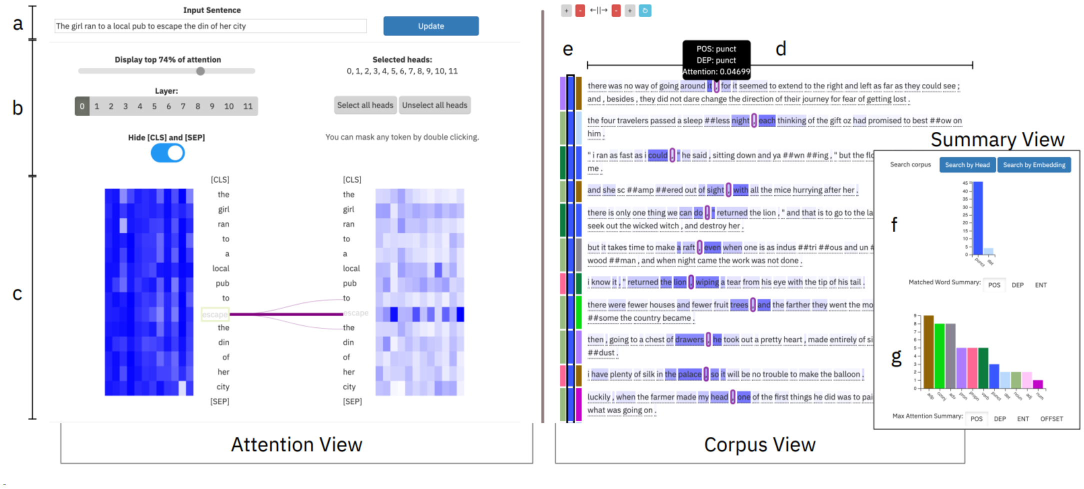
Introduction to Artificial Intelligence HW4 Report

110550088 李杰穎

May 22, 2022

1 Attention Mechanism of BERT



Input Sentence: The girl ran to a local pub to escape the din of her city

Display top 74% of attention

Layer: 0 1 2 3 4 5 6 7 8 9 10 11

Selected heads: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11

Hide [CLS] and [SEP]

You can mask any token by double clicking.

POS: punct
DEP: punct
Attention: 0.04699

there was no way of going around it , for it seemed to extend to the right and left as far as they could see ; and , besides , they did not dare change the direction of their journey for fear of getting lost .

the four travelers passed a sleep ##less night , each thinking of the gift oz had promised to best ##ow on him .

" i ran as fast as i could , " he said , sitting down and ya ##wn ##ing , " but the flo me .

and she sc ##amp ##ered out of sight , with all the mice hurrying after her .

there is only one thing we can do , " returned the lion , " and that is to go to the la seek out the wicked witch , and destroy her .

but it takes time to make a raft , even when one is as indus ##tri ##ous and un # wood ##man , and when night came the work was not done .

i know it , " returned the lion , wiping a tear from his eye with the tip of his tail .

there were fewer houses and fewer fruit trees , and the farther they went the mo ##some the country became .

then , going to a chest of drawers , he took out a pretty heart , made entirely of si ##dust .

i have plenty of silk in the palace , so it will be no trouble to make the balloon .

luckily , when the farmer made my head , one of the first things he did was to pai what was going on .

