UNIVERSITÄT PADERBORN

DATA SCIENCE RESEARCH GROUP

RECENT ADVANCES IN NATURAL LANUGAGE PROCESSING

TOPIC: BERT MODEL FOR LANGUAGE REPRESENTATION

BY: SAI NIKHIL MENON

UNIVERSITÄT PADERBORN

AGENDA

- What is BERT
- Architecture and Variants of BERT
- Attention Mechanism
- BERT Implementation – Input representation, Pre-training and Fine Tuning
- Experimental Results
- Discussion
- Applications of BERT

UNIVERSITÄT PADERBORN

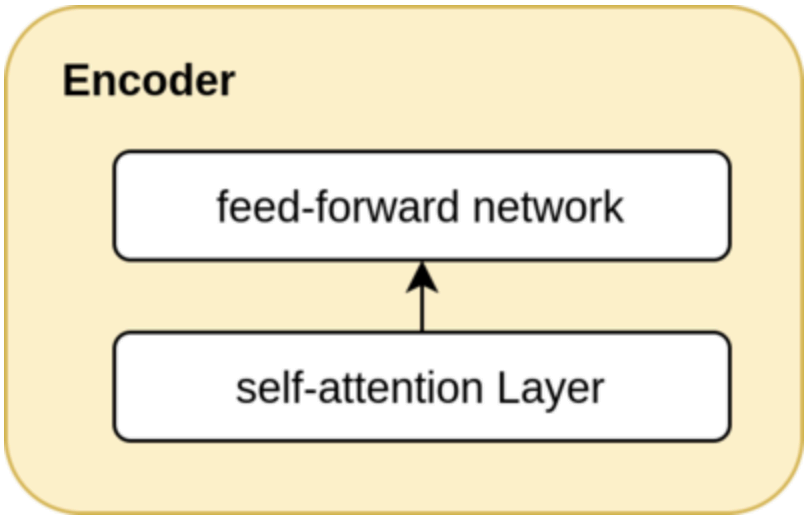
What is BERT ?

- **B**idirectional **E**ncoder **R**epresentations from **T**ransformer
- Language model developed by Google in 2018
- Capable of performing many NLP tasks – Sentiment Analysis, Machine Translation
- Captures contextual nuances bidirectionally using *Attention mechanism*
- Leverages *Transfer Learning* for realising downstream tasks

UNIVERSITÄT PADERBORN

Architecture of BERT

- Transformer neural network architecture at its core – enables **Attention!**
- Consists of stacked Encoders
- Encoders have two parts:
  1. Multi Headed Self Attention
  2. Positional Feed Forward layer

