Contextual Embeddings – BERT

Apr 9th, 2019

Applied Deep Learning

YUN-NUNG (VIVIAN) CHEN HTTP://ADL.MIULAB.TW

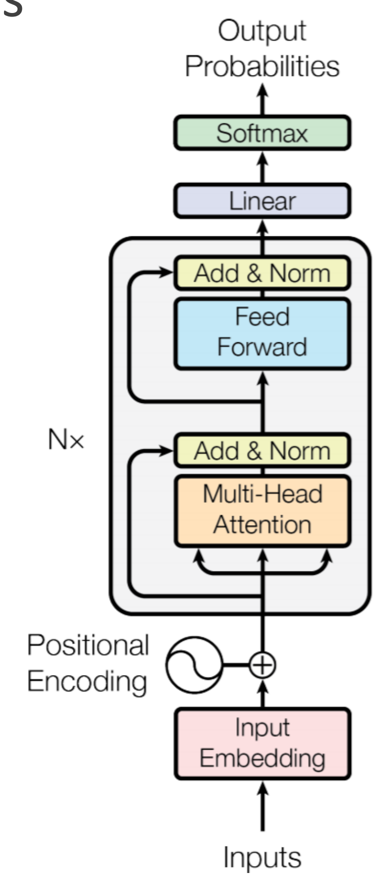
國立臺灣大學
National Taiwan University

Slides credited from Jacob Devlin

BERT: Bidirectional Encoder Representations from Transformers

Idea: contextualized word representations

- Learn word vectors using long contexts using Transformer instead of LSTM



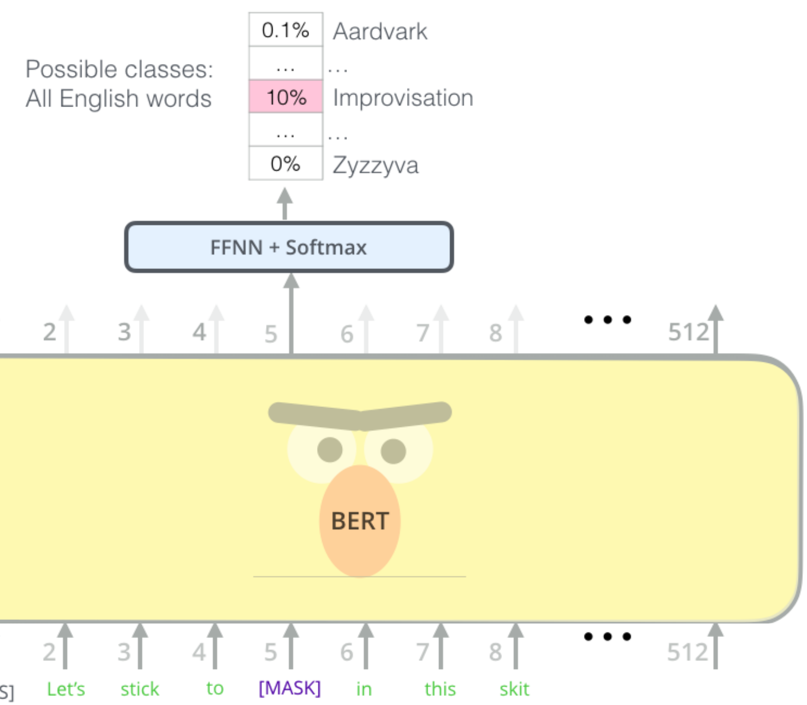
Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, in *NAACL-HLT*, 2019.

BERT #1 – Masked Language Model

Idea: language understanding is bidirectional while LM only uses *left* or *right* context

