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
 - lacktriangledown e.g., head j at layer l
- ullet For each input token ${f x}_{(t)}$
- ullet 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

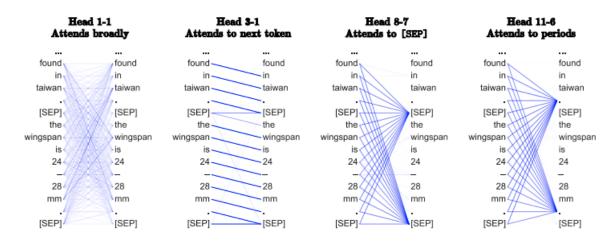


Figure 1: Examples of heads exhibiting the patterns discussed in Section 3. The darkness of a line indicates the strength of the attention weight (some attention weights are so low they are invisible).

Or, certain "linguistic phenomena"

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

Attention to linguistic phenomena

Head 8-10 Head 8-11 - Direct objects attend to their verbs - Noun modifiers (e.g., determiners) attend to their noun - 86.8% accuracy at the dobj relation - 94.3% accuracy at the det relation [CLS] [CLS] [CLS] [CLS] Ιt Ιt The goes declined declined [CLS] [CLS] aoes 45-year-old on on to to formercomplicated complicated to to discuss discuss General language language Electric its its plug plug Co. plans plans a the the executive few few for for huge huge figures new diversified diversified upgrading upgrading new. Fidelity Fidelity its its will has has funds be funds current current muddied muddied (easier product product by by thethisname name line line fight fiaht time [SEP] [SEP] [SEP] [SEP] [SEP] [SEP] [SEP] Head 7-6 Head 4-10 - Possessive pronouns and apostrophes - Passive auxiliary verbs attend to the attend to the head of the corresponding NP verb they modify - 80.5% accuracy at the poss relation - 82.5% accuracy at the auxpass relation many employees employees [CLS] [CLS] [CLS] [CLS] the This This absence working working Not \ Not market market his at has panicky has autograph autograph its trading been been giant giant very very power-hitter power-hitter Renton Renton badly badly its McGwire McGwire damaged 4 damaged presence Wash. Wash. never [SEP] SEP] overtly [SEP] plant plant felt 4 [SEP]-[SEP] [SEP] Head 9-6 Head 5-4 - Prepositions attend to their objects - Coreferent mentions attend to their antecedents - 76.3% accuracy at the pobj relation - 65.1% accuracy at linking the head of a coreferent mention to the head of an antecedent [CLS] [CLS] [CLS] [CLS] Short-term-Short-term with with Prices Prices interest interest Kim Kim ioininaof rates rates today today peace Treasury -Treasury fell fell talks as as bonds bonds yesterday yesterday she she between tumbled tumbled at got got Israel some some and the moderate moderate government government expert expert the opinions Palestinians opinions

active

[SEP]

trading

active

trading

[SEP]

weekly

auction

[SEP]

bill

Treasury

weekly

Treasury

auction

[SEP]

on-

to

her home

the

damage

on-

-the

-to

her

home

[CLS]

former

General

Electric

executive

figures

will

easier

this

time

SEP]

[CLS]

But

the

its

absence

panicky

trading

presence was

never

overtly

felt

[SEP]

ioinina

peace

between

Palestinians

negotiations

talks

Israel

and

the

The

are

The

are

damage negotiations

45-year-old

The

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

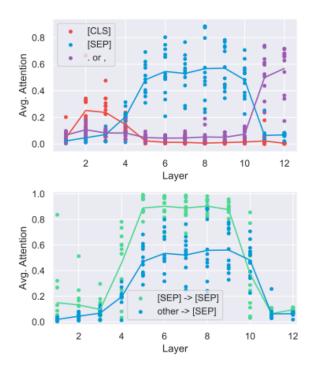


Figure 2: Each point corresponds to the average attention a particular BERT attention head puts toward a token type. Above: heads often attend to "special" tokens. Early heads attend to [CLS], middle heads attend to [SEP], and deep heads attend to periods and commas. Often more than half of a head's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

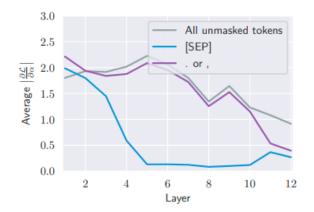


Figure 3: Gradient-based feature importance estimates for attention to [SEP], periods/commas, and other tokens.

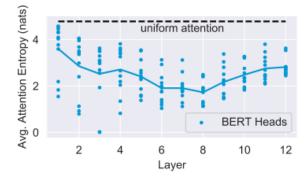


Figure 4: Entropies of attention distributions. In the first layer there are particularly high-entropy heads that produce bag-of-vector-like representations.

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]