BERT

(**B**idirectional **E**ncoder **R**epresentations from **T**ransformers)

Content By:

Jacob Devlin

Presenter:

Mohammad Amin Abbasi

Mohammad Amin Abbasi
NLP Engineer

Email:
m_abbasi1378@comp.iust.ac.ir

Education
4th Semester MSc. Software Engineering

Role
Technical Lead at Native LLM Development at the National Center for AI Navigation

Academic Experience
- PersianLLaMA: Towards Building First Persian Large Language Model
- AriaBERT: A Pre-trained Persian BERT Model for Natural Language Understanding

Pre-training in NLP

- Word embeddings are the basis of deep learning for NLP

king ↓ [-0.5, -0.9, 1.4, …]

queen ↓ [-0.6, -0.8, -0.2, …]

- Word embeddings (`word2vec`, `GloVe`) are often *pre-trained* on text corpus from co-occurrence statistics

Inner Product

the king wore a crown

Inner Product

the queen wore a crown

Contextual Representations

- **Problem**: Word embeddings are applied in a context free manner

```
open a bank account          on the river bank

          [0.3, 0.2, -0.8, …]
```

- **Solution**: Train *contextual* representations on text corpus

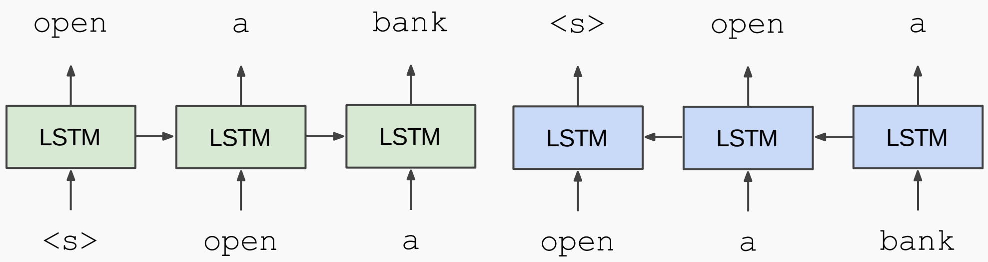
```
[0.9, -0.2, 1.6, …]          [-1.9, -0.4, 0.1, …]

open a bank account          on the river bank
```

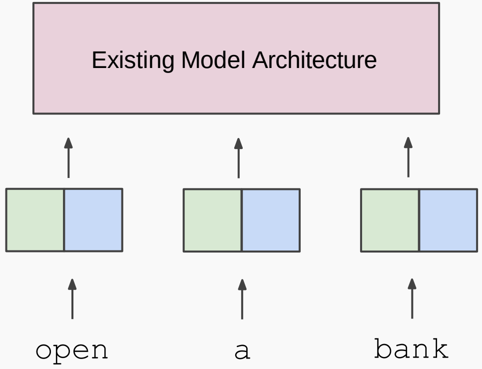
History of Contextual Representations

- *ELMo: Deep Contextual Word Embeddings*, AI2 & University of Washington, 2017

Train Separate Left-to-Right and Right-to-Left LMs



Apply as "Pre-trained Embeddings"



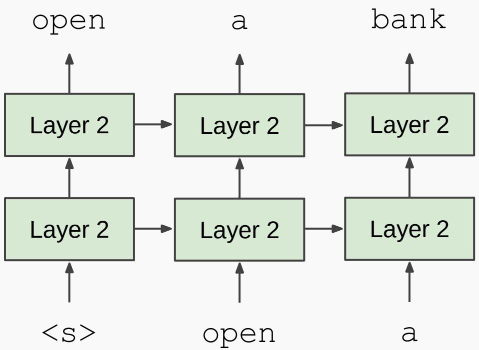
Problem with Previous Methods

- **Problem**: Language models only use left context *or* right context, but language understanding is bidirectional.
- Why are LMs unidirectional?
- <u>Reason 1</u>: Directionality is needed to generate a well-formed probability distribution.
  - We don't care about this.
- <u>Reason 2</u>: Words can "see themselves" in a bidirectional encoder.

