Machine Learning (CE 40717)

Fall 2024

Ali Sharifi-Zarchi

CE Department
Sharif University of Technology

December 14, 2024

1. Encoder Architecture

1. Encoder Architecture

Introduction to Language Modeling

Language Modeling:

- Language modeling involves predicting the probability of a sequence of words.
- Given a sequence $x = \{x_1, x_2, \ldots, x_n\}$, the probability of the entire sequence can be decomposed into the product of conditional probabilities of each word, given the context.

Mathematical Representation:

$$P(x) = \prod_{i=1}^{n} P(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$

- $P(x)$: The probability of the entire sequence x.
- Each word x_i depends on all other words in the sequence, including its left and right context.
- This approach captures the dependencies between words, which is essential for understanding language semantics.

Encoder Language Model

Encoder language models, like BERT, use masked tokens to learn bidirectional representations of text.

- **Masked Language Modeling (MLM):** Predicts randomly masked tokens in a sequence.
- **Bidirectional Context:** Considers information from both directions for each token.
- **Applications:** Used for classification, NER, and other NLP tasks.

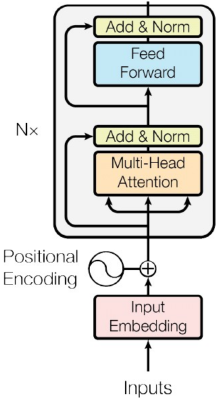
BERT: Key Contributions

- It is a fine-tuning approach based on a deep Transformer Encoder.
- The key: learn representations based on **bidirectional context**

  Why? Because both left and right contexts are important to understand the meaning of words.

  Example #1: we went to the river bank.
  Example #2: I need to go to bank to make a deposit.
- **Pre-training objectives:** masked language modeling + next sentence prediction
- State-of-the-art performance on a large set of **sentence-level** and **token-level** tasks

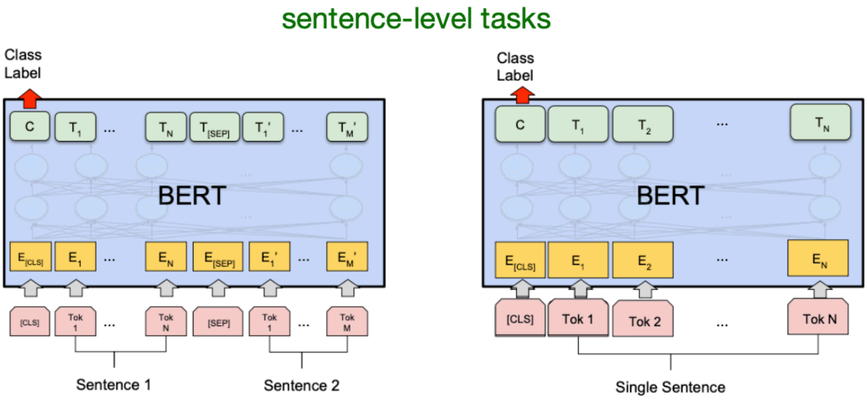
BERT Models



- **BERT-Base:** 12 layers, 768 hidden size, 12 attention heads, 110M parameters
- **BERT-Large:** 24 layers, 1024 hidden size, 16 attention heads, 340M parameters
- **Training corpus:** Wikipedia (2.5B) + BooksCorpus (0.8B)
- **Max sequence size:** 512 word pieces (roughly 256 and 256 for two non-contiguous sequences)
- **Trained for:** 1M steps, batch size 128k

https://arxiv.org/abs/1706.03762

Sentence-level tasks



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

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

https://arxiv.org/abs/1810.04805

Sentence-level tasks(cont.)

- Sentence pair classification tasks:

  MNLI

  - **Premise:** A soccer game with multiple males playing.
  - **Hypothesis:** Some men are playing a sport.
  - Result: {entailment, contradiction, neutral}

  QQP

  - Q1: Where can I learn to invest in stocks?
  - Q2: How can I learn more about stocks?
  - Result: {duplicate, not duplicate}

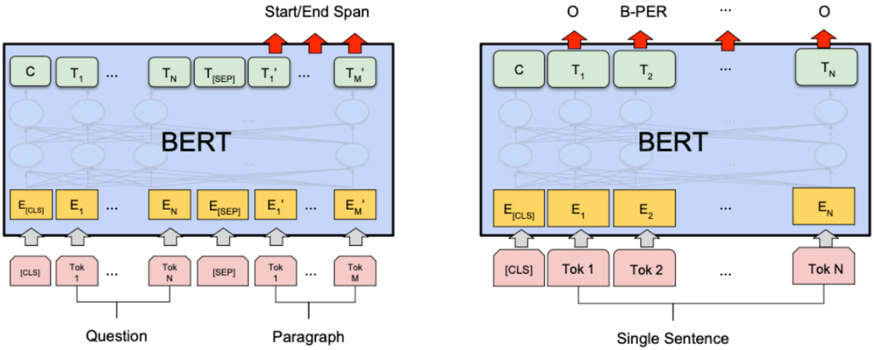
- Single sentence classification tasks:

  SST2

  - Sentence: rich veins of funny stuff in this movie
  - Result: {positive, negative}

Token-level tasks

token-level tasks



(c) Question Answering Tasks: SQuAD v1.1

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

https://arxiv.org/abs/1810.04805

Token-level tasks: Extractive Question Answering

- Extractive question answering e.g., SQuAD (Rajpurkar et al., 2016)

SQuAD

Question: The New York Giants and the New York Jets play at which stadium in NYC ?

