Bert

Tushar B. Kute,
http://tusharkute.com

BERT

- BERT is an open source machine learning framework for natural language processing (NLP).
- BERT is designed to help computers understand the meaning of ambiguous language in text by using surrounding text to establish context.
- The BERT framework was pre-trained using text from Wikipedia and can be fine-tuned with question and answer datasets.

BERT

- BERT, which stands for Bidirectional Encoder Representations from Transformers, is based on Transformers, a deep learning model in which every output element is connected to every input element, and the weightings between them are dynamically calculated based upon their connection. (In NLP, this process is called attention.)

BERT

- Historically, language models could only read text input sequentially -- either left-to-right or right-to-left -- but couldn't do both at the same time.
- BERT is different because it is designed to read in both directions at once. This capability, enabled by the introduction of Transformers, is known as bidirectionality.
- Using this bidirectional capability, BERT is pre-trained on two different, but related, NLP tasks: Masked Language Modeling and Next Sentence Prediction.

BERT: Background

- Transformers were first introduced by Google in 2017. At the time of their introduction, language models primarily used recurrent neural networks (RNN) and convolutional neural networks (CNN) to handle NLP tasks.
- Although these models are competent, the Transformer is considered a significant improvement because it doesn't require sequences of data to be processed in any fixed order, whereas RNNs and CNNs do.
- Because Transformers can process data in any order, they enable training on larger amounts of data than ever was possible before their existence.
- This, in turn, facilitated the creation of pre-trained models like BERT, which was trained on massive amounts of language data prior to its release.

BERT: Background

- In 2018, Google introduced and open-sourced BERT. In its research stages, the framework achieved groundbreaking results in 11 natural language understanding tasks, including sentiment analysis, semantic role labeling, sentence classification and the disambiguation of polysemous words, or words with multiple meanings.
- Completing these tasks distinguished BERT from previous language models such as word2vec and GloVe, which are limited when interpreting context and polysemous words.
- BERT effectively addresses ambiguity, which is the greatest challenge to natural language understanding according to research scientists in the field. It is capable of parsing language with a relatively human-like "common sense".

BERT: Background

- it’s easy to get that BERT stands for Bidirectional Encoder Representations from Transformers.
- Each word here has a meaning to it and we will encounter that one by one in this article. For now, the key takeaway from this line is – BERT is based on the Transformer architecture.
- Second, BERT is pre-trained on a large corpus of unlabelled text including the entire Wikipedia(that’s 2,500 million words!) and Book Corpus (800 million words).

