Attention: What is being attended to ?

What does BERT look at ? (https://arxiv.org/pdf/1906.04341.pdf) is a paper that attempts to interpret how attention is being used in the BERT Transformer model.

We give a very brief overview of the author's theories.

The authors answer the question through the mechanism of an *attention map*

- For each attention head in the model
    - e.g., head j at layer l
- For each input token $\mathbf{x}_{(t)}$
- Record the attention weight $\mathbf{w}_{j,l,t,t'}$ of the head on input token t'

By using pre-determined relationships between input tokens

- the authors hope to uncover the "concepts" that the head is attending to

For example

- attending to the previous or subsequent token
- attention to periods ("line end" character)

Attention to position

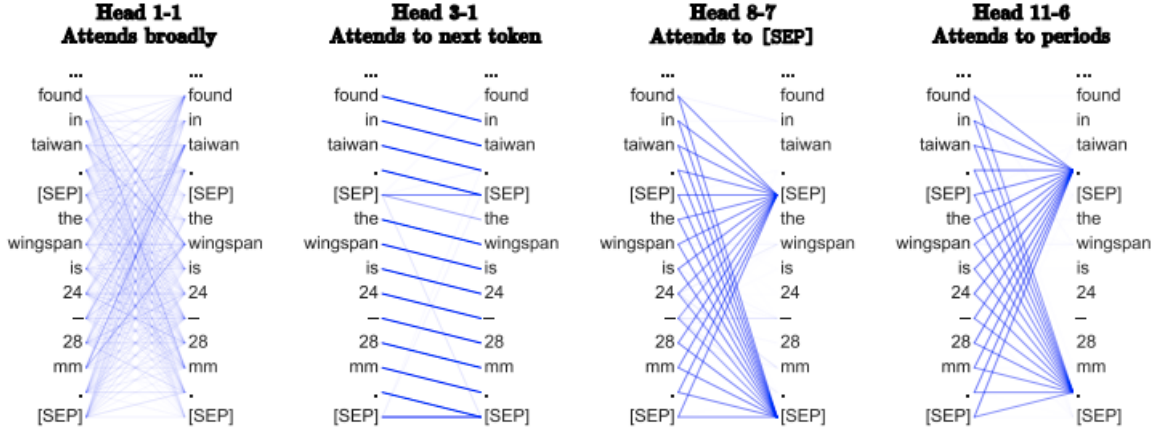


Figure 1: Examples of heads exhibiting the patterns discussed in Section 3. The darkness of a line indicates the strength of the attention weight (some attention weights are so low they are invisible).

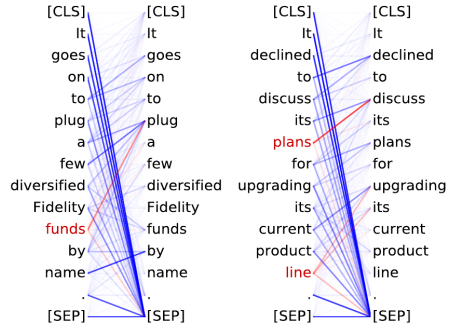
Or, certain "linguistic phenomena"

- an object attending to their verb
- a noun modifier attending to the noun

Attention to linguistic phenomena

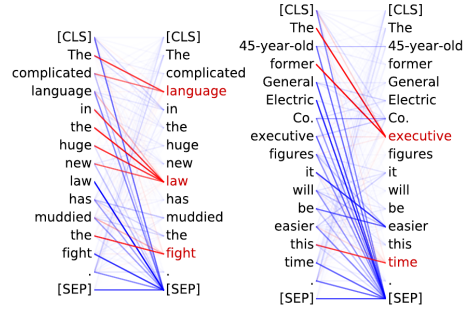
Head 8-10

- **Direct objects** attend to their verbs
- 86.8% accuracy at the `dobj` relation



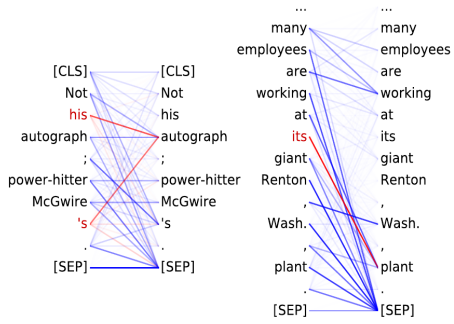
Head 8-11

- **Noun modifiers** (e.g., determiners) attend to their noun
- 94.3% accuracy at the `det` relation



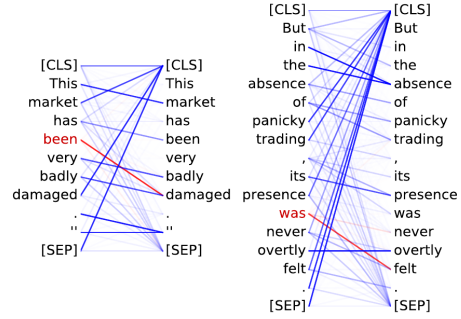
Head 7-6

- **Possessive pronouns** and apostrophes attend to the head of the corresponding NP
- 80.5% accuracy at the `poss` relation