Context: The city is represented in the National Football League by the New York Giants and the New York Jets , although both teams play their home games at MetLife Stadium in nearby East Rutherford , New Jersey , which hosted Super Bowl XLVIII in 2014 .

(Training example 29,883)

Example Result: MetLife Stadium

Token-level tasks: Named Entity Recognition

Token-level tasks

- Named entity recognition (Tjong Kim Sang and De Meulder, 2003)

CoNLL 2003 NER

John	Smith	lives	in	New	York
B-PER	I-PER	O	O	B-LOC	I-LOC

Masked Language Modeling (MLM)

- **Q:** Why we can't do language modeling with bidirectional models?

- **Solution:** Mask out a percentage k of the input words, and then predict the masked words.

$$\text{the man went to} \quad [\underset{\downarrow}{\overset{\textbf{store}}{}}\textit{MASK}] \quad \text{to buy a} \quad [\overset{\textbf{gallon}}{}\textit{MASK}] \text{ of milk}$$

MLM: Masking Rate and Strategy

- **Q: What is the value of k?**
  - They always use $k = 15\%$.
  - Too little masking: computationally expensive (we need to increase # of epochs)
  - Too much masking: not enough context
  - See (Wettig et al., 2022) for more discussion of masking rates:
    - Masking 40% outperforms 15% for BERT-large size models on GLUE and SQuAD
    - High masking rate of 80% can still preserve 95% fine-tuning performance

- **Q: How are masked tokens selected?**
  - 15% tokens are uniformly sampled
  - Is it optimal? See span masking (Joshi et al., 2020) and PMI masking (Levine et al., 2021)

Example: He **[MASK]** from Kuala **[MASK]**, Malaysia.

Next Sentence Prediction (NSP)

- Motivation: many NLP downstream tasks require understanding the relationship between two sentences (natural language inference, paraphrase detection, QA).
- NSP is designed to reduce the gap between pre-training and fine-tuning.

[CLS]: a special token always at the beginning

[SEP]: a special token used to separate two segments

Input = `[CLS] the man went to [MASK] store [SEP]`
`he bought a gallon [MASK] milk [SEP]`

Label = `IsNext`

Input = `[CLS] the man [MASK] to the store [SEP]`
`penguin [MASK] are flight ##less birds [SEP]`

Label = `NotNext`

They sample two contiguous segments for 50% of the time and another random segment from the corpus for 50% of the time

BERT Training

Dataset: Let $\mathscr{D}$ be a set of examples $(x_{1:L}, c)$ constructed as follows:

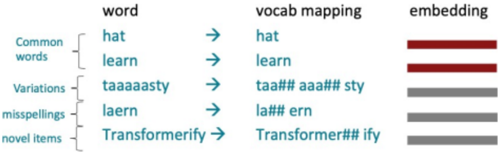
- Let A be a sentence from the corpus.
- With probability 0.5, let B be the next sentence.
- With probability 0.5, let B be a random sentence from the corpus.
- Let $x_{1:L} = [\text{CLS}], A, [\text{SEP}], B$.
- Let c denote whether B is the next sentence or not.

Objective. Then the BERT objective is:

$$\mathcal{O}(\theta) = \sum_{(x_{1:L},c)\in D} \underbrace{\mathbb{E}_{I,\tilde{x}_{1:L}\sim A(\cdot|x_{1:L},I)}\left[\sum_{i\in I} -\log p_\theta(\tilde{x}_i \mid x_{1:L})\right]}_{\text{masked language modeling}} + \underbrace{-\log p(c \mid \phi(x_{1:L})_1)}_{\text{next sentence prediction}}.$$

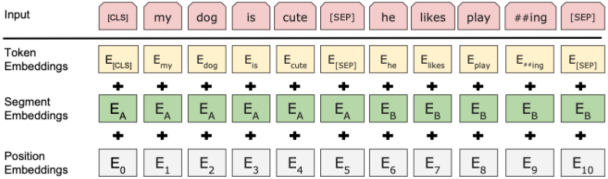
BERT Pre-training: Putting Together

- **Vocabulary size:** 30,000 wordpieces (common sub-word units) (Wu et al., 2016)



(Image: Stanford CS224N)

- **Input embeddings:**



Separate two segments

- Just two possible "segment embeddings": EA and EB.
- Positional embeddings are learned vectors for every possible position between 0 and 512-1.

https://cs330.stanford.edu/fall2019/presentations/presentation-10.23-1.pdf

BERT Pre-training: Putting Together

- MLM and NSP are trained together
- `[CLS]` is pre-trained for NSP
- Other token representations are trained for MLM

