BERT

Bidirectional Encoder Representations from Transformers

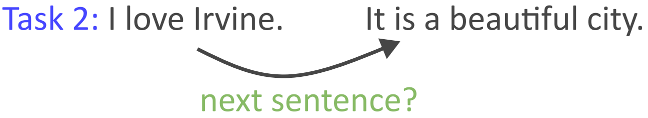
Dr. Alireza Aghamohammadi

BERT as a Pretrained Model

- BERT is a transformer model that has been pretrained on extensive text data.
- It leverages transfer learning.
- The pretrained model can be fine-tuned (e.g., by adding one or two dense layers) for specific tasks such as spam classification.
- During pretraining, BERT learns parameters through two tasks:
  1. Predicting missing words in sentences.
  2. Determining if two sentences were originally adjacent in the text.

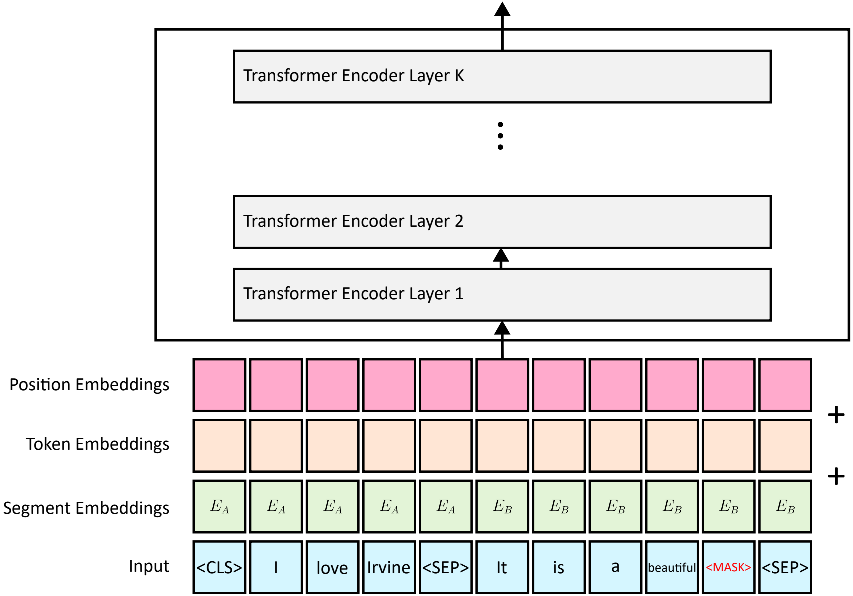
Original Text: I love Irvine. It is a beautiful city.

Task 1: I love Irvine. <MASK> is a beautiful <MASK>.



Task 2: I love Irvine. It is a beautiful city.

BERT Architecture



Transformer Encoder Layer K

⋮

Transformer Encoder Layer 2

Transformer Encoder Layer 1

Position Embeddings											
Token Embeddings											
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B	E_B
Input	<CLS>	I	love	Irvine	<SEP>	It	is	a	beautiful	<MASK>	<SEP>

Special Tokens: CLS and SEP

- ❖ **CLS** (Classification) is a special token added at the beginning of every input sequence.
- ❖ **SEP** (Separator) is a special token used to separate different segments of the input.
- ❖ The **CLS** token is prepended to the input text and passes through the transformer layers along with other tokens.
- ❖ The final hidden state of the **CLS** token represents the entire sentence.
- ❖ For example, in a spam classification task, the representation of the **CLS** token can be fed into a classifier to determine the sentence's class.

BERT Configuration

- BERT is available in two versions: BERT Base and BERT Large.
  - BERT Base:
    - Encoder Layers: 12
    - Feed Forward Hidden Layer Units: 768
    - Attention Heads: 12
    - Total Parameters: 110 million
  - BERT Large:
    - Encoder Layers: 24
    - Feed Forward Hidden Layer Units: 1024
    - Attention Heads: 16
    - Total Parameters: 340 million