- This is not a generation task

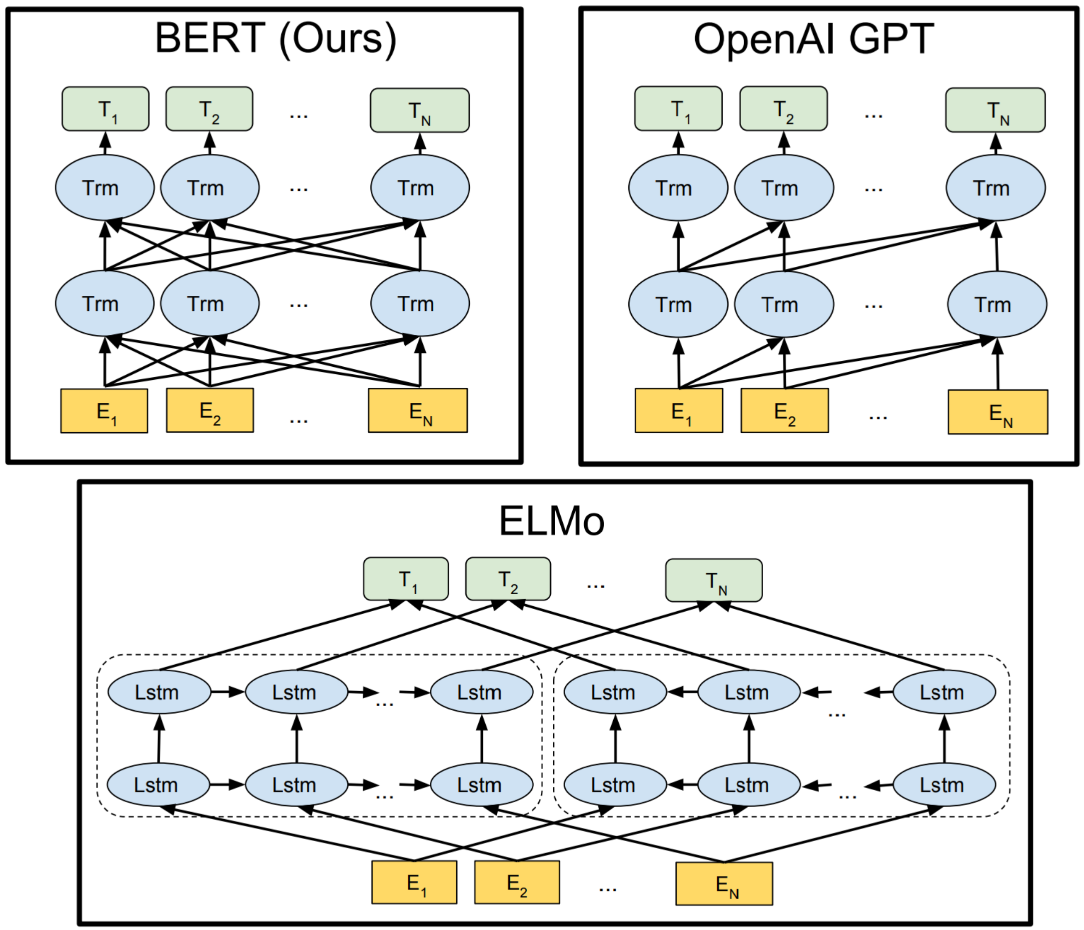
Use the output of the
masked word's position
to predict the masked word



