TECHNISCHE UNIVERSITÄT WIEN

Informatics

Transformers: BERT

VU Research topics in natural language processing (194.135)

1. Paper Presentation - WS2023

Seminar Group

Carlos Vargas R.

December 6, 2023

Outline

- Introduction to Transformers (NLP)
- Related Work

1. BERT
   - What is BERT?
   - Model Architecture
   - Pre-Training Tasks & Procedure + Fine-Tuning
2. Comparison of BERT, ELMo & OpenAI GPT
3. References

Introduction to Transformers

Attention is All you Need!

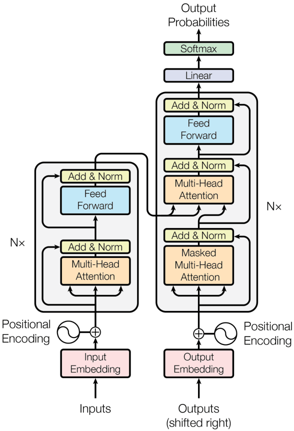


Figure: The Transformer - model architecture (Vaswani et al. 2017)

Related Work

The state-of-the-art (SOA) on NLP tasks.

- **Unsupervised Feature-based Approaches:** Pre-trained word embeddings are an integral part of modern NLP systems.
- ELMo (2018): **traditional word embedding** (multidimensional) and context-sensitive features from left-to-right and right-to-left language model.
- **Unsupervised Fine-tuning Approaches:** Fine-tuned for a supervised downstream task.
- OpenAI GPT (2018): **sentence-level tasks** SOA on GLUE
- **Transfer Learning from Supervised Data** (Pirge and Follow 2019)

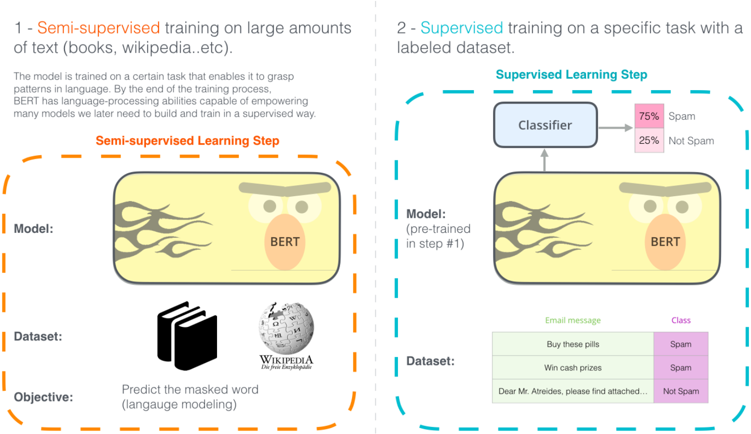
BERT

Bidirectional Encoder Representations from Transformers

- **Why is BERT so important?**
- Broke several bechmarks on NLP tasks.
- Open-source.
- It is available to download (on already pre-trained datasets).
- **What is BERT?**
  BERT is basically a trained Transformer Encoder stack.
- B.base (L=12, H=768, A=12, Total Parameters=110M)
- B.large (L=24, H=1024, A=16, Total Parameters=340M)
  Layers (L), Hidden size (H) and self-attention heads (A)
- For the **Pre-training**: Understand Language
  BooksCorpus (800M words) & EN Wikipedia (2500M words)
  1. Masked LM (MLM) & 2. Next Sentence Prediction (NSP)
- **Fine-tuning:** Specific NLP tasks

Pre-Training & Fine-tuning

Understanding Language and Using it for specific NLP tasks.



1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

2 - Supervised training on a specific task with a labeled dataset.

Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

Figure: Pre-Training & Fine-tuning - BERT (*Fine-Tuning BERT for text classification with LoRA Karkar Nizar · Follow* 2019)

