Applied Deep Learning

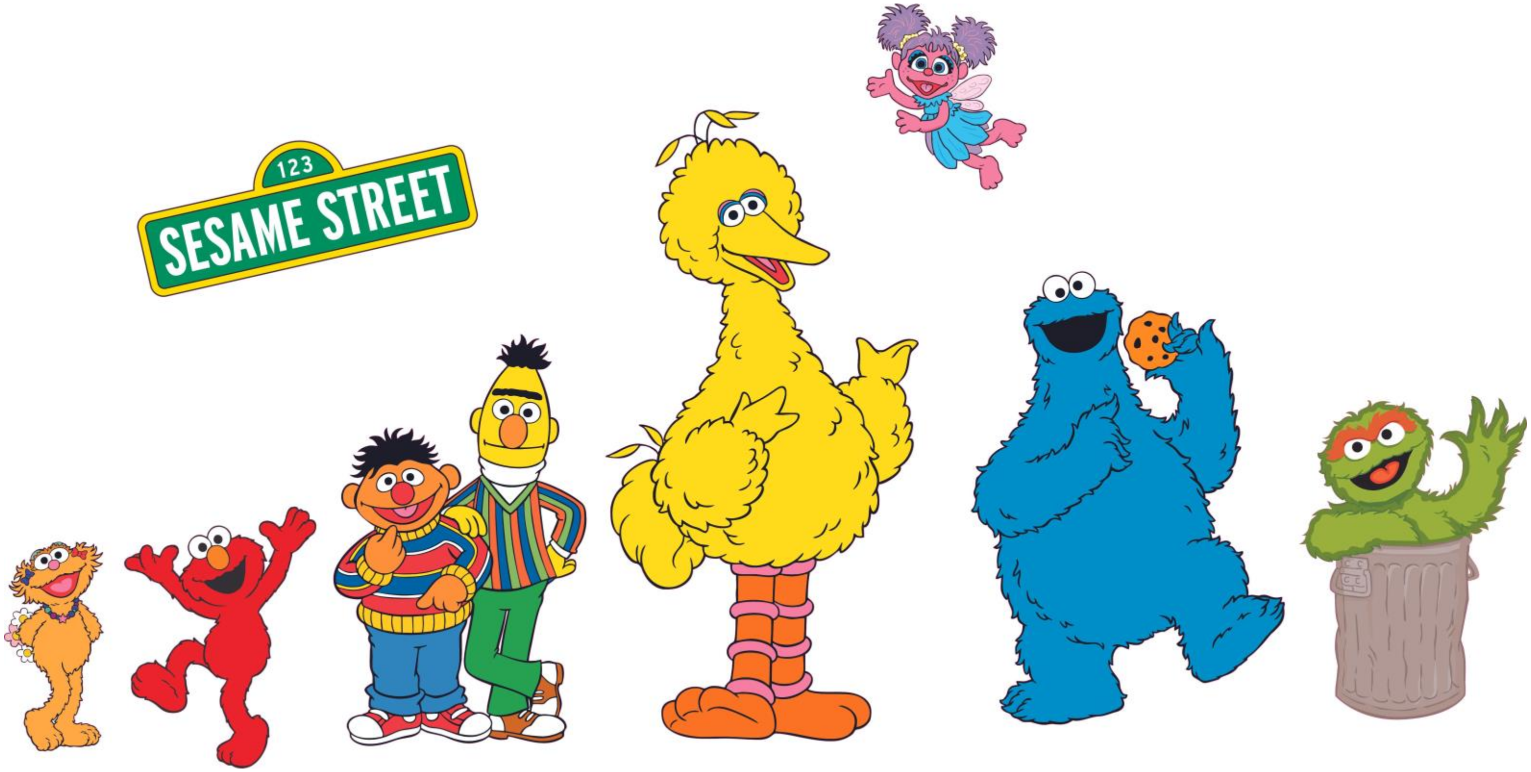
BERT

Bidirectional Encoder Representations from Transformers

October 13th, 2022 http://adl.miulab.tw

National
Taiwan
University
國立臺灣大學

Sesame Street



ELMO

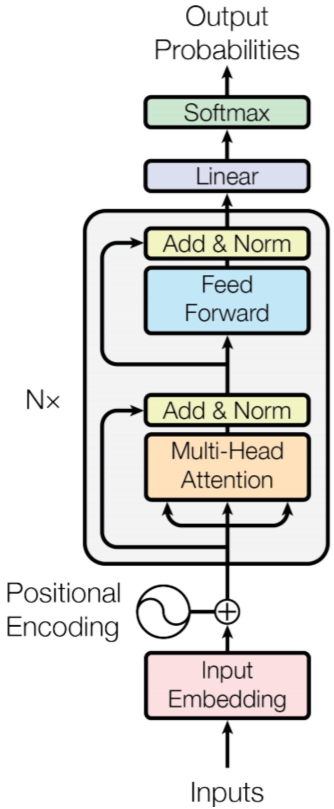
BERT

Picture from https://imagizer.imageshack.com/img924/8457/xhILHR.jpg

3

BERT: Bidirectional **E**ncoder **R**epresentations **from T**ransformers

- Idea: contextualized word representations
  - Learn word vectors using long contexts using Transformer instead of LSTM

