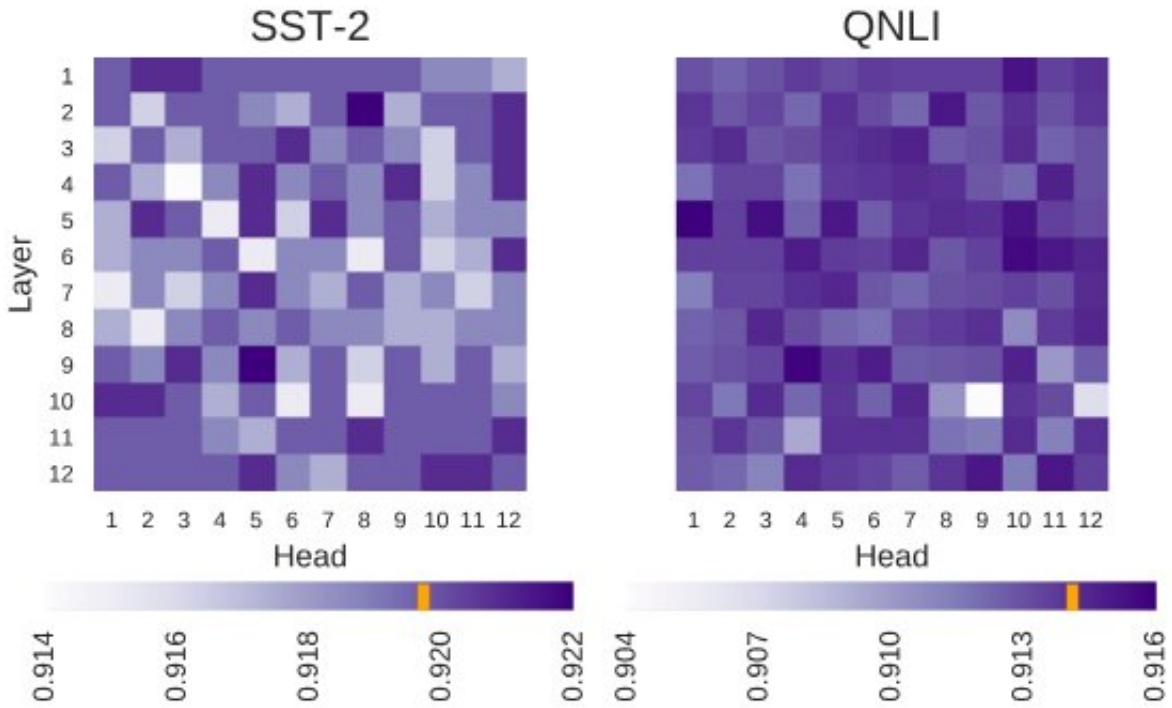
Investigating the self-attention mechanism behind BERT-based architectures

September 11 2019, by Ingrid Fadelli



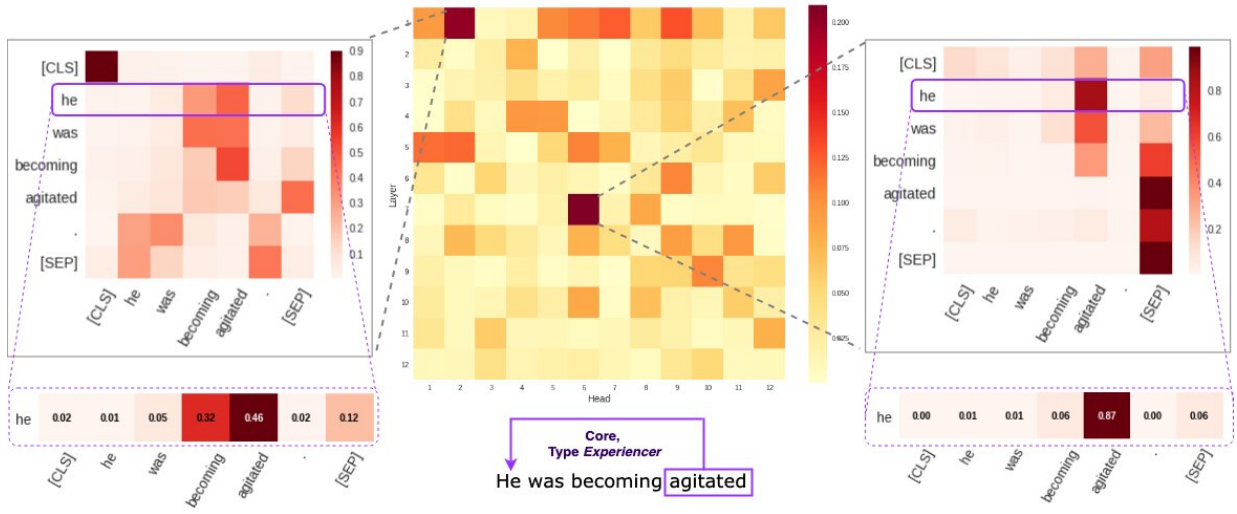
Investigated BERT architecture has the architecture of 12 layers by 12 heads. Each cell in this figure shows the performance of BERT if the corresponding head is switched off. Darker colors indicate higher performance, and white cells indicate heads without which BERT's performance decreases. Stanford Sentiment Treebank (SST-2): There are multiple heads encoding information that is necessary for the task. Question Natural Language Inference (QNLI): Most heads improve the overall performance when they are switched off. Credit: Kovaleva et al.