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

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

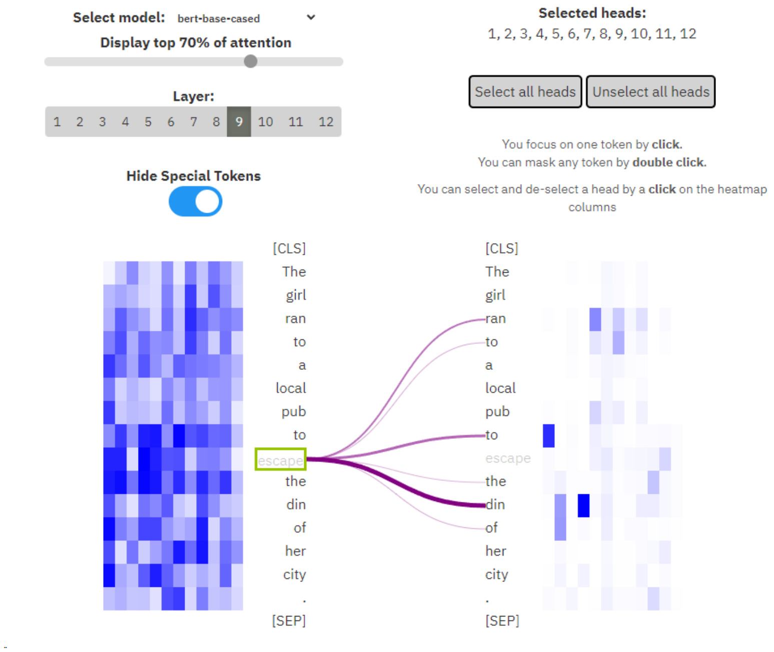
Next sentence prediction is a task that model will be given two sentences, let's say sentence A and B. It needs to tell us is B the next sentence of A.

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 BERT$_{\text{base}}$ as language model.



Select model: bert-base-cased

Display top 70% of attention

Layer: 9

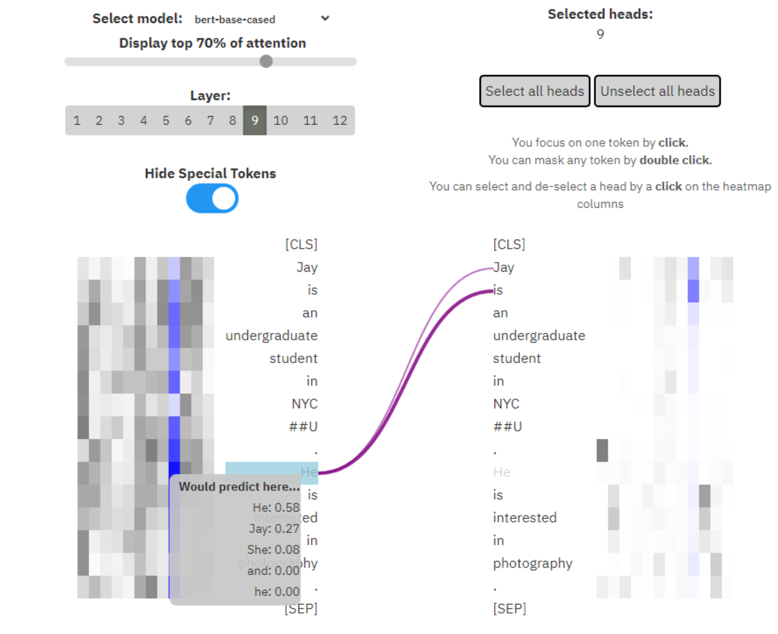
Selected heads: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

You focus on one token by **click**.
You can mask any token by **double click**.

You can select and de-select a head by a **click** on the heatmap columns

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

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.