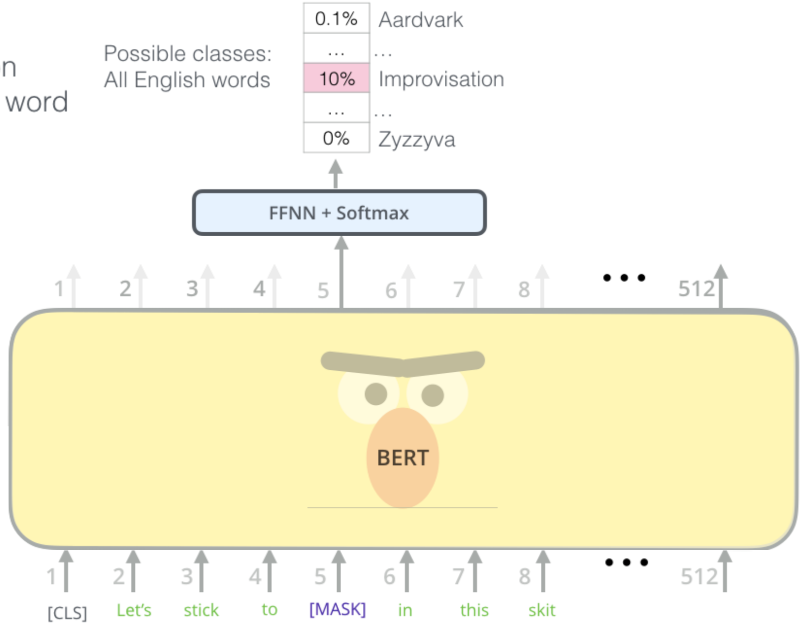


Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, in *NAACL-HLT*, 2019.