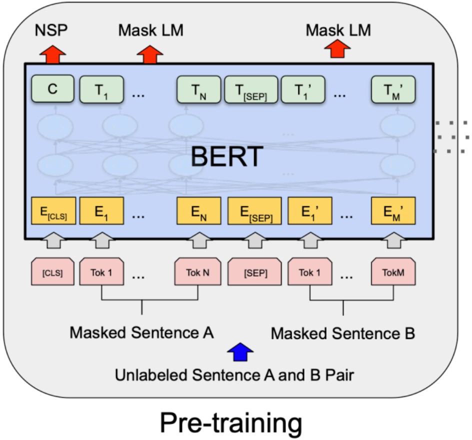


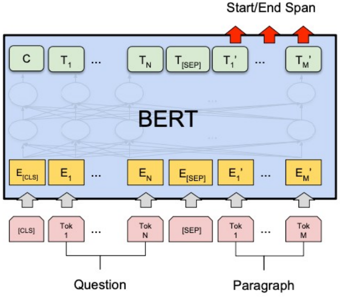
Figure 1: *

https://arxiv.org/abs/1810.04805

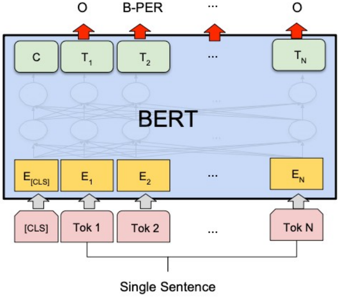
Fine-tuning BERT

“Pre-train once, finetune many times.”

token-level tasks



(c) Question Answering Tasks: SQuAD v1.1



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

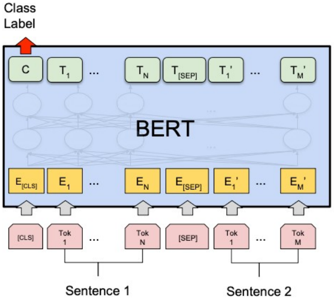
For token-level prediction tasks, add linear classifier on top of hidden representations

Q: How many new parameters?

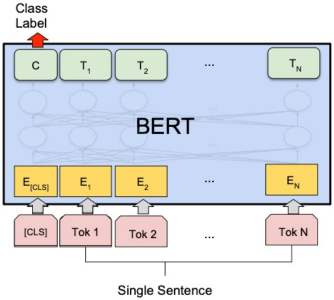
Fine-tuning BERT

"Pre-train once, finetune many times."

sentence-level tasks



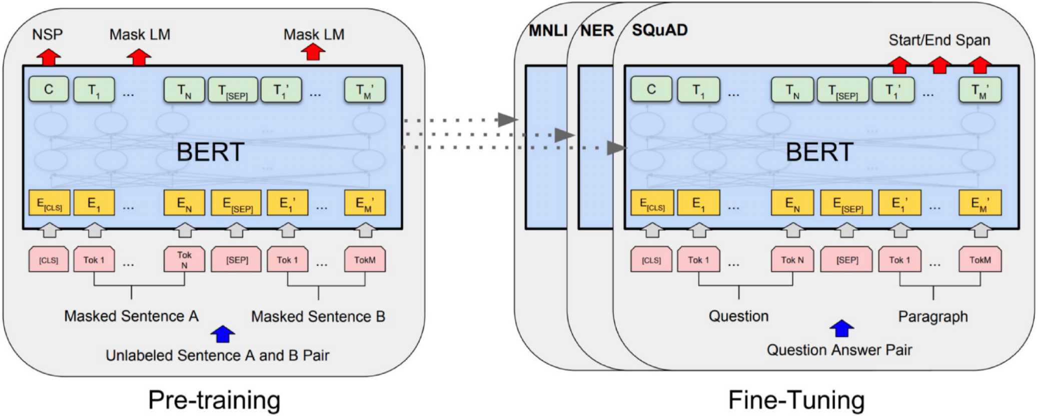
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



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

For sentence pair tasks, use [SEP] to separate the two segments with segment embeddings and add a linear classifier on top of [CLS] representation.

Finetuning Paradigm in NLP



BERT Extensions

- Models that handle long contexts (> 512 tokens)
  - Longformer, Big Bird, ...
- Multilingual BERT
  - Trained single model on 104 languages from Wikipedia. Shared 110k WordPiece vocabulary
- BERT extended to different domains
  - SciBERT, BioBERT, FinBERT, ClinicalBERT, ...
- Making BERT smaller to use
  - DistillBERT, TinyBERT, ...

BERT Extensions

- **RoBERTa** (Liu et al., 2019)
  - Trained on 10x data & longer, no NSP
  - Much stronger performance than BERT (e.g., 94.6 compared to 90.9 on SQuAD)
  - Still one of the most popular models to date
- **ALBERT** (Lan et al., 2020)
  - Increasing model sizes by sharing model parameters across layers
  - Less storage, much stronger performance but runs slower