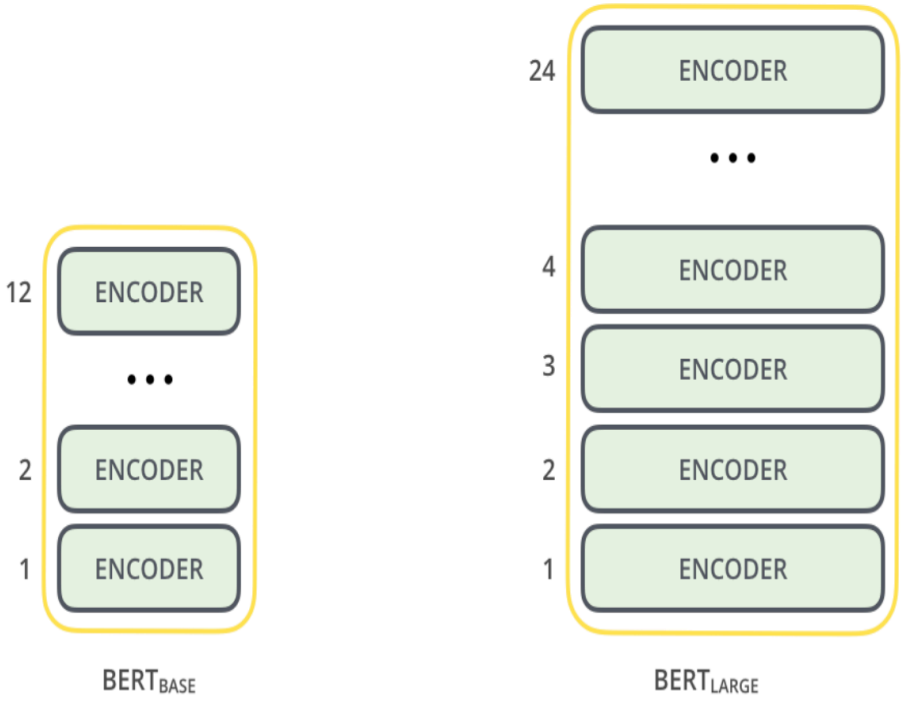


Source: https://lionbridge.ai/articles/what-are-transformer-models-in-machine-learning/

UNIVERSITÄT PADERBORN

Variants of BERT

- Two variants of BERT – 1) BERT-Base
  2) BERT-Large
- BERT-Base: 12 encoder layers
- BERT-Large: 24 encoder layers