Select model: bert-base-cased

Display top 70% of attention

Layer: 9

Selected heads: 9

Hide Special Tokens

You focus on one token by **click.**
You can mask any token by **double click.**

You can select and de-select a head by a **click** on the heatmap columns

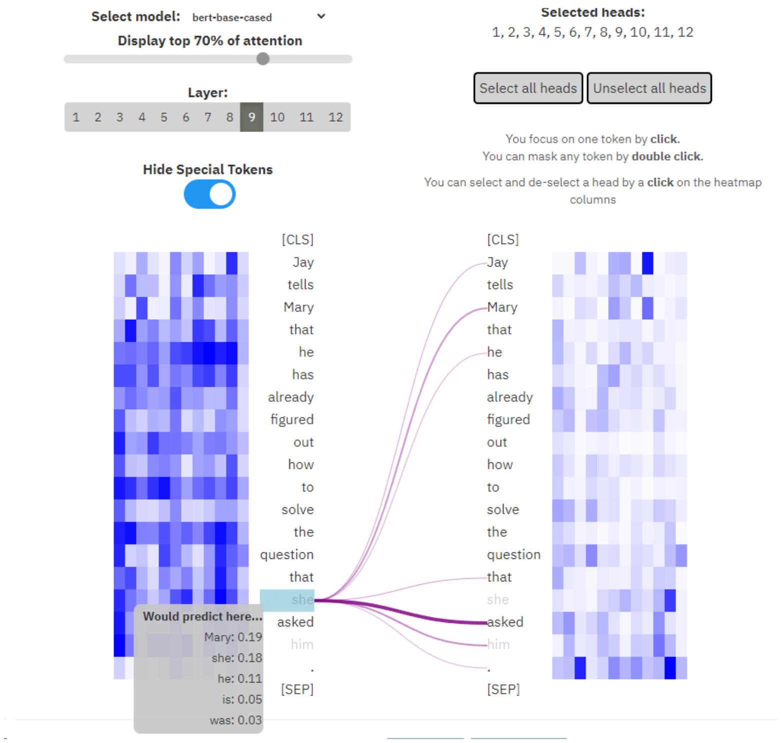
Would predict here...
He: 0.58
Jay: 0.27
She: 0.08
and: 0.00
he: 0.00

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

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 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.



Select model: bert-base-cased

Display top 70% of attention

Layer: 9

Hide Special Tokens

Selected heads:
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

You focus on one token by **click.**
You can mask any token by **double click.**

You can select and de-select a head by a **click** on the heatmap columns

```
Would predict here...
Mary: 0.19
she: 0.18
he: 0.11
is: 0.05
was: 0.03
```

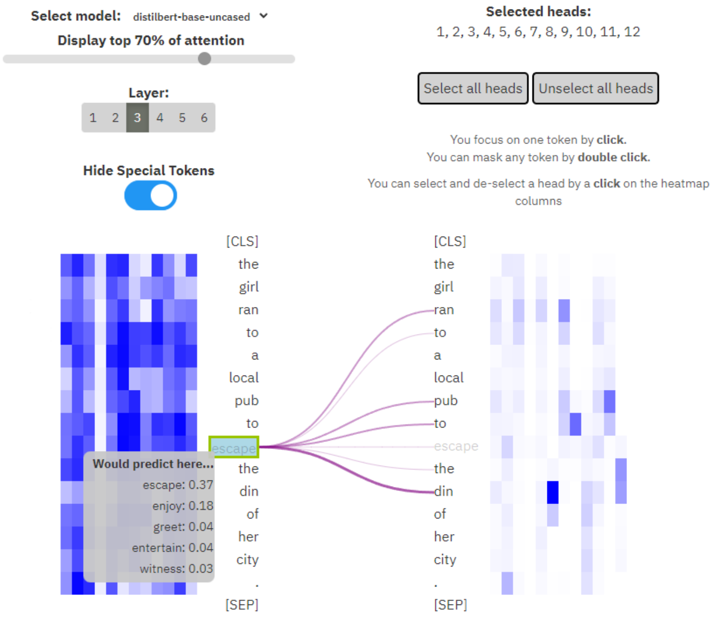
Figure 4: Using the sentence “Jay tells Mary that he has already figured out how to solve the question that she asked him.” and mask “she” and “him” in the second sentence.

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.



Select model: distilbert-base-uncased

Display top 70% of attention

Layer: 3

Selected heads: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

You focus on one token by **click.**
You can mask any token by **double click.**

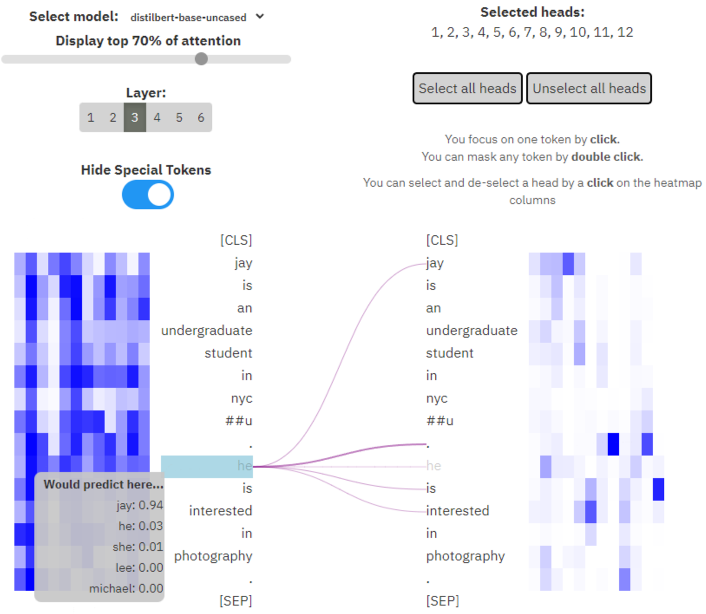
You can select and de-select a head by a **click** on the heatmap columns