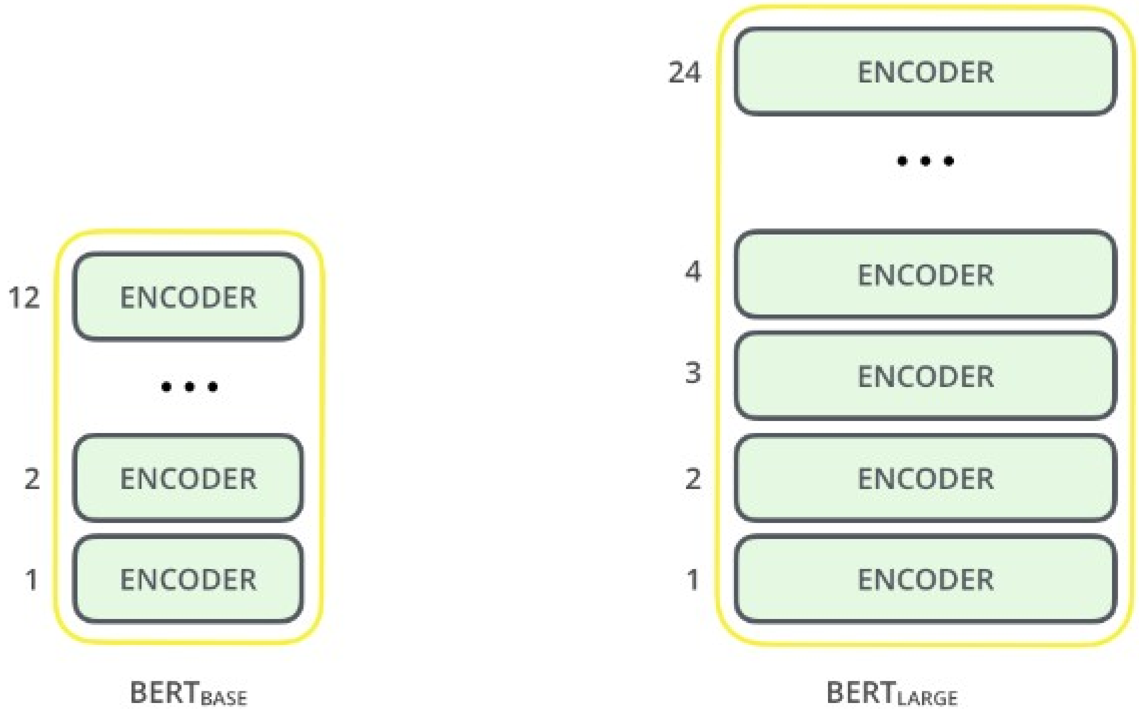
BERT: Background

- This pre-training step is half the magic behind BERT's success. This is because as we train a model on a large text corpus, our model starts to pick up the deeper and intimate understandings of how the language works. This knowledge is the swiss army knife that is useful for almost any NLP task.
- Third, BERT is a "deeply bidirectional" model. Bidirectional means that BERT learns information from both the left and the right side of a token's context during the training phase.

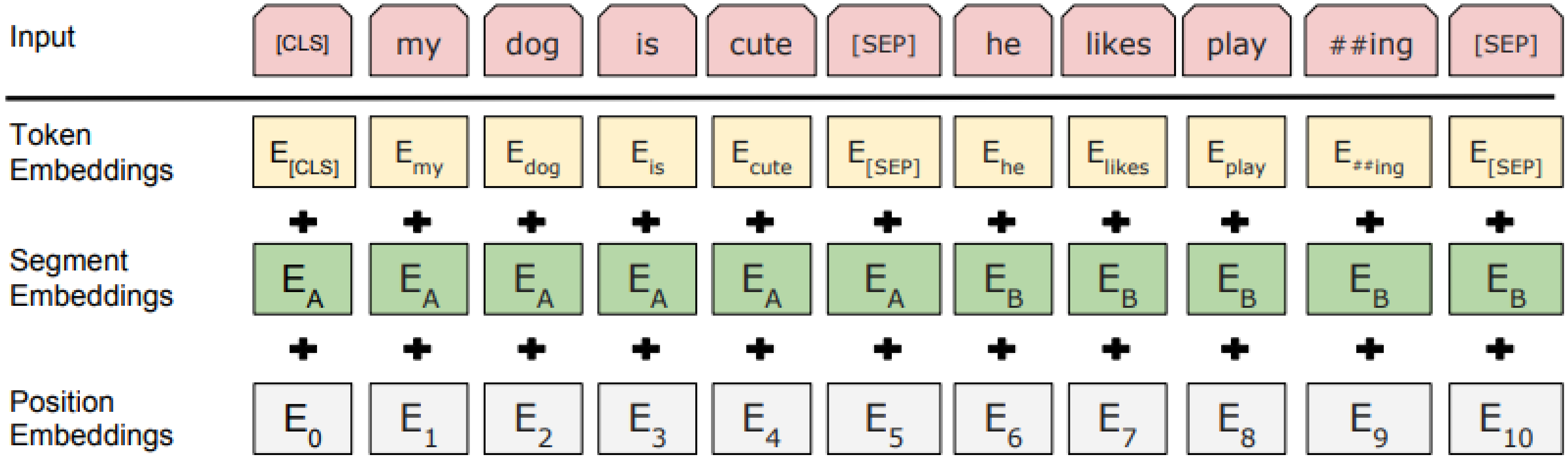
BERT: Architecture

- The BERT architecture builds on top of Transformer. We currently have 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

BERT: Architecture



Text Preprocessing



Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{\#\#ing}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}