Randomly mask 15% of tokens

- Too little: expensive to train
- Too much: not enough context

BERT #1 – Masked Language Model



BERT #2 – Next Sentence Prediction

Idea: modeling ***relationship*** between sentences

- QA, NLI etc. are based on understanding inter-sentence relationship

Input = `[CLS] the man went to [MASK] store [SEP]`

`he bought a gallon [MASK] milk [SEP]`

Label = `IsNext`

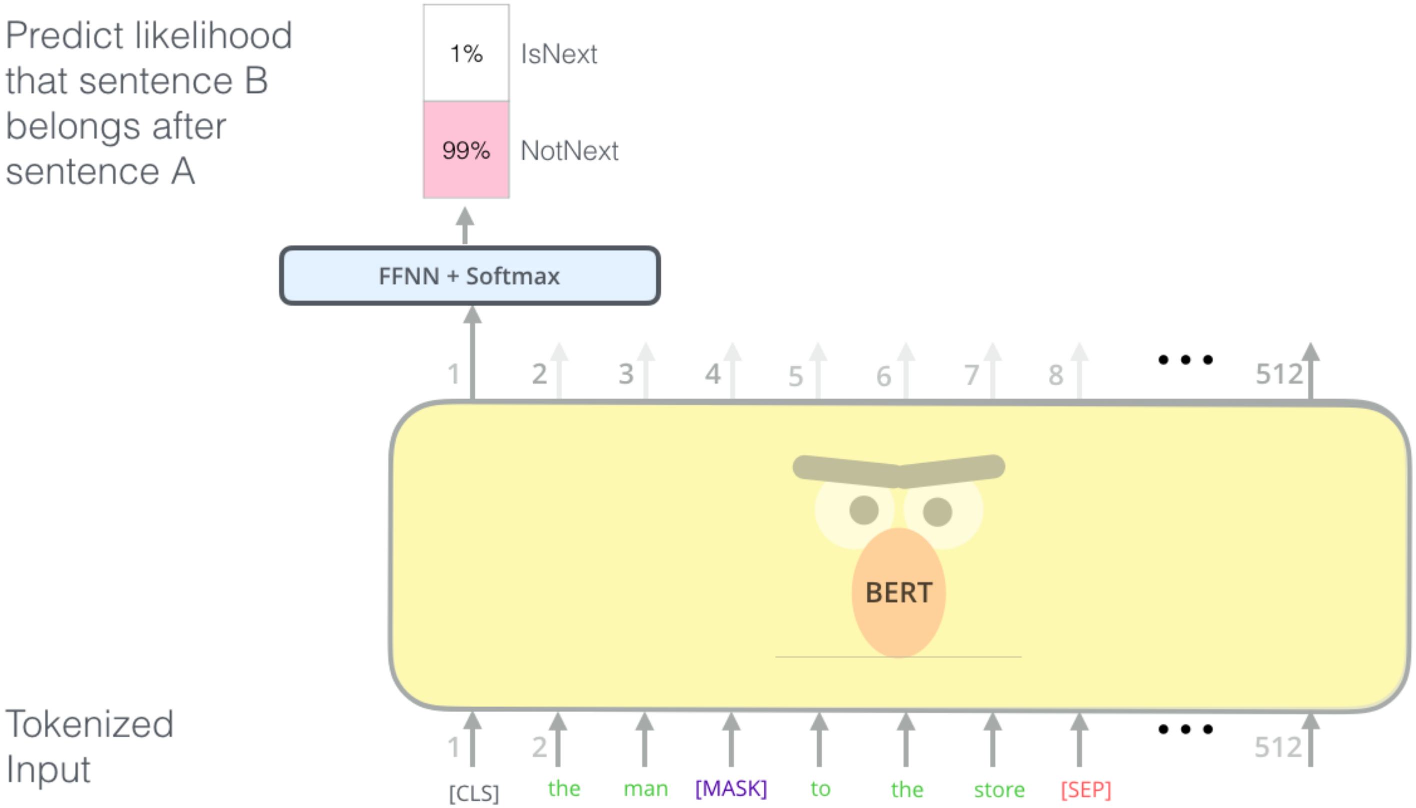
Input = `[CLS] the man [MASK] to the store [SEP]`

`penguin [MASK] are flight ##less birds [SEP]`

Label = `NotNext`

BERT #2 – Next Sentence Prediction

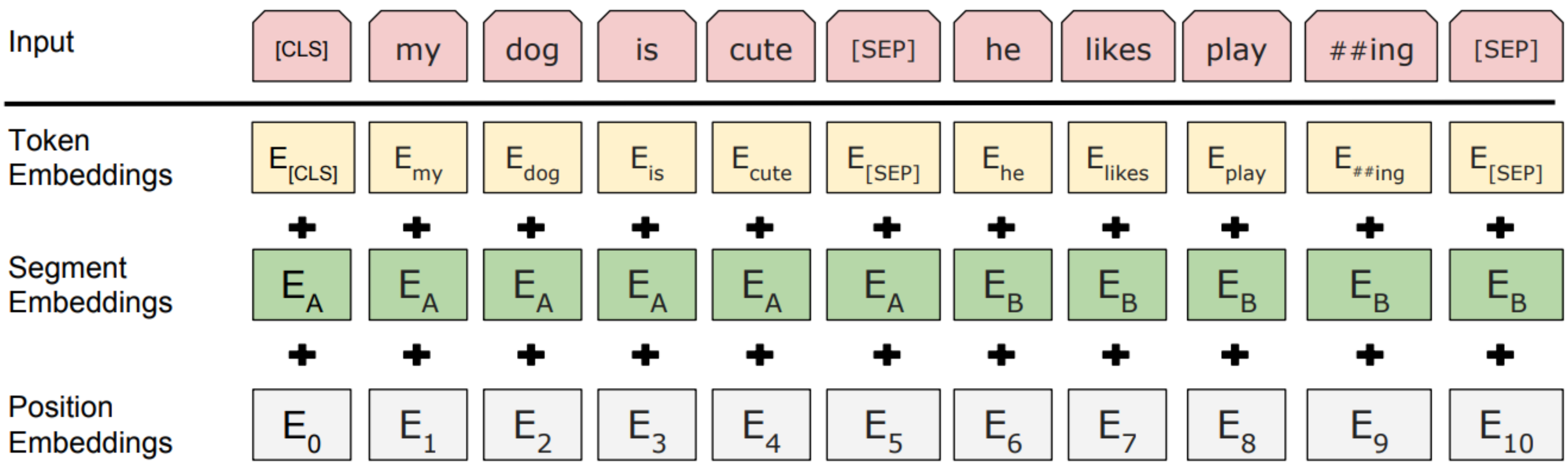
Idea: modeling ***relationship*** between sentences



