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, similar to original Attention is all you need paper but with slight modifications.

RETRO

Retrieval-Enhanced Transformer.
Enhanced by large text database.
In total 4% of GPT-3 size.

Flash Attention

Self-attention is the bottleneck when training/inferencing transformers (quadratic time). During Attention, tensors are moved from HBM (Hidh Bandwith (GPU) Memory) to SRAM (Shared random-access (GPU) memory) multiple times: Before each operation, the tensors are moved from HBM to SRAM and then written back to HBM.
Flash Attention fuses multiple operations into one kernel, thus saving load and write operations. It is becoming the standard for transformers.

Fine-Tuning

Traditional fine-tuning: Re-train all parameters with a small labelled dataset for example on domain knowledge.
Traditional fine-tuning with freezing: Freeze a part of the model.

Prompting

Removes the need for additional layers for a downstream task. Uses the same format as the pretraining objective of LLMs: We ask the LLM to generate the answer based on a task-specific input.
Requires no new parameter, can be practiced on a closed-source LLM.

Zero-Shot

Simple prompting is equivalent to zero-shot learning. The model predicts the answer based on the question.
Example: "Translate English to French: cheese = "