



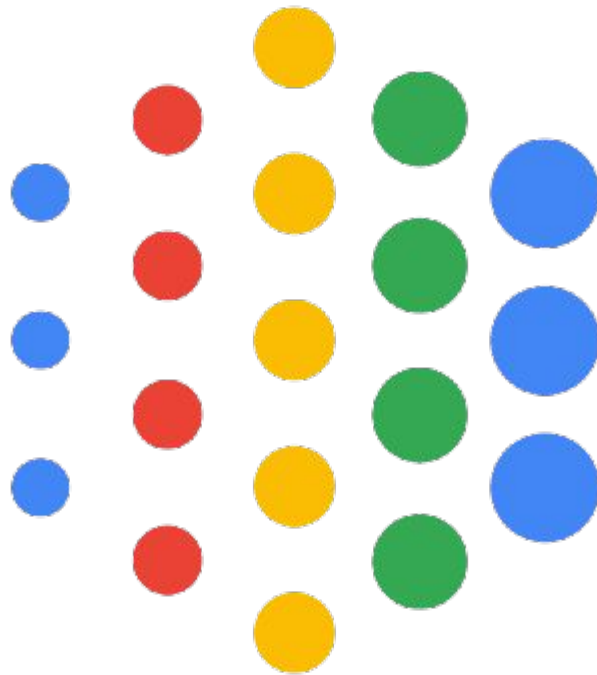
BERT

Presented by :
Justin, Ananya, Nikhil, Urvi, Priyanka,
Sanik



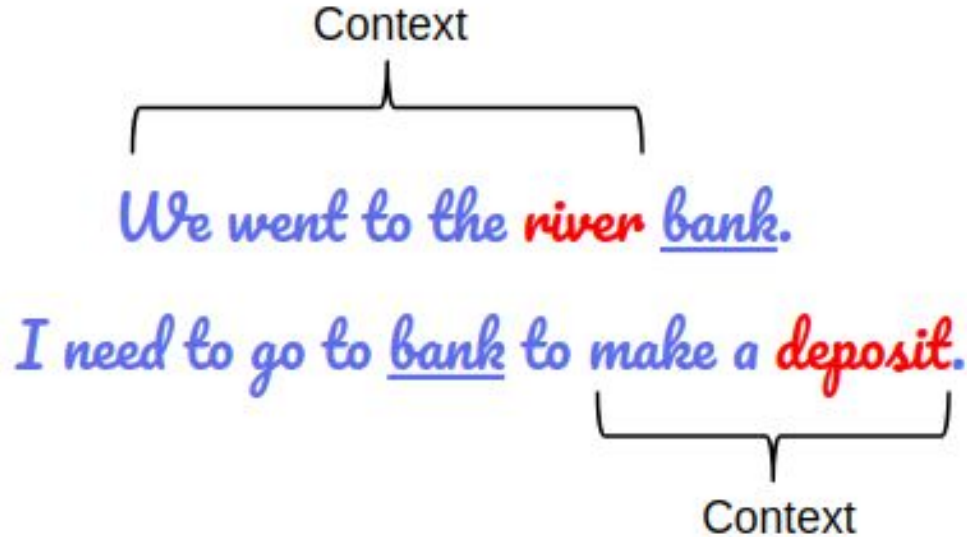
Overview

1. What is BERT?
2. BERT's Architecture
3. Text Preprocessing
4. Masked Language Modeling
5. Next Sentence Prediction



What is BERT?

- Bidirectional Encoder Representations from Transformers (BERT)
- State of the art language model for natural language processing (NLP)
- Developed by researchers at Google AI Language
- BERT “deeply bidirectional” - looks at text data in both directions to have a deeper sense of language context and flow
- Based on transformer architecture
- Trained on unlabeled text: Wikipedia and Book Corpus



BERT captures both the left and right context

Why is BERT Better?