BERT – Input Representation

Input embeddings contain

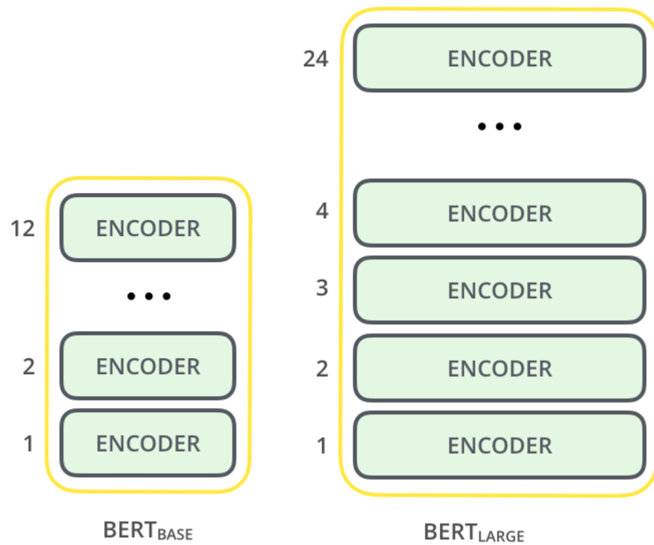
- Word-level token embeddings
- Sentence-level segment embeddings
- Position embeddings

