2.4 BERT

BERT stands for Bidirectional Encoder Representations from Transformers and is about self-supervised learning in NLP.

One major limitation of standard language models is that they are unidirectional. OpenAI GPT, for example, uses a left-to-right architecture where can every token can only attend to previous tokens in the self-attention layers of the Transformer. BERT, on the other hand, takes a bidirectional approach using the following two pre-training tasks:

- **Masked language model (MLM).** Some input tokens are masked at random and the objective is to predict the original vocabulary id of those tokens.
- **Next sentence prediction (NSP).** To train a model that understands sentence relationships, given sentences `A` and `B`, we predict whether `B` follows `A` in the original corpus.

BERT jointly trains on these two tasks.