Advanced Natural Language Processing

1 LLMs

BERT

Bidirectional Encoder Representations from Transformers.
Basically a transformer encoder.
Training: BERT was trained on Books and Wikipedia.
Training objectives: Masked Language Modelling. Predict randomly masked tokens (by [MASK] token). Next Sentence Prediction (with [CLS] token).
Versions:
- BERT-Base: 12 transformer blocks, embedding dim: 768, 12 attention heads.
- BERT-Large: 24 transformer blocks, embedding dim: 1024, 16 attention heads.
Extensions:
- RoBERTa: Larger batches, removed NSP.
- SentenceBERT: BERT + siamese architecture (Sentence A and B in the same model, outputs are compared by some metric) + triplet loss (tune network such that distance between anchor sentence and positive sentence is smaller than anchor sentence and negative sentence)
- DistillBERT: Larger teacher network creates soft probability distribution labels. Smaller student model tries to replicate these distributions.

GPT

Generative Pretrained Transformer.
Only a transformer decoder, since its task is not seq2seq, where you might want to refer to another part of the input sentence, but instead just generate next tokens (still the decoder receives already generated tokens as inputs).
Goal is to use few-shot learning and prompting instead of fine-tuning and task specific architectures.
Autoregressive: Feeds back generated tokens in decoder input.

Other LLMs

T5

Text-to-Text-Transfer-Transformer.

Encoder and decoder block.