Unidirectional vs. Bidirectional Models

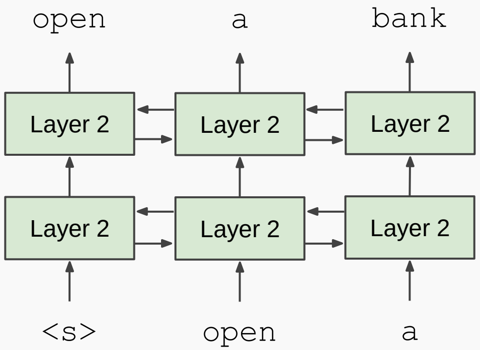
Unidirectional context

Build representation incrementally



Bidirectional context

Words can “see themselves”



Masked LM

- **Solution**: Mask out *k*% of the input words, and then predict the masked words
  - We always use *k* =15%

```
                        store             gallon
                          ↑                 ↑
the man went to the [MASK] to buy a [MASK] of milk
```

- Too little masking: Too expensive to train
- Too much masking: Not enough context

Masked LM

- Problem: Mask token never seen at fine-tuning
- Solution: 15% of the words to predict, but don't replace with `[MASK]` 100% of the time. Instead:
- 80% of the time, replace with `[MASK]`

  `went to the store → went to the [MASK]`
- 10% of the time, replace random word

  `went to the store → went to the running`
- 10% of the time, keep same

  `went to the store → went to the store`

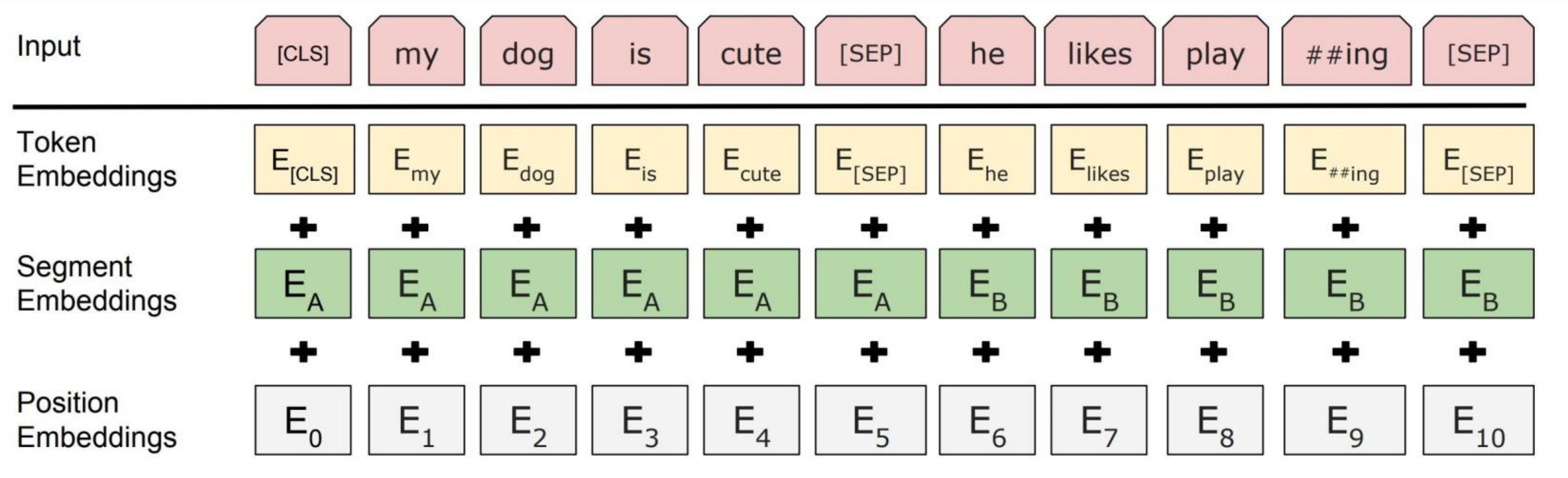
Next Sentence Prediction

- To learn *relationships* between sentences, predict whether Sentence B is actual sentence that proceeds Sentence A, or a random sentence

Sentence A = `The man went to the store.`
Sentence B = `He bought a gallon of milk.`
Label = `IsNextSentence`

Sentence A = `The man went to the store.`
Sentence B = `Penguins are flightless.`
Label = `NotNextSentence`

Input Representation

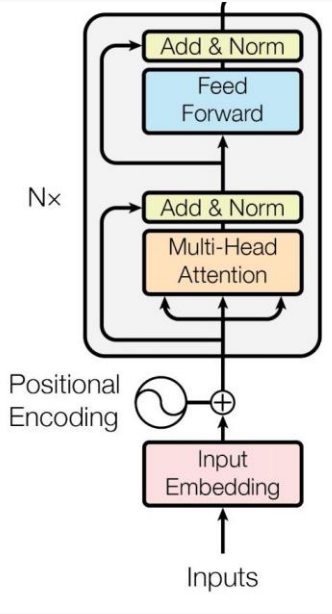


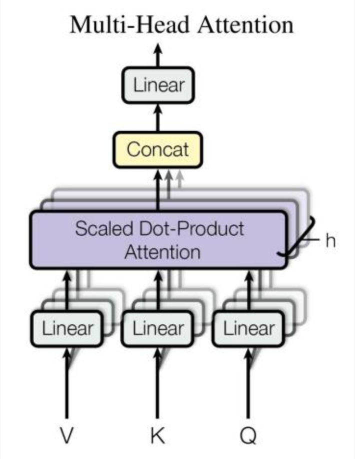
- Use 30,000 WordPiece vocabulary on input.
- Each token is sum of three embeddings
- Single sequence is much more efficient.