Would predict here...
escape: 0.37
enjoy: 0.18
greet: 0.04
entertain: 0.04
witness: 0.03

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

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.

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



Select model: distilbert-base-uncased

Display top 70% of attention

Layer: 3

Hide Special Tokens

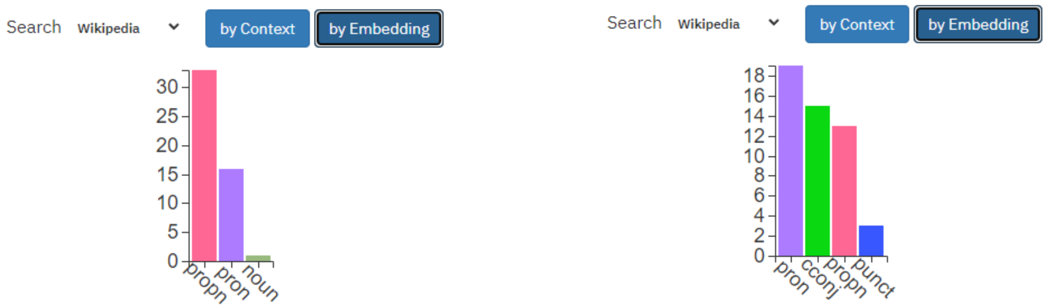
Selected heads: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

You focus on one token by **click.**
You can mask any token by **double click.**

You can select and de-select a head by a **click** on the heatmap columns

Would predict here...
jay: 0.94
he: 0.03
she: 0.01
lee: 0.00
michael: 0.00

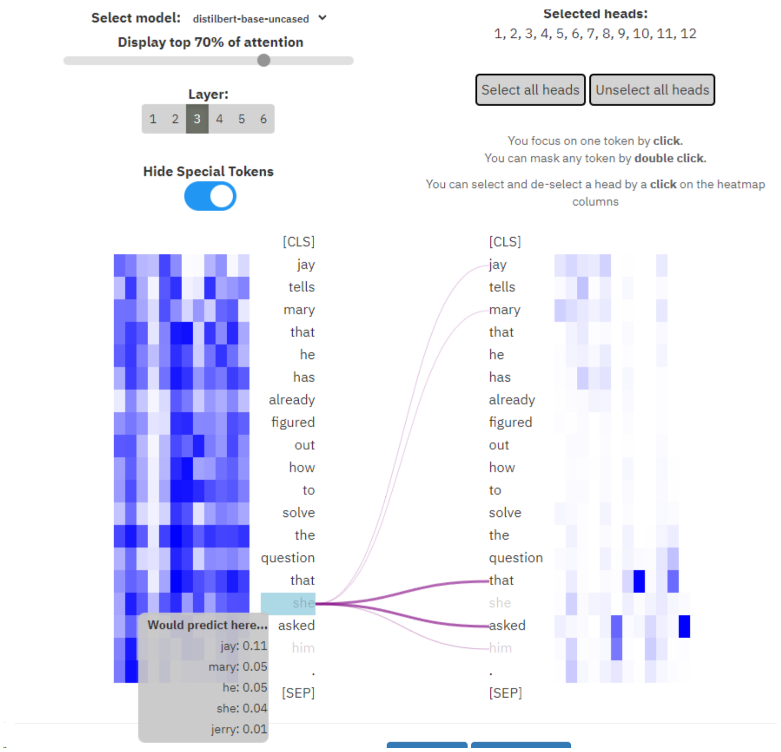
Figure 6: Using the sentence “Jay is an undergraduate student in NYCU. He is interested in photography.” and mask the word “He”.