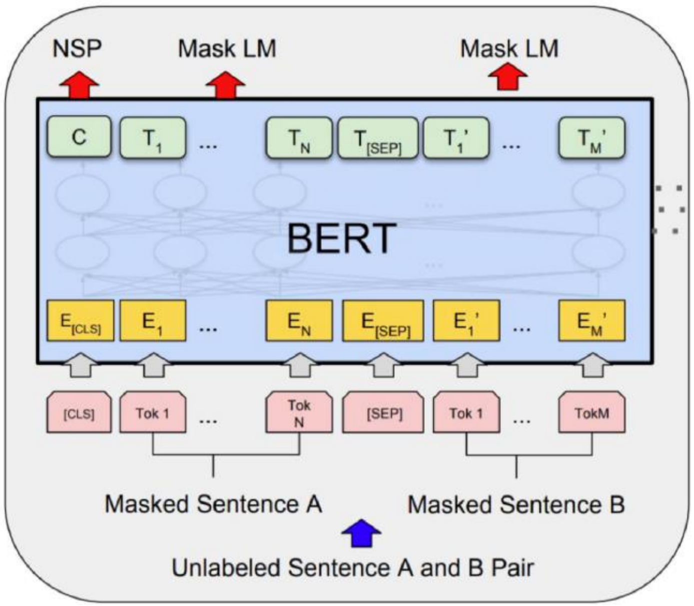


BERT – Training

Training data: Wikipedia + BookCorpus

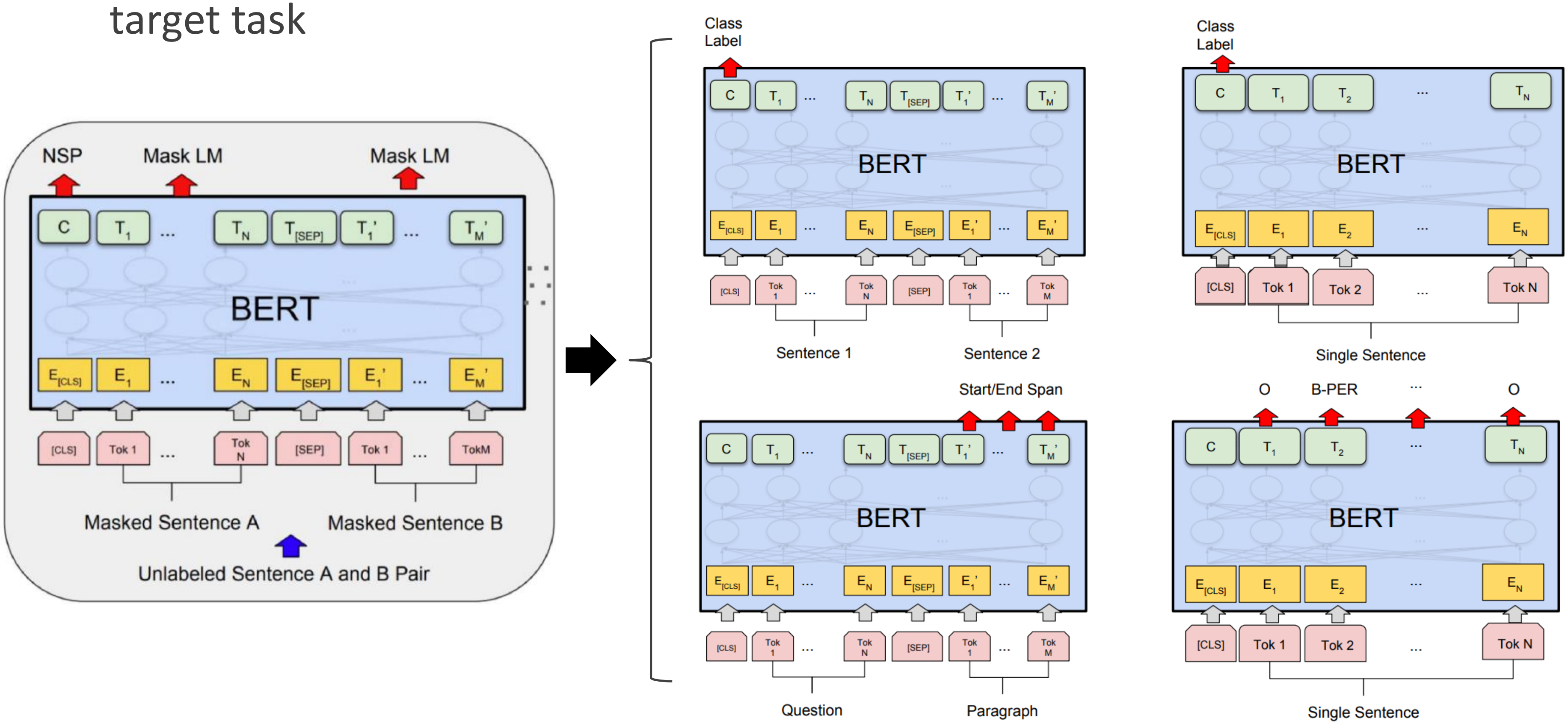
2 BERT models
- BERT-Base: 12-layer, 768-hidden, 12-head
- BERT-Large: 24-layer, 1024-hidden, 16-head





BERT for Fine-Tuning Understanding Tasks

Idea: simply learn a classifier/tagger built on the top layer for each target task

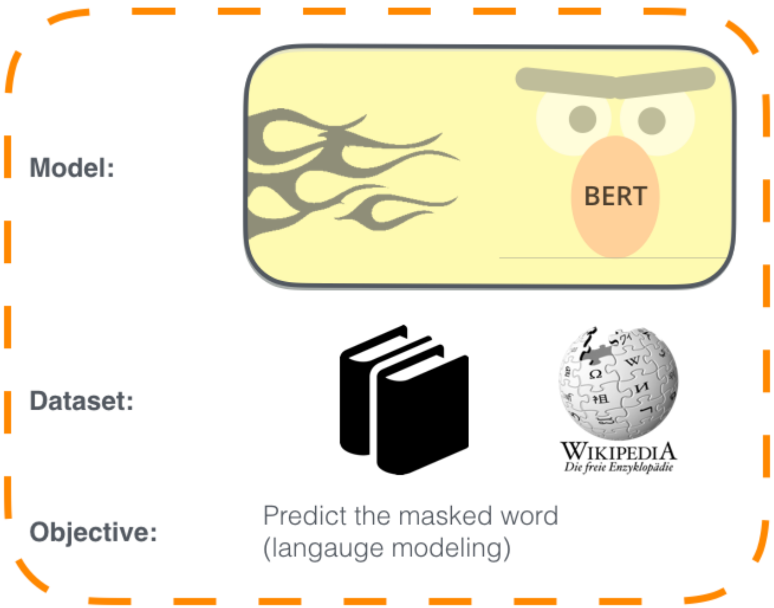


BERT Overview

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

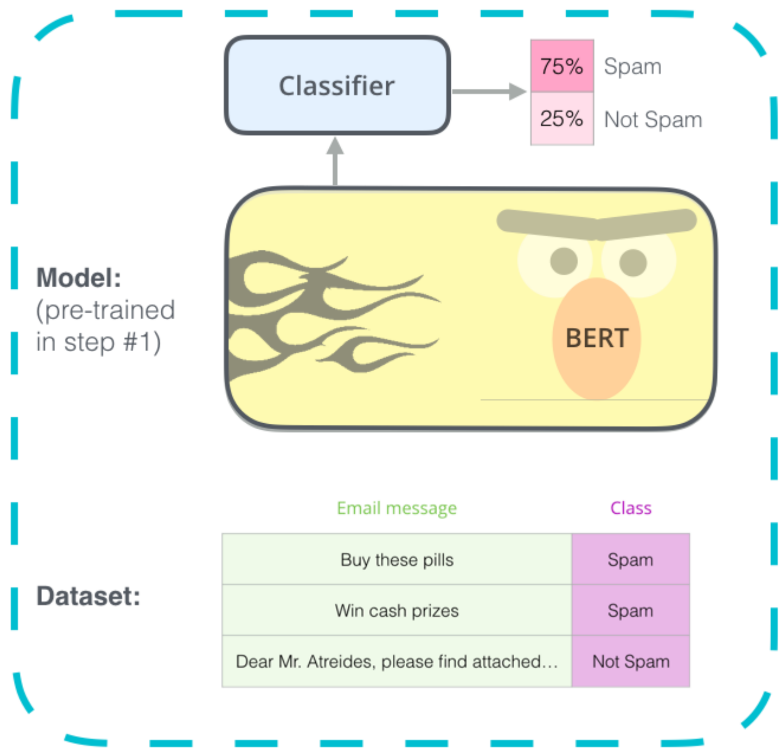
The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