Source: http://jalammar.github.io/illustrated-bert/

UNIVERSITÄT PADERBORN

Attention mechanism

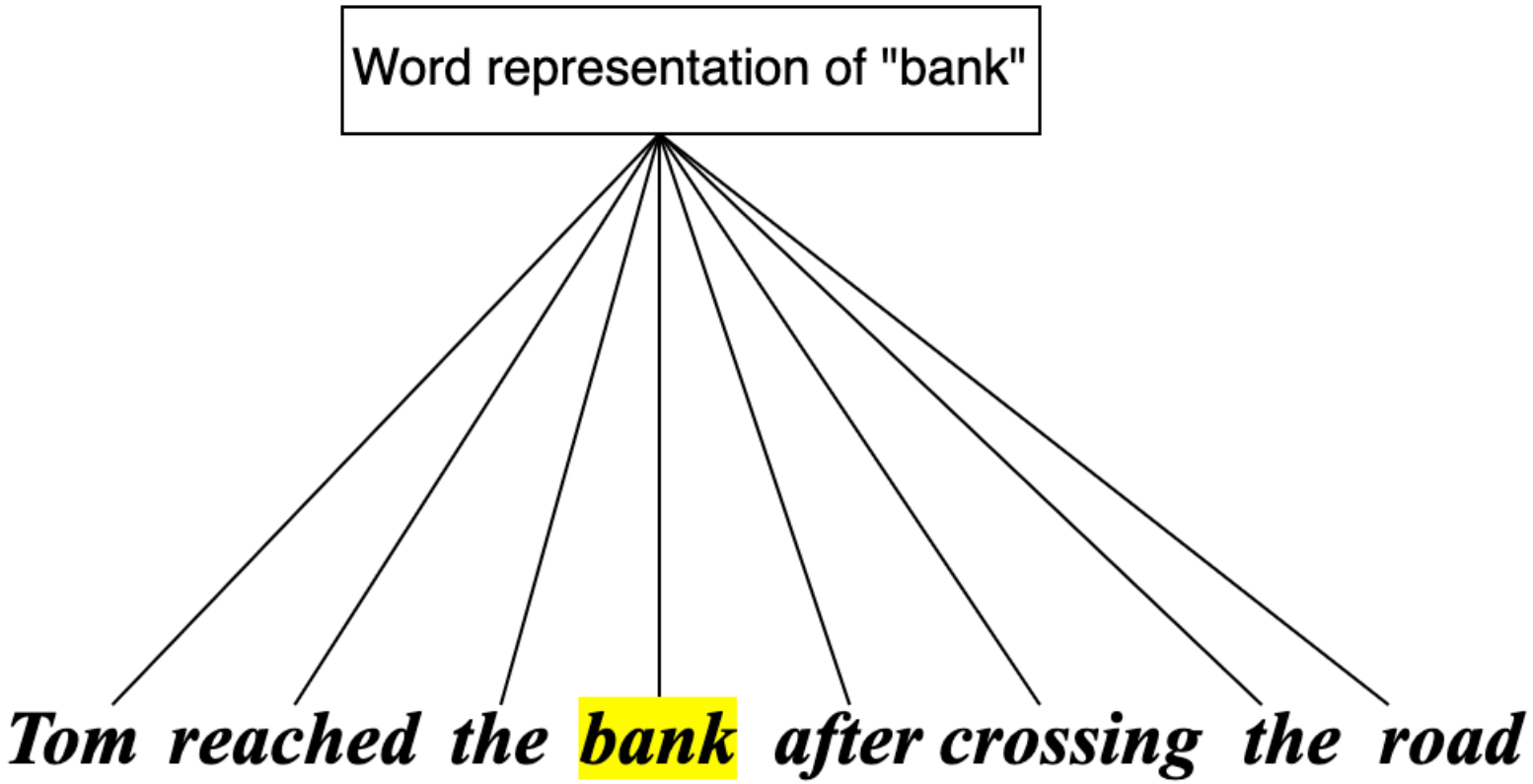
TOM REACHED THE BANK AFTER CROSSING THE

UNIVERSITÄT PADERBORN

Attention mechanism

UNIVERSITÄT PADERBORN

Contextual Word Representation



UNIVERSITÄT PADERBORN

BERT Implementation

Problems to Solve

- Machine Translation
- Question Answering
- Sentiment Analysis
- Text Summarisation

Requires Language Understanding

BERT Training

1. Pretrain BERT – for lanugage and context understanding
2. Fine Tuning BERT – for learning specific task

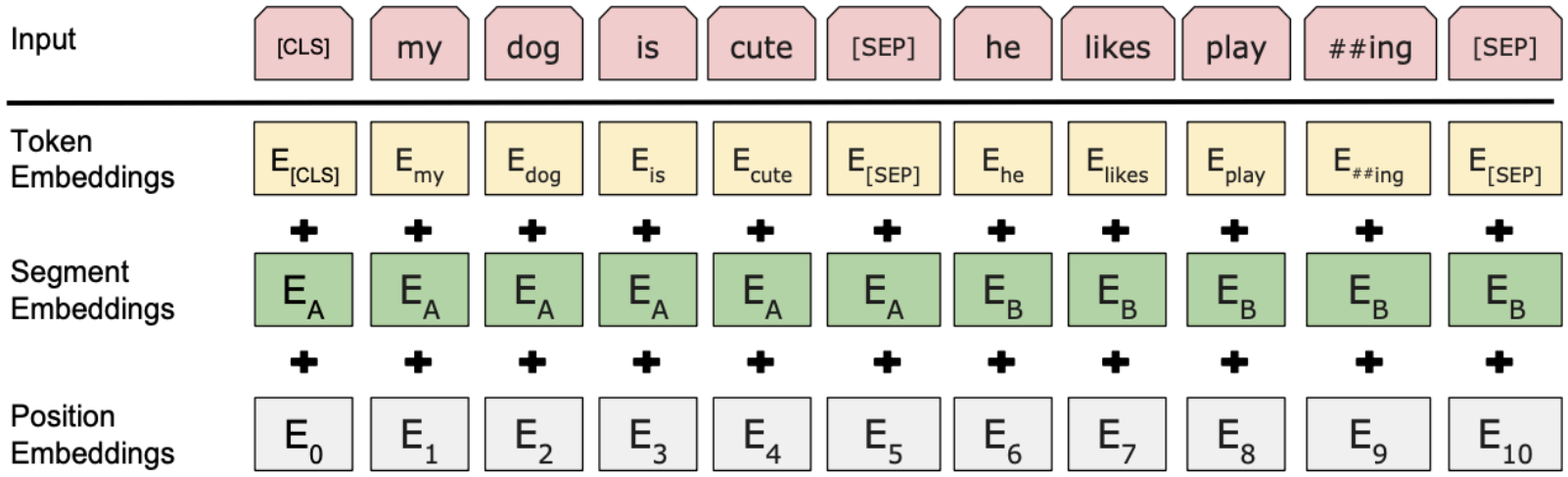
UNIVERSITÄT PADERBORN

Input Representation in BERT

Token Embeddings: Generated using WordPiece embeddings for each input word

Segment Embedding: Incorporates information about which part of sentence token belongs to