- BERT is a pre-trained model that only needs one more output layer to be able to solve a wide variety of NLP problems
- BERT has been trained on large datasets that allow it to have a deeper understanding of how a language works
- Older models have limitations to the amount of data they can take in making many models too complex causing problems like overfitting
- Older models don't take in the context, or even if they do it's one sided causing many words with different meanings in different contexts to have the same vector

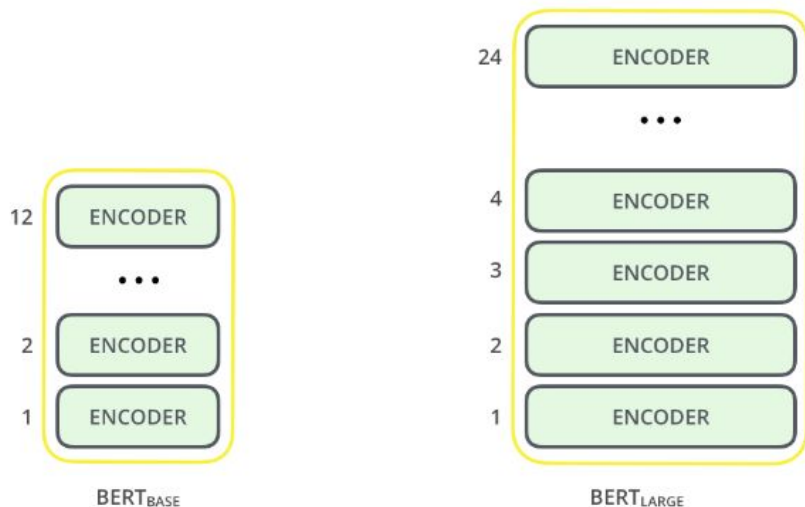
BERT's Architecture

The BERT architecture builds on top of Transformer. Two variants available:

- BERT Base: 12 layers (transformer blocks), 12 attention heads, and 110 million parameters
- BERT Large: 24 layers (transformer blocks), 16 attention heads and, 340 million parameters

	Training Compute + Time	Usage Compute
BERT _{BASE}	4 Cloud TPUs, 4 days	1 GPU
BERT _{LARGE}	16 Cloud TPUs, 4 days	1 TPU