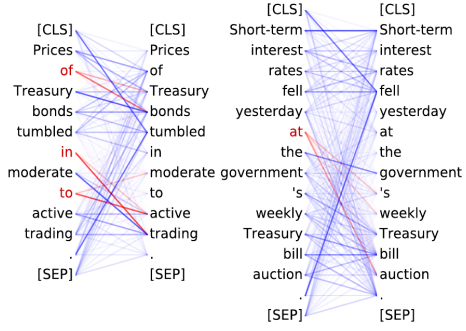
Head 4-10

- **Passive auxiliary verbs** attend to the verb they modify
- 82.5% accuracy at the `auxpass` relation



Head 9-6

- **Prepositions** attend to their objects
- 76.3% accuracy at the `pobj` relation



Head 5-4

- **Coreferent** mentions attend to their antecedents
- 65.1% accuracy at linking the head of a coreferent mention to the head of an antecedent

Also interesting is attention to "special" tokens added to the raw text input.

A typical input is bracketed by special tokens: `[CLS], [SEP]`

```
[CLS] This is the first part [SEP] This is the second part [SEP]
```

- Attention to `[SEP]`
- Attention to `[CLS]`

Attention to special tokens

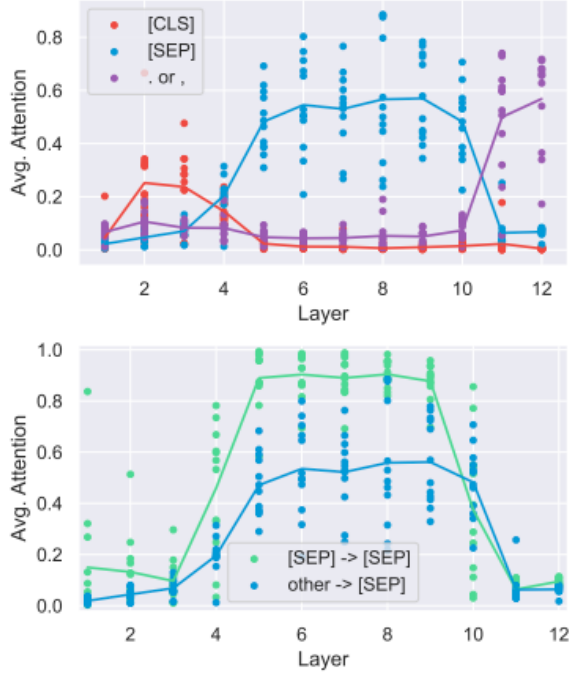


Figure 2: Each point corresponds to the average attention a particular BERT attention head puts toward a token type. Above: heads often attend to "special" tokens. Early heads attend to [CLS], middle heads attend to [SEP], and deep heads attend to periods and commas. Often more than half of a head's total attention is to these tokens. Below: heads attend to [SEP] tokens even more when the current token is [SEP] itself.

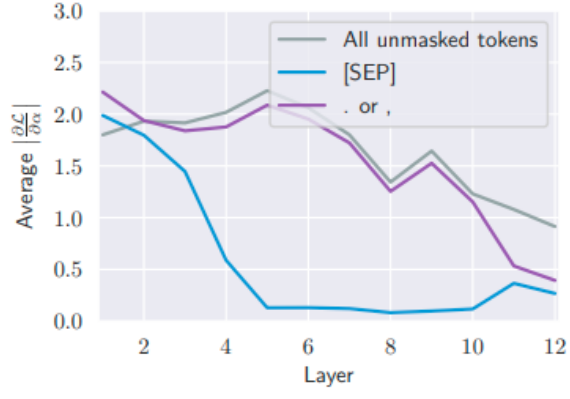


Figure 3: Gradient-based feature importance estimates for attention to [SEP], periods/commas, and other tokens.

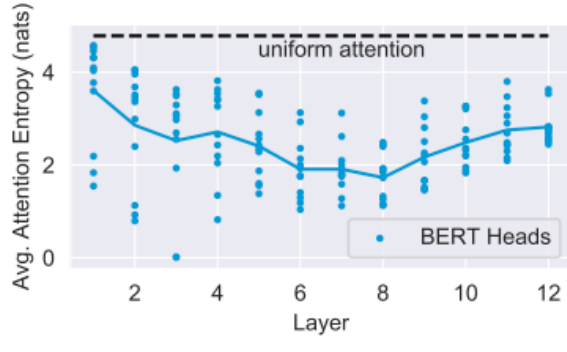


Figure 4: Entropies of attention distributions. In the first layer there are particularly high-entropy heads that produce bag-of-vector-like representations.

Setup. We extract the attention maps from BERT

Attention to [SEP] was interpreted in a particularly interesting manner

- it is a "no op": when there is nothing in particular to attend to, attend to [SEP]