in photography." and mask the word "He".

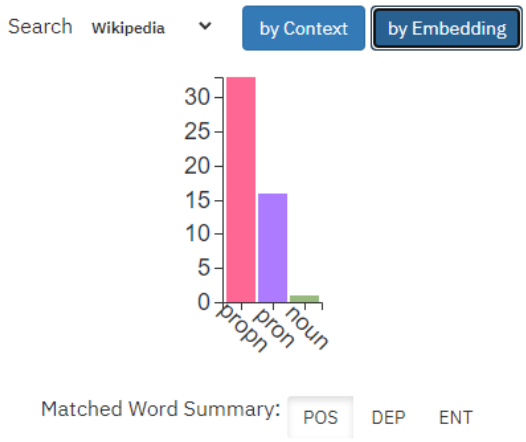

Figure 7: The matched word summary of DistilBERT using the sentence "Jay is an undergraduate student in NYCU. He is interested in photography." and mask the word "He".

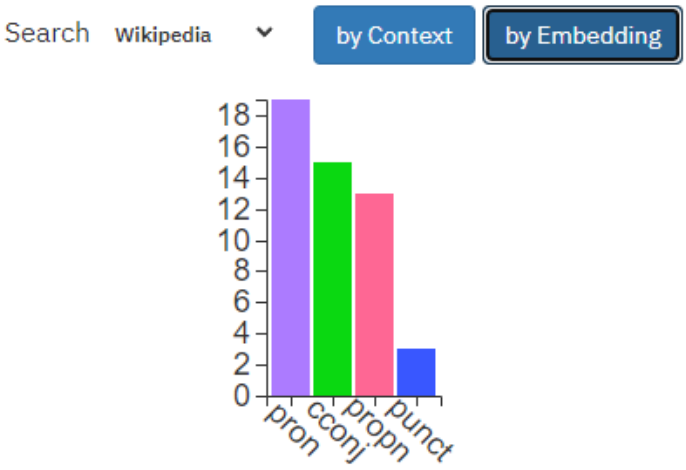

Figure 8: The matched word summary of BERT using the sentence "Jay is an undergraduate student in NYCU. He is interested in photography." and mask the word "He".

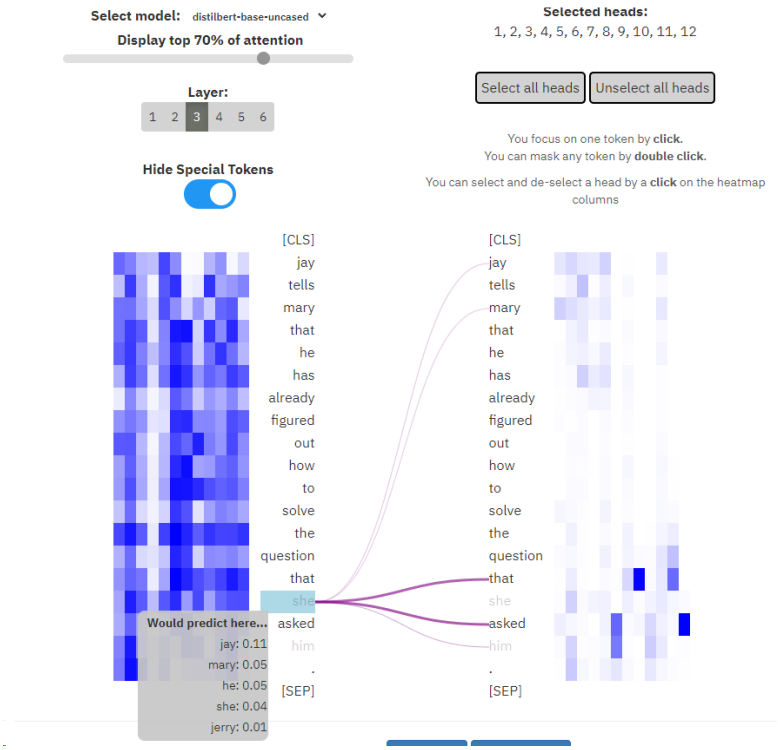

Figure 9: Using the sentence "Jay tells Mary that he has already figured out how to solve the question that she asked him." and mask "she" and "him" in the second sentence.

or “He”. Although both of these is correct answer and model knows that Jay is a male name. But the better answer is “He”, which is closer to daily English usage. But DistilBERT only has 3% of probability to fill “He” here. This shows that DistilBERT perform worse than original BERT. We can also use “by Embedding” to see the matched word. Furthermore, Figure 7 shows that the matched word summary in the last layer of DistilBERT. In contrast, Figure 8 shows that the matched word summary in the last layer of BERT. We can see that BERT matched more pronoun than DistilBERT does, which is a better matched compared with proper noun. This also shows the difference between BERT and DistilBERT.

Figure 9 is the last example. Compared with the result shown in Figure 4, DistilBERT’s prediction is completely wrong, the most possible word is “Jay”, which is the last word should be filled in. This example shows that there has significant performance difference between BERT and DistilBERT.

3 Comparison of the Explanation of LIME and SHAP

4 Implementation of Other Explanation Technique

5 Attack NLP Model

6 Encountered Problems