Using the Transformer

BERT – Implications for future work

Learning goals

- Understand how impactful this architecture was
- See how this changed research in the field

ENGLISH CENTRICITY OF NLP

- BERT trained on a corpus of English text
- More importantly: Also only evaluated on English benchmarks (obviously)
  ▸ GLUE ▸ SQUAD ▸ RACE
- Devlin et al. (2019) published different (monolingual) models, but only varying in size, not in language
- Later: Multilingual BERT model ▸ mBERT for 100+ languages
- This leads to a shared embedding space for all the languages included in the model
- Before this: Need for alignment of separately learned embedding spaces

BERTS FOR ALL LANGUAGES

- The breakthrough performance of BERT in the English Language triggered a wave of new BERT models in different languages. Just to name a few:
  - German BERT
  - FlauBERT (French)
  - BETO (Spanish)
  - BERTje (Dutch)
  - Chinese BERT
  - RuBERT (Russian)
  - Italian BERT
  - ...

PRETRAIN-FINETUNE + TRANSFORMER BACKBONE

- Before BERT:
  - ELMo (and other specialized architectures) very popular
  - Examples (also CNNs): ▸ Kim, 2014 ▸ Zhang et al., 2016
- After BERT:
  - Using a pre-trained model and fine-tuning it is the de-facto standard
  - CNNs and RNNs rarely used, different variants of the transformer or other self-attention based mechanisms are the backbone of nearly every architecture

BERTOLOGY ▸ RODGERS ET AL., 2020

Post-BERT architectures:

- Most architectures still rely on either an encoder- *or* a decoder-style type of model (e.g. ▸ GPT2, ▸ XLNet)
- *BERTology:* Many papers/models which aim at ..
  - .. explanining BERT (e.g. ▸ Coenen et al., 2019, ▸ Michel et al., 2019)
  - .. improving BERT (▸ RoBERTa, ▸ ALBERT)
  - .. making BERT more efficient (▸ ALBERT, ▸ DistilBERT)
  - .. modifying BERT (▸ BART)
- Overview on many different papers: https://github.com/tomohideshibata/BERT-related-papers

BERTOLOGY – EXAMPLE

Examining/Interpreting Attention patterns:

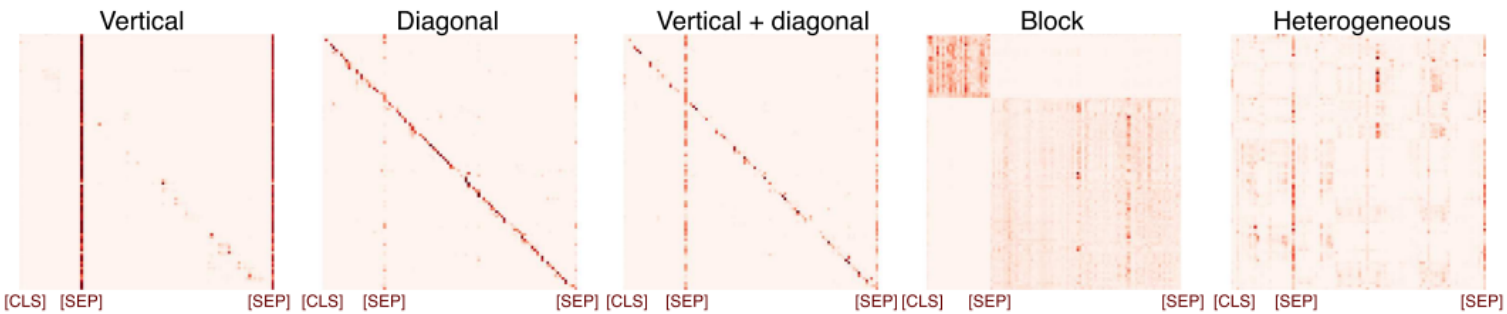

Figure 3: Attention patterns in BERT (Kovaleva et al., 2019).

- Attempt to "understand" what the model has learned
- Still relevant today when seeking interpretability