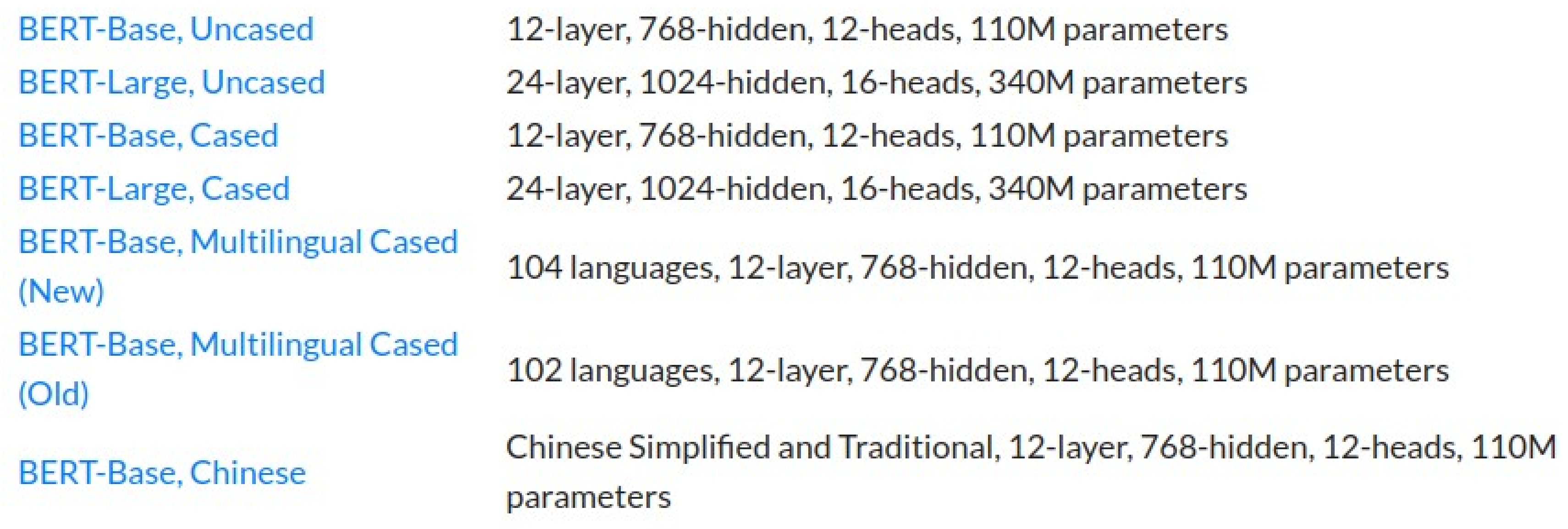
Input Embedding

- 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 Transformer which, unlike an RNN, is not able to capture “sequence” or “order” information
- Segment Embeddings: BERT can also take sentence pairs as inputs for tasks (Question-Answering). That’s why it learns a unique embedding for the first and the second sentences to help the model distinguish between them. In the above example, all the tokens marked as EA belong to sentence A (and similarly for EB)
- Token Embeddings: These are the embeddings learned for the specific token from the WordPiece token vocabulary.

Pre-training tasks

- BERT is pre-trained on two NLP tasks:
  - Masked Language Modeling
  - Next Sentence Prediction

Pre-training BIRT Models



BERT-Base, Uncased	12-layer, 768-hidden, 12-heads, 110M parameters
BERT-Large, Uncased	24-layer, 1024-hidden, 16-heads, 340M parameters
BERT-Base, Cased	12-layer, 768-hidden, 12-heads, 110M parameters
BERT-Large, Cased	24-layer, 1024-hidden, 16-heads, 340M parameters
BERT-Base, Multilingual Cased (New)	104 languages, 12-layer, 768-hidden, 12-heads, 110M parameters
BERT-Base, Multilingual Cased (Old)	102 languages, 12-layer, 768-hidden, 12-heads, 110M parameters
BERT-Base, Chinese	Chinese Simplified and Traditional, 12-layer, 768-hidden, 12-heads, 110M parameters

Pre-trained models

- patentBERT - a BERT model fine-tuned to perform patent classification.
- docBERT - a BERT model fine-tuned for document classification.
- bioBERT - a pre-trained biomedical language representation model for biomedical text mining.
- VideoBERT - a joint visual-linguistic model for process unsupervised learning of an abundance of unlabeled data on Youtube.
- SciBERT - a pretrained BERT model for scientific text

Pre-trained models

- G-BERT - a BERT model pretrained using medical codes with hierarchical representations using graph neural networks (GNN) and then fine-tuned for making medical recommendations.
- TinyBERT by Huawei - a smaller, "student" BERT that learns from the original "teacher" BERT, performing transformer distillation to improve efficiency. TinyBERT produced promising results in comparison to BERT-base while being 7.5 times smaller and 9.4 times faster at inference.
- DistilBERT by HuggingFace - a supposedly smaller, faster, cheaper version of BERT that is trained from BERT, and then certain architectural aspects are removed for the sake of efficiency.

Thank you

This presentation is created using LibreOffice Impress 7.4.1.2, can be used freely as per GNU General Public License

@mitu_skillologies

@mITuSkillologies

@mitu_group

@mitu-skillologies

@MITUSkillologies

kaggle

@mituskillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

@mituskillologies

contact@mitu.co.in

tushar@tusharkute.com