Pre-Training & Fine-tuning - 2

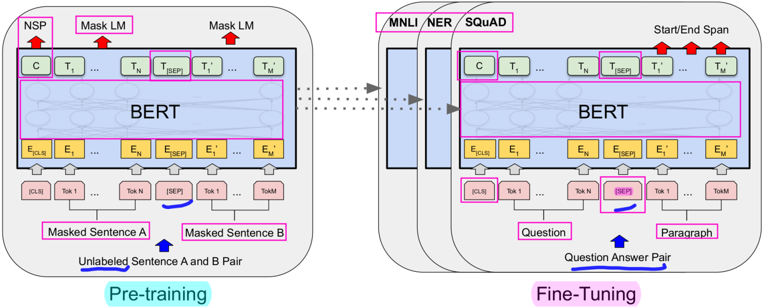


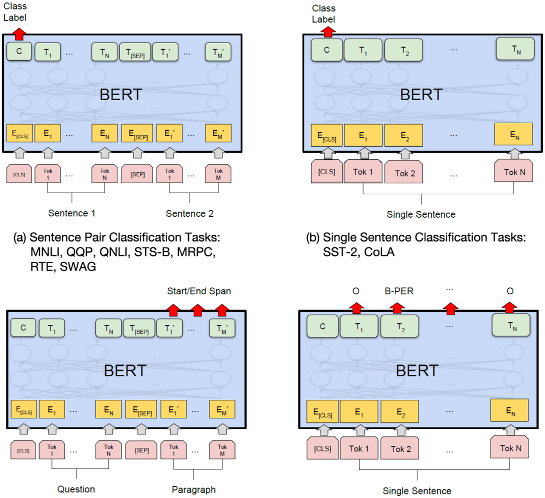
Figure: Pre-Training & Fine-tuning - Layers (Devlin et al. 2019)

Evaluation

Evaluation with specific among NLP datasets (Devlin et al. 2019)

- (MNLI) Multi-Genre Natural Language Inference:
  Given a pair of sentences then, predict the second sentence.
- (QQP) Quora Question Pairs:
  Given two questions then, semantically equivalent
- (QNLI) Question Natural Language Inference:
  Determine the correct answer
- (SST-2) The Stanford Sentiment Treebank
- (CoLA) The Corpus of Linguistic Acceptability
- (STS-B) The Semantic Textual Similarity Benchmark
- (MRPC) Microsoft Research Paraphrase Corpus: new sources
- (RTE) Recognizing Textual Entailment: Similar to MNLI
- (WNLI) Winograd NLI

Fine-tuning - 2



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Figure: Fine-tuning - Tasks (Devlin et al. 2019)

Evaluation

Evaluation with specific among NLP datasets

- **GLUE** General Language Understanding Evaluation
- **SQuAD v1.1**: Stanford Question Answering Dataset
- **SQuAD v2.0**: Extended v1.1 with more realistic answers
- **SWAG**: Situations With Adversarial Generations
  113k sentences-pair completion to evaluate common sense.

Comparison among the SOA Models

- GPT is trained on the BooksCorpus (800M words)
- BERT is trained on the BooksCorpus (800M words) and Wikipedia (2,500M words).
- GPT uses a sentence separator ([SEP]) and classifier token ([CLS]) which are only introduced at fine-tuning time
- BERT learns [SEP], [CLS] and sentence A/B embeddings during pre-training.
- GPT was trained for 1M steps with a batch size of 32,000 words;
- BERT was trained for 1M steps with a batch size of 128,000 words.
- GPT used the same learning rate of 5e-5 for all fine-tuning experiments
- BERT chooses a task-specific fine-tuning learning rate which performs the best on the development set (Rogers, Kovaleva, and Rumshisky 2019).

Comparison among the SOA Models

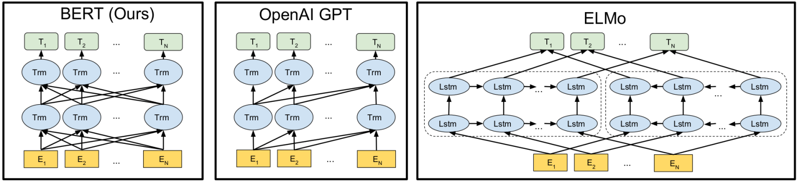


Figure: Comparison among the SOA (Rogers, Kovaleva, and Rumshisky 2019)

References

Vaswani, Ashish et al. (June 2017). “Attention Is All You Need”. In: URL: http://arxiv.org/abs/1706.03762.

Devlin, Jacob et al. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: URL: https://github.com/tensorflow/tensor2tensor.

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky (2019). “A Primer in BERTology: What We Know About How BERT Works”. In: DOI: 10.1162/tacl. URL: https://doi.org/10.1162/tacl.