Model Architecture

Transformer encoder

- Multi-headed self attention
  - Models context
- Feed-forward layers
  - Computes non-linear hierarchical features
- Layer norm and residuals
  - Makes training deep networks healthy
- Positional embeddings
  - Allows model to learn relative positioning

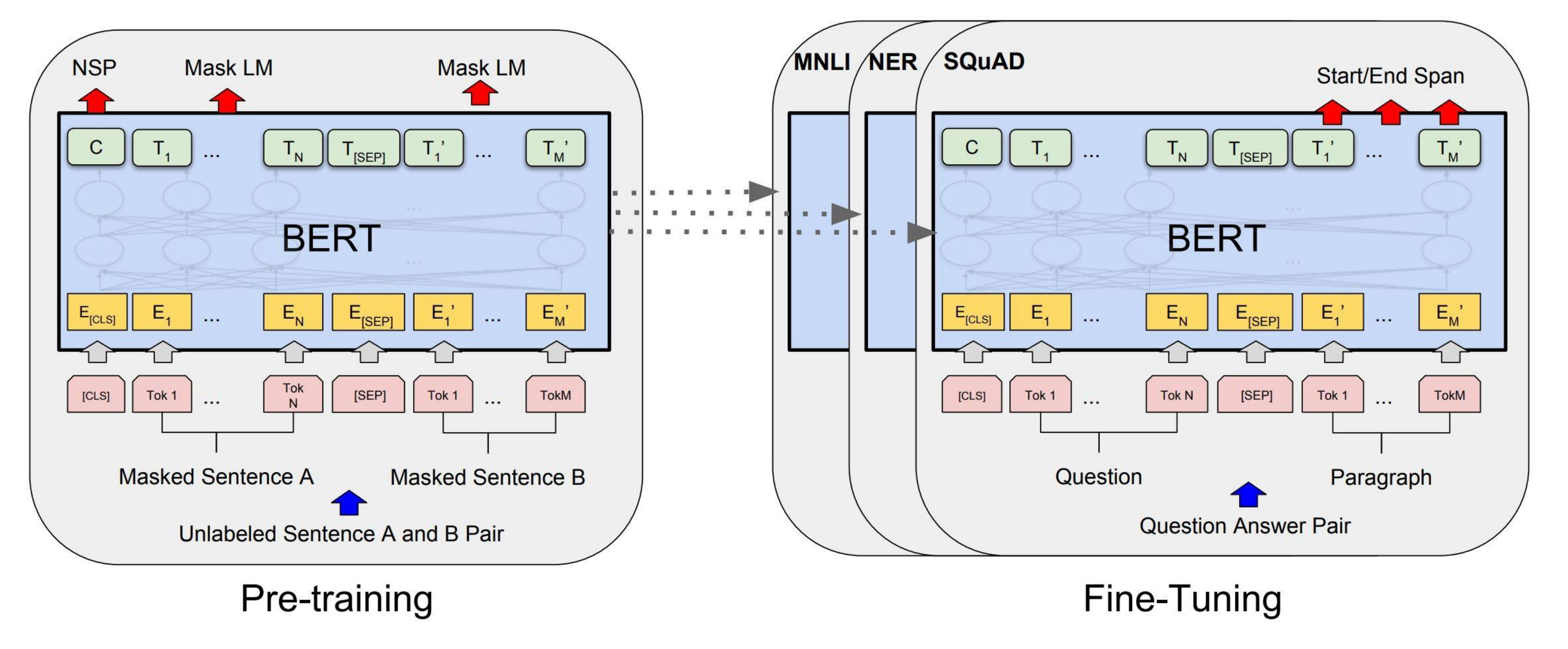




Bert Model Details

- Data: Wikipedia (2.5B words) + BookCorpus (800M words)
- Batch Size: 131,072 words (1024 sequences * 128 length or 256 sequences * 512 length)
- Training Time: 1M steps (~40 epochs)
- Optimizer: AdamW, 1e-4 learning rate, linear decay
- `BERT-Base`: 12-layer, 768-hidden, 12-head
- `BERT-Large`: 24-layer, 1024-hidden, 16-head
- Trained on 4x4 or 8x8 TPU slice for 4 days

Fine-Tuning Procedure



Effect of Model Size