Semi-supervised Learning Step

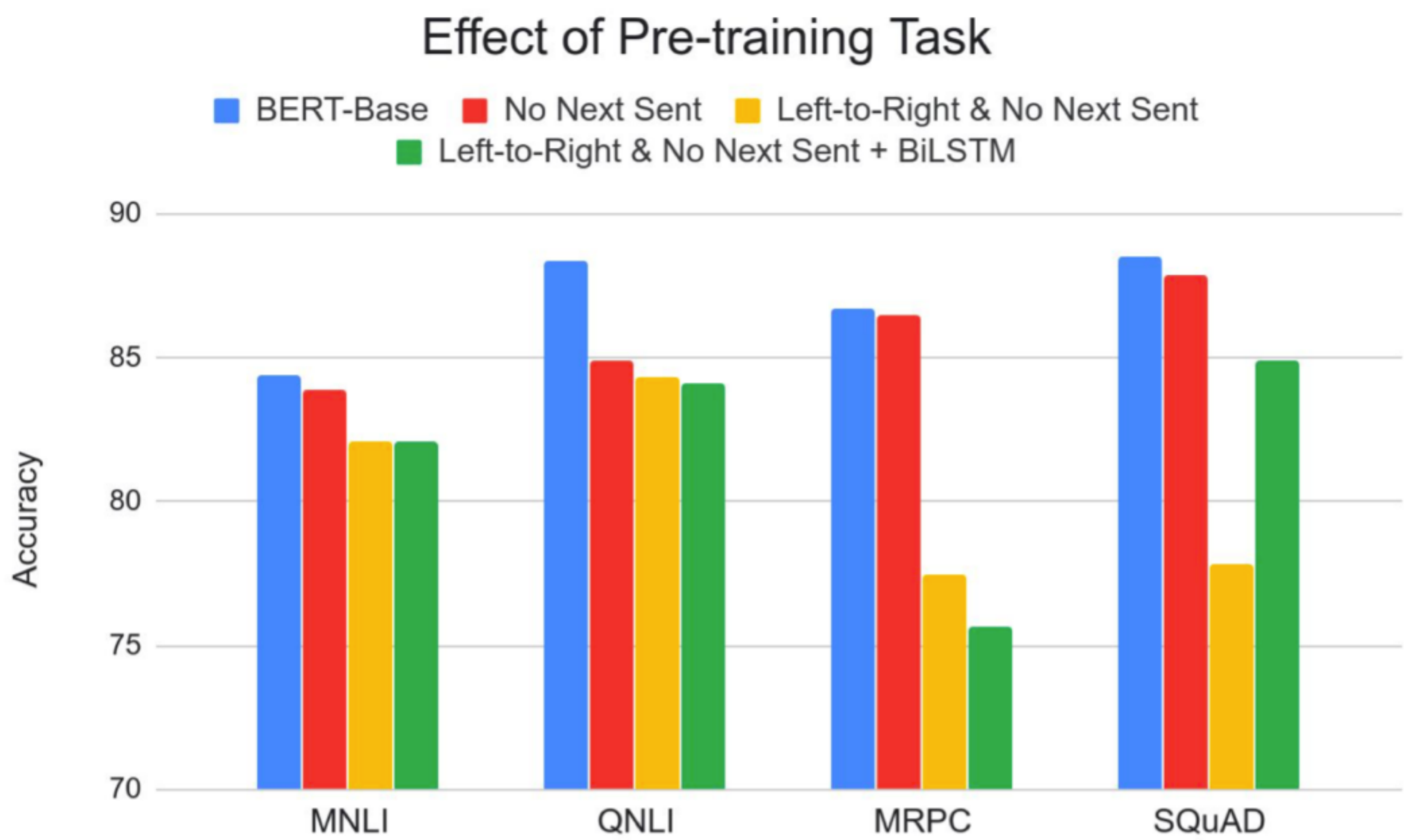


2 - Supervised training on a specific task with a labeled dataset.