Positional Embedding: To incorporate the positional/ordering information of that token



Source: BERT: Pre-training of deep bidirectional Transformers for language understanding (Devlin et al,..2019)

UNIVERSITÄT PADERBORN

Pre-training BERT

- **Masked Language Modelling:**
  - Approximately 15% words in every input sequence is masked as [MASK]

THE QUICK BROWN FOX ⟶ THE [MASK] BROWN FOX

  - Task is to predict what the masked word is based on context.

UNIVERSITÄT PADERBORN

Pre-training BERT

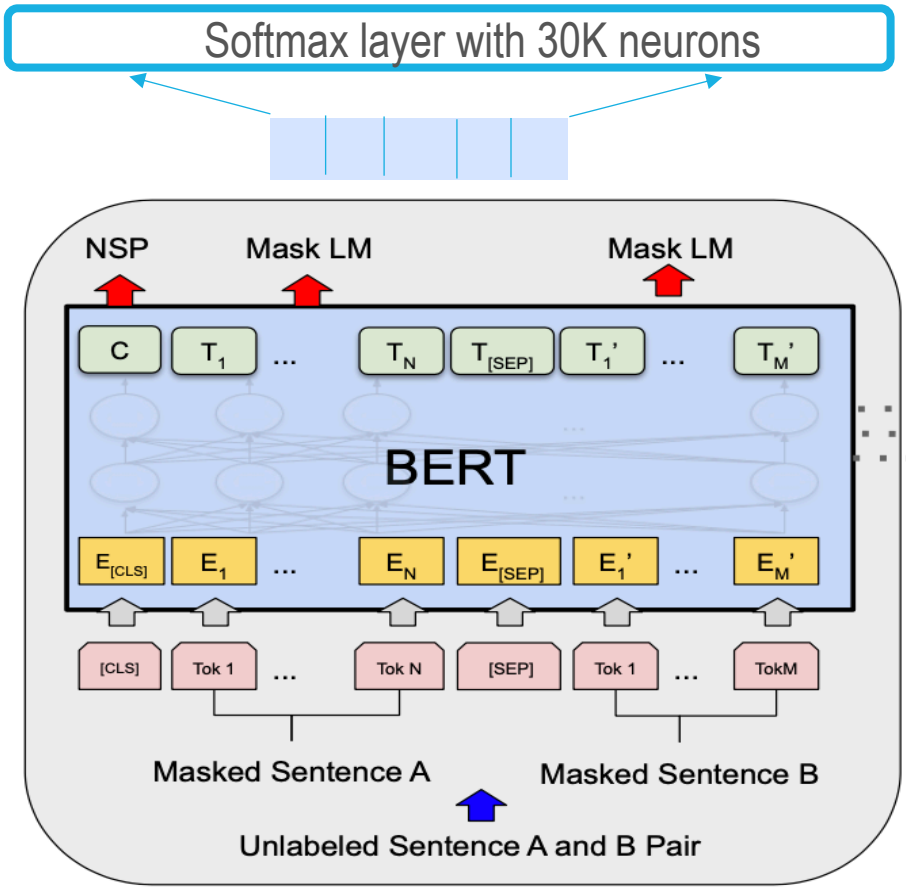
- **Next Sentence Prediction:**

- Task is to predict the relationship between sentencce A and B.

UNIVERSITÄT PADERBORN

Pre-training BERT

In practice, Masked Language Modelling and Next Sentence Prediction tasks are trained for simulatnouesly.