BERT's Architecture Visualized

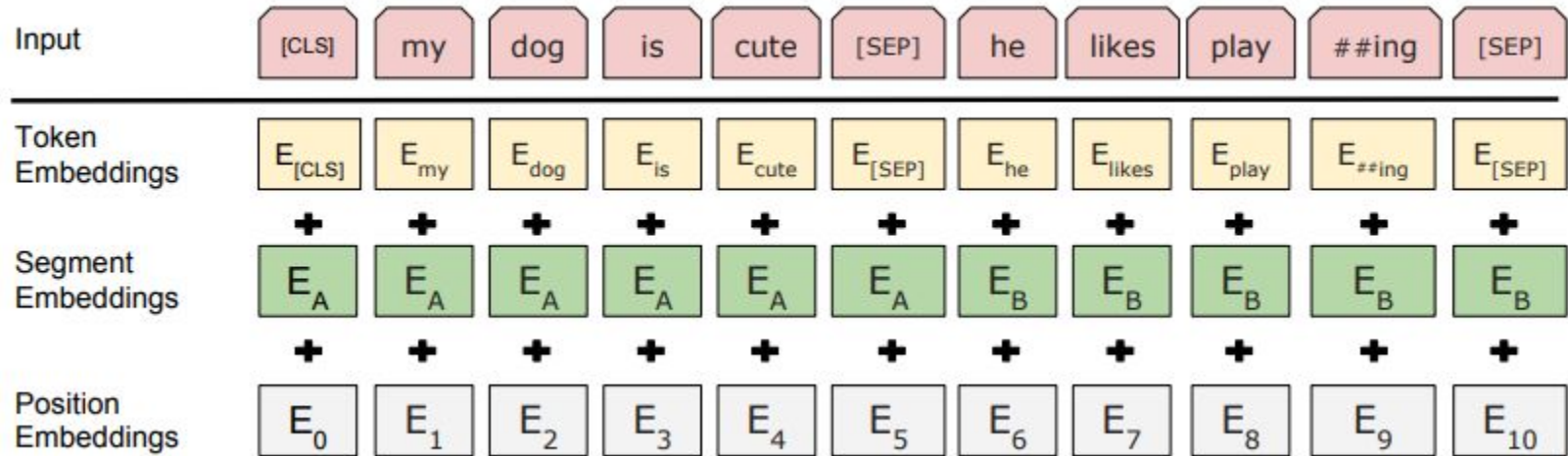


- **BERT base** was explicitly created to compare results with **OpenAI GPT** and control for model size*.
- As such, **BERT base** has exactly the same parameters as **OpenAI GPT**: $L=12$, $H=768$, $A=12$ where L is the number of stacked encoders, H is the hidden size and A is the number of heads in the MultiHead Attention layers.
- **BERT large** is basically larger and more compute-intensive: $L=24$, $H=1024$, $A=16$.

Text Preprocessing

1. **Position Embeddings:** BERT learns and uses positional embeddings to express the position of words in a sentence. These are added to overcome the limitation of the Transformer
2. **Segment Embeddings:** BERT can also take sentence pairs as inputs for tasks (Question-Answering). It learns a unique embedding for the first and the second sentences to help the model distinguish between them. In the example on the next slide, all the tokens marked as EA belong to sentence A (and similarly for EB)
3. **Token Embeddings:** These are the embeddings learned for the specific token from the WordPiece token vocabulary

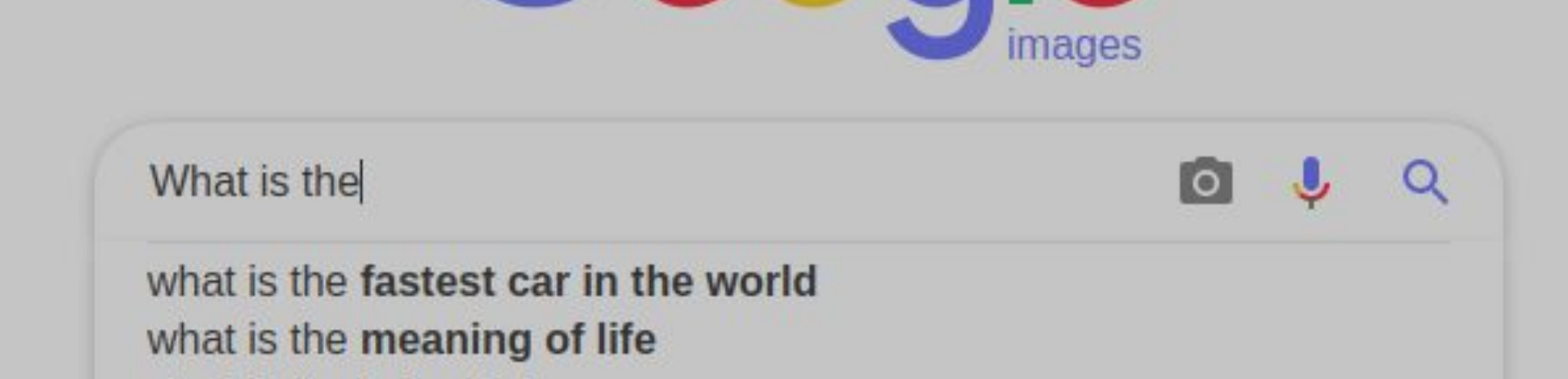
Text Preprocessing Visualized



Pre-training Tasks

BERT is pre-trained on two NLP tasks:

1. Masked Language Modeling
2. Next Sentence Prediction

A blurred screenshot of a Google search page. At the top, the Google logo is partially visible, with the word 'images' to its right. Below the logo is a search bar containing the text 'What is the'. To the right of the search bar are icons for a camera, a microphone, and a magnifying glass. Below the search bar, two search suggestions are visible: 'what is the fastest car in the world' and 'what is the meaning of life'.