BERT, a transformer-based model characterized by a unique self-attention mechanism, has so far proved to be a valid alternative to recurrent neural networks (RNNs) in tackling natural language processing (NLP) tasks. Despite their advantages, so far, very few researchers have studied these BERT-based architectures in depth, or tried to understand the reasons behind the effectiveness of their self-attention mechanism.

Aware of this gap in the literature, researchers at the University of Massachusetts Lowell's Text Machine Lab for Natural Language Processing have recently carried out a study investigating the interpretation of self-attention, the most vital component of BERT models. The lead investigator and senior author for this study were Olga Kovaleva and Anna Rumshisky, respectively. Their paper pre-published on arXiv and set to be presented at the EMNLP 2019 conference, suggests that a limited amount of attention patterns are repeated across different BERT sub-components, hinting to their over-parameterization.

"BERT is a recent model that made a breakthrough in the NLP community, taking over the leaderboards across multiple tasks. Inspired by this recent trend, we were curious to investigate how and why it works," the team of researchers told TechXplore via email. "We hoped to find a correlation between self-attention, the BERT's main underlying mechanism, and linguistically interpretable relations within the given input text."

BERT-based architectures have a layer structure, and each of its layers consists of so called "heads." For the model to function, each of these heads is trained to encode a specific type of information, thus contributing to the overall model in its own way. In their study, the researchers analyzed the information encoded by these individual heads, focusing on both its quantity and quality.



Each cell in the middle figure reflects how individual heads pay attention to core semantic links within a given sentence (on average). We identified two specific heads that tend to encode semantic information more than the others. The two images on the sides demonstrate how these two heads assign weights to individual words within a random sentence of our dataset. Credit: Kovaleva et al.

"Our methodology focused on examining individual heads and the patterns of attention they produced," the researchers explained. "Essentially, we were trying to answer the question: "When BERT encodes a single word of a sentence, does it pay attention to the other words in a way meaningful to humans?"

The researchers carried out a series of experiments using both basic pretrained and fine-tuned BERT models. This allowed them to gather numerous interesting observations related to the self-attention mechanism that lies at the core of BERT-based architectures. For instance, they observed that a limited set of attention patterns is often repeated across different heads, which suggests that BERT models are

over-parameterized.

"We found that BERT tends to be over-parameterized, and there is a lot of redundancy in the information it encodes," the researchers said. "This means that the computational footprint of training such a large model is not well justified."

A further interesting finding gathered by the team of researchers at the University of Massachusetts Lowell is that depending on the task tackled by a BERT model, randomly switching off some of its heads can lead to an improvement, rather than a decline, in performance. In addition, the researchers did not identify any linguistic patterns that are of particular importance in determining BERT's performance in downstream tasks.

"Making deep learning interpretable is important for both fundamental and applied research, and we will continue working in this direction," the researchers said. "New BERT-based models have recently been released, and we plan to extend our methodology to investigate them as well."

More information: Revealing the dark secrets of BERT. arXiv:1908.08593 [cs.CL]. arxiv.org/abs/1908.08593