Source: BERT: Pre-training of deep bidirectional Transformers for language understanding (Devlin et al,..2019)

UNIVERSITÄT PADERBORN

Fine Tuning BERT

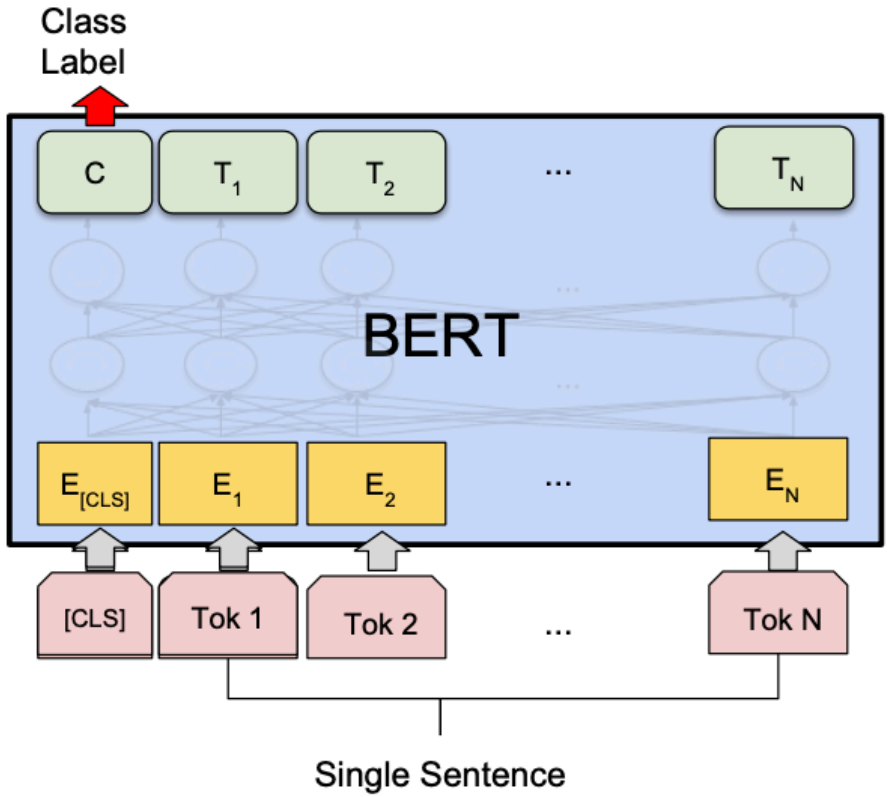
- Extend model to perform specific tasks – Question Answering, Sentiment Analysis etc
- Pre-trained model parameters are adjusted.
- Input to BERT are modified based on task.
- Output layer is modified based on task.
- Allows leveraging Transfer Learning in real world scenario

UNIVERSITÄT PADERBORN

Fine Tuning BERT

Modify output layer for classification task

Modify inputs to accept a single sentence



Source: BERT: Pre-training of deep bidirectional Transformers for language understanding (Devlin et al,..2019)

UNIVERSITÄT PADERBORN

Experimental Results

- BERT model was tested for GLUE tasks and recorded state-of-the art results on each of them.
- BERT-Large outperforms BERT-Base on each task.

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

Source: BERT: Pre-training of deep bidirectional Transformers for language understanding (Devlin et al,..2019)

UNIVERSITÄT PADERBORN

Discussion

PROS:

- Simultaneous processing of tokens
- Deeply bi-directional context awareness
- Unified architecture for Pre-training and Fine-tuning

CONS:

- Pre-training is extremely conputationally intensive
- Convergence is slow during pre-training due to masking

UNIVERSITÄT PADERBORN

BERT in Action

BERT has been tweaked and tuned leading to many variants of BERT for different applications:

- **RoBERTa**: Used by Facebook for content moderation
- **bioBERT**: Used in Biomedical industry for biomedical text mining
- **docBERT**: Fine-tuned BERT model for document classification
- **videoBERT**: Visual Linguistic model based on BERT for video captioning and other such tasks used by YouTube.

UNIVERSITÄT PADERBORN

Thank you