(a) DistilBERT

(b) BERT

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



Select model: distilbert-base-uncased

Display top 70% of attention

Layer: 3

Hide Special Tokens

Selected heads:
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

You focus on one token by **click**.
You can mask any token by **double click**.

You can select and de-select a head by a **click** on the heatmap columns

Would predict here...
jay: 0.11
mary: 0.05
he: 0.05
she: 0.04
jerry: 0.01

Figure 8: 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.

But if we look at more complicate examples, like sentences in Figure 6 and Figure 8, 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" 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.

Furthermore, we can also use the "by Embedding" button to see the matched word. Figure 7a shows that the matched word summary in the last layer of DistilBERT. In contrast, Figure 7b 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 8 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 a significant performance difference between BERT and DistilBERT.

In conclusion, although DistilBERT can predict words correctly in some simple sentences, but if the sentences are too complicate, DistilBERT may not perform very well.

3 Explainable AI: LIME and SHAP

Nowadays, people have trained lots of AI models to do various of tasks. But most of these models are black boxes, we don't know how machine makes its decision. This makes us hard to trust the prediction of machine. Even worse, if the models make some wrong decision, it's nearly impossible to debug those AI models. Therefore, we need to conceive methods to explain behavior of AI models. In this section, I will discuss two methods that can explain AI models. One of them is local interpretable model-agnostic explanations (LIME) and another is shapley additive explanations (SHAP).

3.1 Local Interpretable Model-agnostic Explanations (LIME)

LIME is an method to explain AI model. Its name clearly shows its properties.

1. **Local**: LIME explain the given prediction locally.
2. **Interpretable**: The explanation given by LIME is easy enough for human to understand.
3. **Model-agnostic**: Treat the model as a black box so that LIME works on every model.

Given a prediction we want to explain, LIME first creates a datasets by perturbed the given input. Then it uses this datasets to train a simple (often a linear classifier) and explainable classifier. The performance of this classifier need to be good enough on the new datasets. In this way, we can achieved the so-called "local fidelity". Figure 9 illustrates above procedures. $g(z')$ is the classifier we want to train.

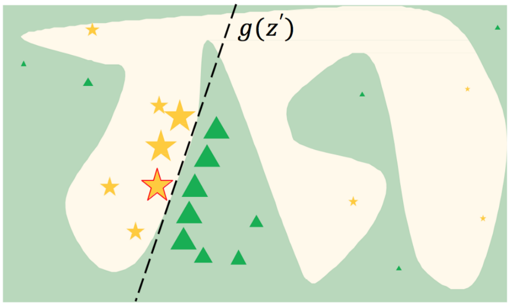


Figure 9: How LIME works.[3]

We can noticed that LIME only works on a single prediction. We can't get a global explanation using LIME. On the contrast, the next method I will introduce works both on local and global explanation.

3.2 Shapley Additive Explanations (SHAP)

SHAP is an method based on Shapley value. Shapley value is first proposed by an American mathematician Lloyd Stowell Shapley. Shapley value can be used to compute the contribution to the final output of model for a given feature. The following equation is for computing the Shapley value for a given feature j.

[3] From `https://reurl.cc/3o4YE8`

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} \left[f_{S \cup \{i\}} \left(x_{S \cup \{i\}} \right) - f_S \left(x_S \right) \right] \tag{1}$$

Where F is all the features used to build the model. $f_S(x_S)$ is the output of the model trained with feature set S.[4]

We can also noticed that SHAP deals with features, it can also give the global explanation. In the next subsection, I will compare the local explanation given by LIME and SHAP. And I will also show the global explanation given by SHAP using the IMDB datasets.

3.3 Explanation of LIME and SHAP

4 Other Explanation Technique

5 Attack NLP Model

6 Encountered Problems

[4] For detailed explanation, please refer to the original paper or this article. https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html