What is the|

what is the **fastest car in the world**

what is the **meaning of life**

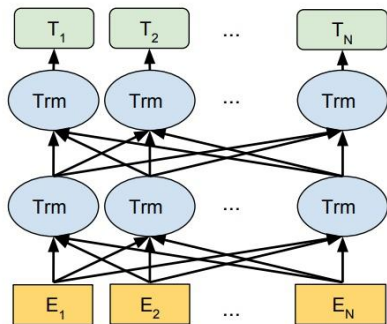
Masked Language Modeling (Bi-directionality)

Traditionally, we had language models either trained to predict the next word in a sentence (right-to-left context used in GPT) or language models that were trained on a left-to-right context.

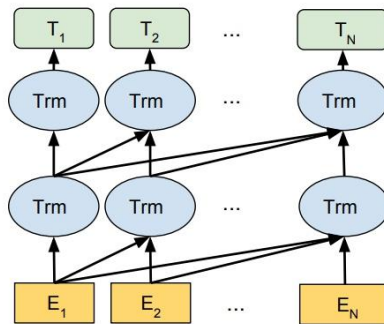
This made previous models susceptible to errors due to loss in information.

BERT vs GPT and ELMo

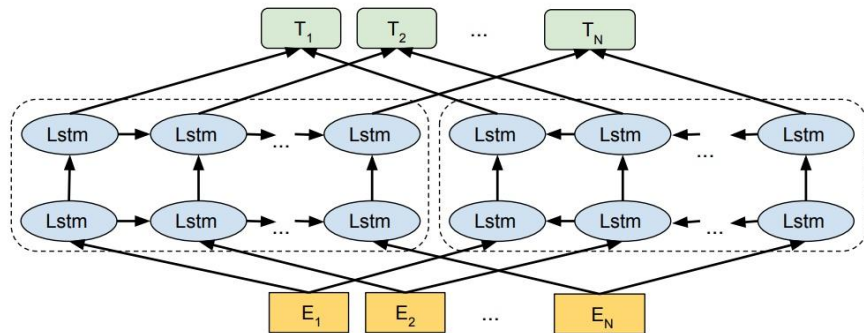
BERT (Ours)



OpenAI GPT



ELMo



About Masked Language Models

To prevent the model from focusing too much on a particular position or tokens that are masked, the researchers **randomly masked 15% of the words**

The masked words were not always replaced by the masked tokens [MASK] because the [MASK] token would never appear during fine-tuning

So, the researchers used the below technique:

- 80% of the time the words were replaced with the masked token [MASK]
- 10% of the time the words were replaced with random words
- 10% of the time the words were left unchanged

Next Sentence Prediction

- BERT is also trained on the task of Next Sentence Prediction for tasks that require an understanding of the relationship between sentences
- Question answering systems
- Inputs are given in pairs
- Output in the form of Loss function and Binary Value.

- **Input Question:**

Where do water droplets collide with ice
crystals to form precipitation?

- **Input Paragraph:**

... Precipitation forms as smaller droplets
coalesce via collision with other rain drops
or ice crystals within a cloud. ...

- **Output Answer:**

within a cloud

Beyond BERT: Current State-of-the-Art in NLP

BERT has inspired great interest in the field of NLP, especially the application of the Transformer for NLP tasks.

This has led to a spurt in the number of research labs and organizations that started experimenting with different aspects of pre-training, transformers and fine-tuning.

Many of these projects outperformed BERT on multiple NLP tasks.

Some of the most interesting developments were RoBERTa, which was Facebook AI's improvement over BERT and DistilBERT, which is a compact and faster version of BERT.