Supervised Learning Step



BERT Fine-Tuning Results



BERT Results on SQuAD 2.0

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	**87.147**	**89.474**
2 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
3 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147
4 Mar 16, 2019	BERT + DAE + AoA (single model) *Joint Laboratory of HIT and iFLYTEK Research*	85.884	88.621
5 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	85.082	87.615
5 Mar 13, 2019	BERT + ConvLSTM + MTL + Verifier (single model) *Layer 6 AI*	84.924	88.204
5 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (single model) *Google AI Language* https://github.com/google-research/bert	85.150	87.715

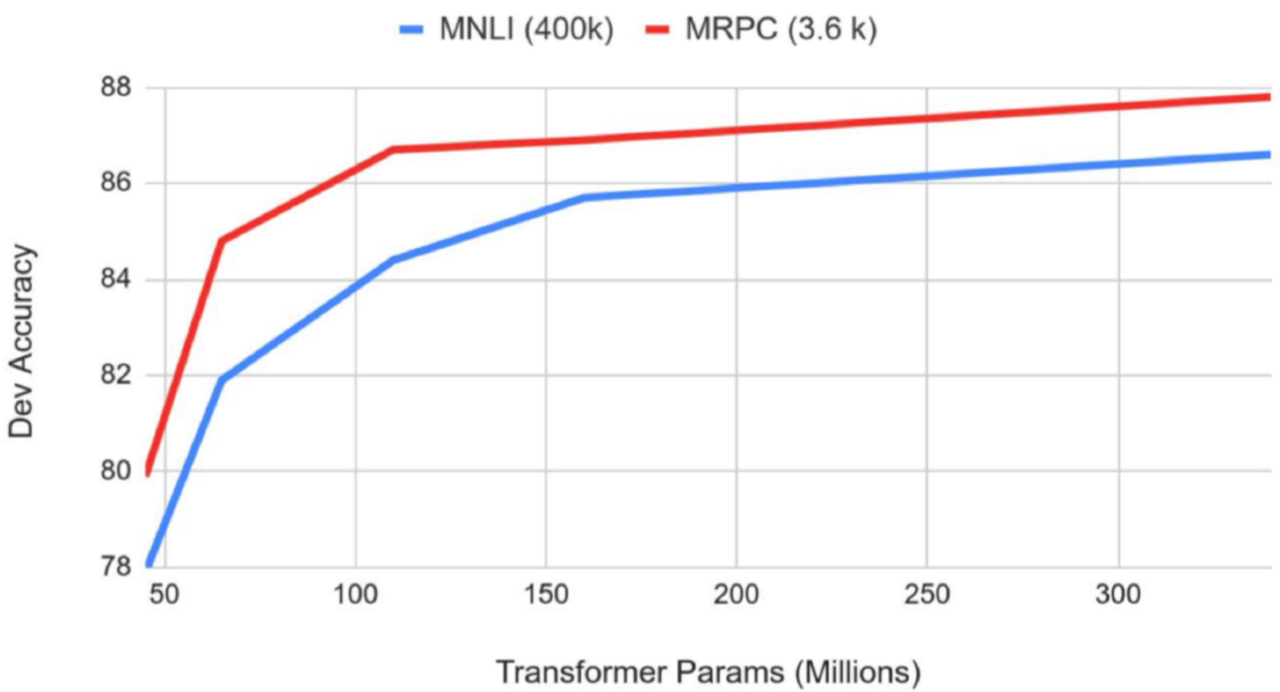
BERT Results on NER

Model	Description	CONLL 2003 F1
TagLM (Peters+, 2017)	LSTM BiLM in BLSTM Tagger	91.93
ELMo (Peters+, 2018)	ELMo in BLSTM	92.22
BERT-Base (Devlin+, 2019)	Transformer bidi LM + fine tune	92.4
CVT Clark	Cross-view training + multitask learn	92.61
BERT-Large (Devlin+, 2019)	Transformer bidi LM + fine tune	92.8
Flair	Character-level language model	93.09

Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", in *NAACL-HLT*, 2019.

