Bidirectional Encoder Representations from Transformers (BERT)

Presentation by Connor Silbiger

Presentation Overview

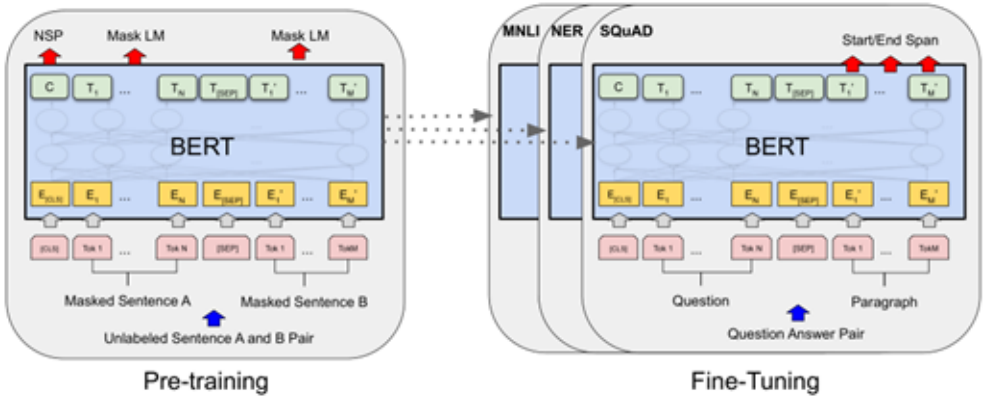
- Part 1: Summary of the BERT paper
- Part 2: Strengths at release and strengths that still stand
- Part 3: Weaknesses at release and weaknesses in hindsight
- Part 4: Extensions and uses beyond the paper

Part 1: Paper Summary

- Older NLP systems often needed a custom model for each task
- ELMo used context, but added features to task specific models
- Original GPT used one Transformer, but read only left to right
- BERT aimed to read both directions and adapt to many tasks

Part 1: How BERT Works

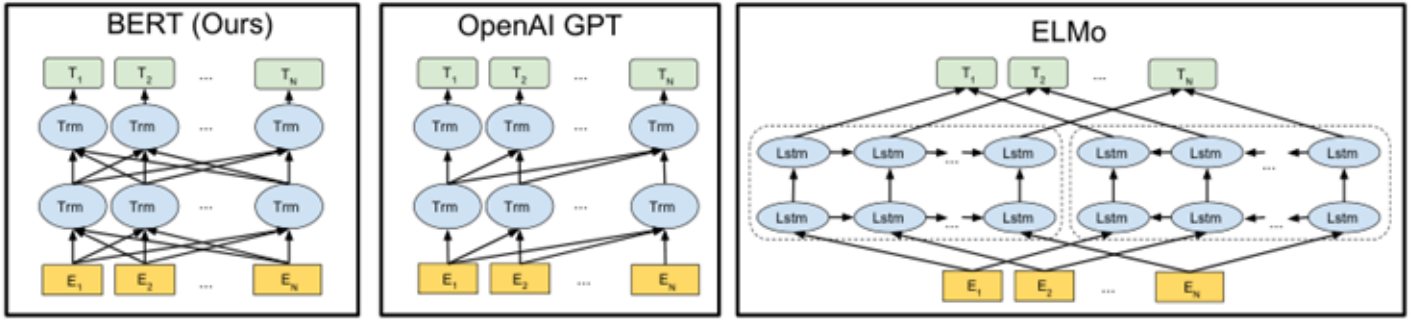
- Transformer encoder
- Masks 15% of tokens and predicts them
- Also predicts whether one text follows another
- Fine tunes the same model for each task



Part 1: Main Results

- State of the art results on 11 NLP tasks
- GLUE: 80.5, up 7.7 points
- MultiNLI: 86.7% accuracy, up 4.6 points
- SQuAD 2.0: 83.1 F1, up 5.1 points
- SQuAD ablation: 88.5 with BERT vs. 77.8 left to right

Part 2, Section 1: Strengths at Release



- ELMo: separate directions and task specific models
- GPT: Transformer fine tuning, but left to right
- BERT: both directions at every layer and simple task outputs

Part 2, Section 2: Strengths That Still Stand

- Strong fit for classification and token labeling
- Useful for extractive question answering and reranking
- Can be smaller and cheaper than large generative models
- Uses the full input at once
- Does not generate free text when a label or score is enough

Part 3, Section 1: Weaknesses at Release

- Pretraining used 16 or 64 TPU chips for more than four days
- Input length was limited to 512 tokens
- Attention cost grows quickly as text gets longer
- [MASK] appears in training but not in normal input
- Fine tuning could be unstable on small datasets
- BERT does not naturally write long answers

Part 3, Section 2: Weaknesses in Hindsight

- RoBERTa found that BERT was undertrained
- It trained longer, with bigger batches and more data
- It changed the mask each time an example was used
- It removed next sentence prediction
- RoBERTa reached 88.5 on GLUE
- Later decoder models were better suited to text generation

Part 4: How BERT Was Extended

- Sentence-BERT: makes sentence vectors for fast search and matching
- BioBERT: trains further on medical papers to understand medical terms
- DistilBERT: a smaller version made for faster and cheaper use
- The same idea can be adapted to other fields and languages

Part 4: What BERT Can Be Used For

- Classify sentiment, spam, topics, or customer requests
- Find names, dates, places, medical terms, and other details
- Answer questions by selecting a passage from a document
- Search by meaning instead of only matching exact words
- Rank search results and retrieve passages for another model
- Group similar documents and find duplicates

Main Takeaway

- BERT made deep bidirectional pre training practical
- One encoder could be adapted to many language tasks
- Its original training recipe was quickly improved
- BERT style encoders still work well for focused tasks

Sources

- Devlin et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. aclanthology.org/N19-1423/
- Peters et al. 2018. Deep Contextualized Word Representations. aclanthology.org/N18-1202/
- Liu et al. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arxiv.org/abs/1907.11692
- Reimers and Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. aclanthology.org/D19-1410/
- Lee et al. 2020. BioBERT: A Pre-trained Biomedical Language Representation Model. doi.org/10.1093/bioinformatics/btz682
- Sanh et al. 2019. DistilBERT: Smaller, Faster, Cheaper and Lighter. arxiv.org/abs/1910.01108