Paper: What Does BERT Look At? An Analysis of BERT’s Attention.

Summarization:

The paper proposes methods for analyzing the attention mechanism or attention maps of the pretrained model. BERT's attention head exhibit patterns such as attending to delimiter tokens, specific positional offsets, or broadly attending over the sentence with heads in the same layer often exhibiting similar behaviors. The analysis focuses on 144 attention heads in BERT.

The researchers experimented by treating each head as a simple no-training required classifier which takes a word as input and outputs the most attended to other word. This way they evaluate the ability of heads to classify various syntactic relations. They found particular heads that corresponds well to particular relations such as objects of verbs, determiners of nouns, objects of verbs of nouns, objects of verbs, determiners of nouns prepositions, and objects of possessive pronouns. The heads were able to find these with accuracy of more than 75%.

The researchers propose an attention-based probing classifier that takes attention maps as input. They use the "base" size BERT model, which has 12 layers containing 12 attention heads each. In BERT most of the heads attend to special tokens. Most heads attend on the next or previous token instead of the current token. Over half of BERT's attention in layers 6-10 focuses on [SEP] token.

They use gradient-based measures of feature importance for further investigation.

Attending more or less to [SEP] does not substantially change BERT's output, so Attention over special tokens might be used as a sort of "no-op" when the attention head’s function is not applicable.

The researchers measure weather the attention heads focuses on a few words or attend broadly over many words by computing average entropy of each head’s attention distribution. The high entropy attention heads especially on lower layer typically spend at most 10% of their attention mass on any single word.

They also investigate individual attention heads to see what aspects of language they have learned.

Sine BERT uses byte-pair tokenization, they convert token-token attention maps to word-word attention maps. For attention to a split-up word, they sum up attention weights over its tokens. For attention from a split-up word, they take the mean of attention weights over its token.

Heads can disagree with standard annotation conventions while still performing syntactic behavior. It shows that the syntactic behavior in BERT is learned as by-product of self supervised training.

They also analyze the coreference resolution by computing percent of time the head word of a co-referent mention most attend to the head of one of that mention's antecedents.

They find that one of BERT's attention heads achieve decent performance in this experiment and is good with nominal mentions.

They also measure the models overall knowledge through a novel family of attention based probing classifiers and applying then to dependency parsing. Attention-Only Probe which learns a simple linear combination of attention weights.

and Attention and word probe which sets the weights of the attention heads based on the GloVe embeddings for the input words. They found that Attention with GloVe probing classifier substantially outperforms the baselines. This tells that BERT's attention maps have fairly thorough representation of English syntax.

They also analyze the clustering of the attention heads. They find that several clear clusters of heads that behave similarly. This may be because of attention dropout.

This paper shows that there is substantial amount of linguistic knowledge found in the attention heads.