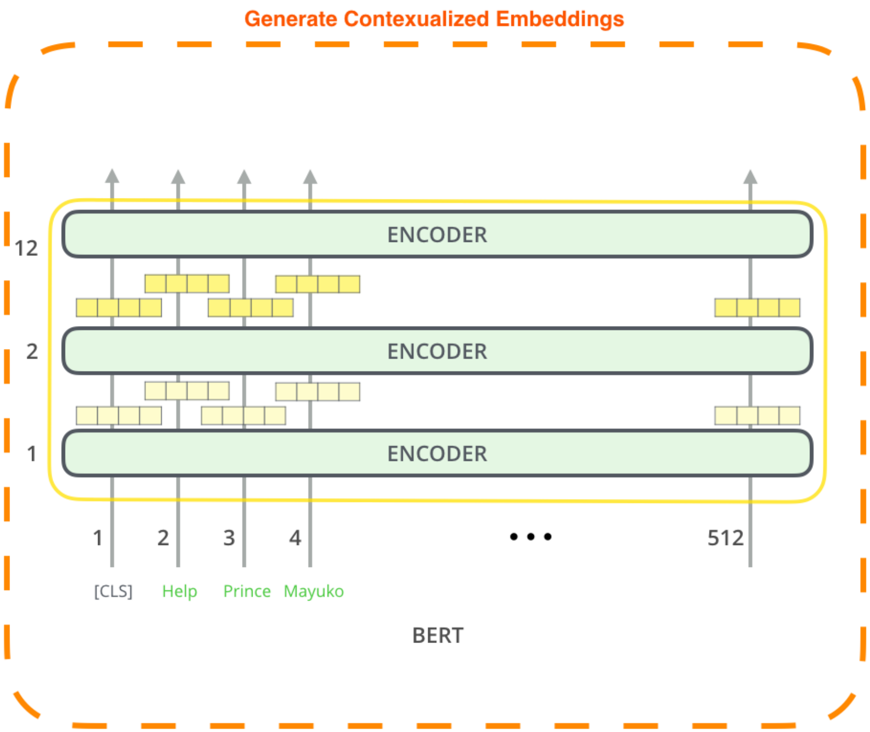
BERT Results with Different Model Sizes

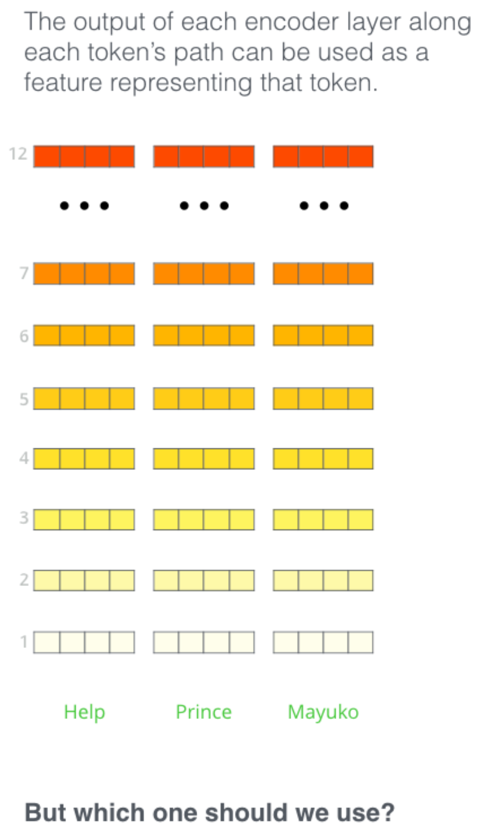
Improving performance by increasing model size



BERT for Contextualized Word Embeddings

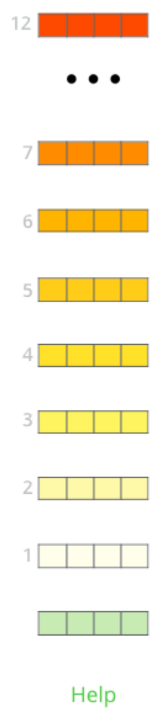
Idea: use pre-trained BERT to get contextualized word embeddings and feed them into the task-specific models





BERT Embeddings Results on NER

What is the best contextualized embedding for "Help" in that context?
For named-entity recognition task CoNLL-2003 NER



	Dev F1 Score
First Layer (Embedding)	91.0
Last Hidden Layer (12)	94.9
Sum All 12 Layers (12 + … + 2 + 1 =)	95.5
Second-to-Last Hidden Layer (11)	95.6
Sum Last Four Hidden (12 + 11 + 10 + 9 =)	95.9
Concat Last Four Hidden (9, 10, 11, 12)	96.1

fine-tune = 96.4

Concluding Remarks

Contextualized embeddings learned from masked LM via Transformers provide informative cues for **transfer learning**

BERT – a general approach for learning contextual representations from Transformers and benefiting language understanding

- Pre-trained BERT: https://github.com/google-research/bert