Interpreting Multi-Head Attention in Pre-Trained Language Representations like BERT

A recent trend that has boosted performance in many NLP areas is pre-training general language representations on huge data sets and fine-tuning them on a specific task like question answering, sentiment analysis or semantic similarity scoring. One of the most popular and successful of these models is the autoencoder language model BERT that is based on Transformer encoder blocks with multi-head attention. Being trained on a simple word prediction task, it seems to learn more useful information than word embeddings like word2vec or GloVe.

Multi-head attention has the nice property that it can offer insights into what the models are learning and how they are modelling the language by testing what its attention heads are attending to. There have been various papers that show that there are attention heads in BERT that specialize on certain tasks, and that heads on different levels of the network build a hierarchy: The heads on the lowest levels learn word-level features, in the middle thee heads mostly learn syntactic features, and on the highest level the heads learn semantic features that can cover larger text spans.

A possible project could be to analyze these papers, select convincing interpretation techniques and apply them to the multilingual BERT model that can be fine-tuned on Swedish or any language(s) other than English that you understand well enough. Further probing experiments based on literature or own ideas can also be added.
Other possibilities are comparing BERT to competing models like XLNet, or studying if/how fine-tuning affects the functions of attention heads.

Related Papers:

Transformers: https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf
BERT: https://arxiv.org/pdf/1810.04805.pdf
XLNet: https://arxiv.org/pdf/1906.08237.pdf

Interpretation Studies:
Clark et al. 2019: https://nlp.stanford.edu/pubs/clark2019what.pdf
Jawahar et al. 2019 https://www.aclweb.org/anthology/P19-1356.pdf

Student profile

A master student in computer science, cognitive science or statistics with experience in machine learning and natural language processing.

Contact

Jenny Kunz, jenny.kunz@liu.se