BERT #1 – Masked Language Model

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

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

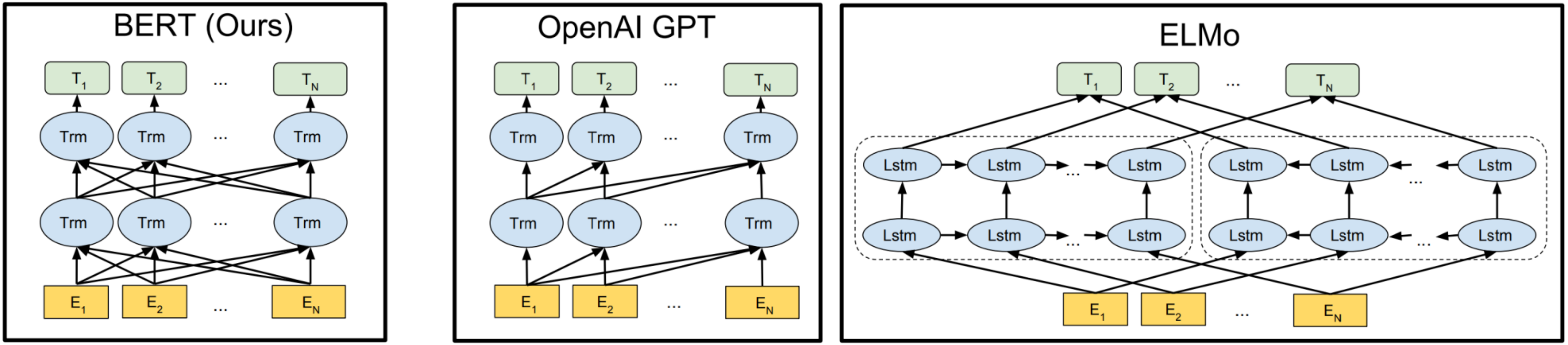


Randomly mask 15% of tokens

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

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

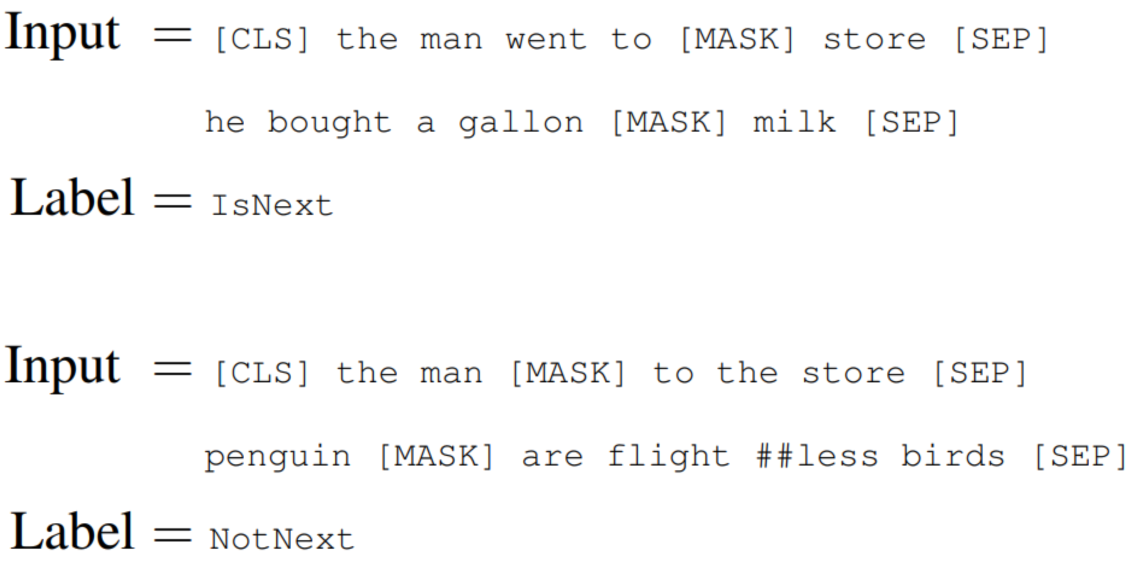
BERT #1 – Masked Language Model



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

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`

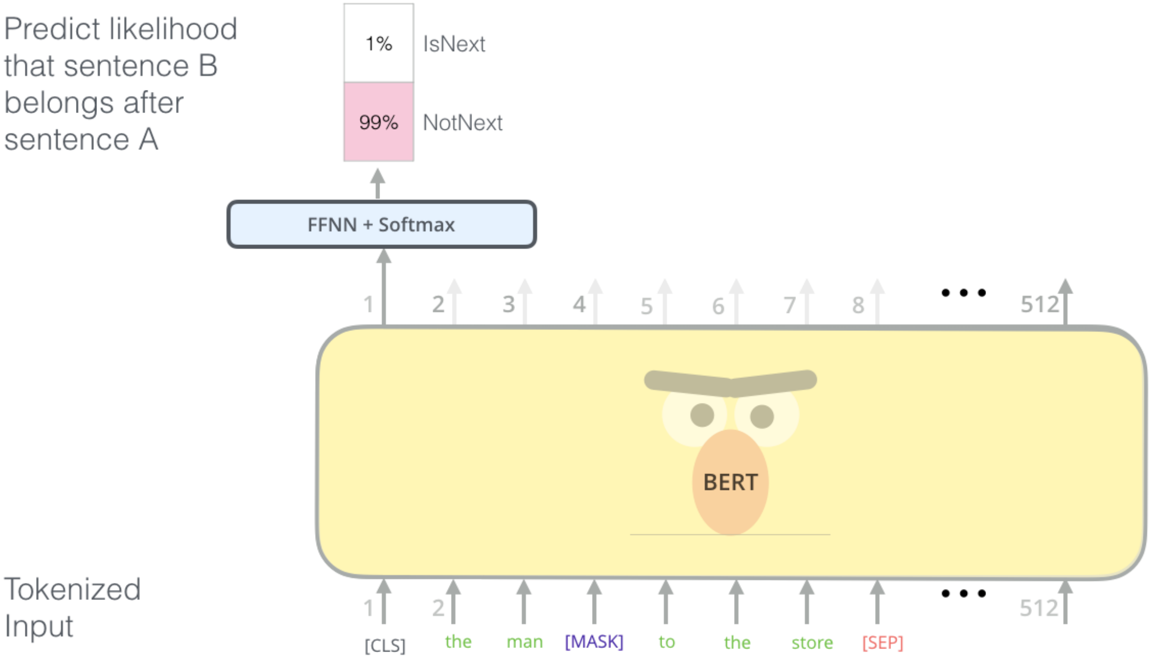
Input = `[CLS] the man [MASK] to the store [SEP]`
`penguin [MASK] are flight ##less birds [SEP]`

Label = `NotNext`

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

BERT #2 – Next Sentence Prediction

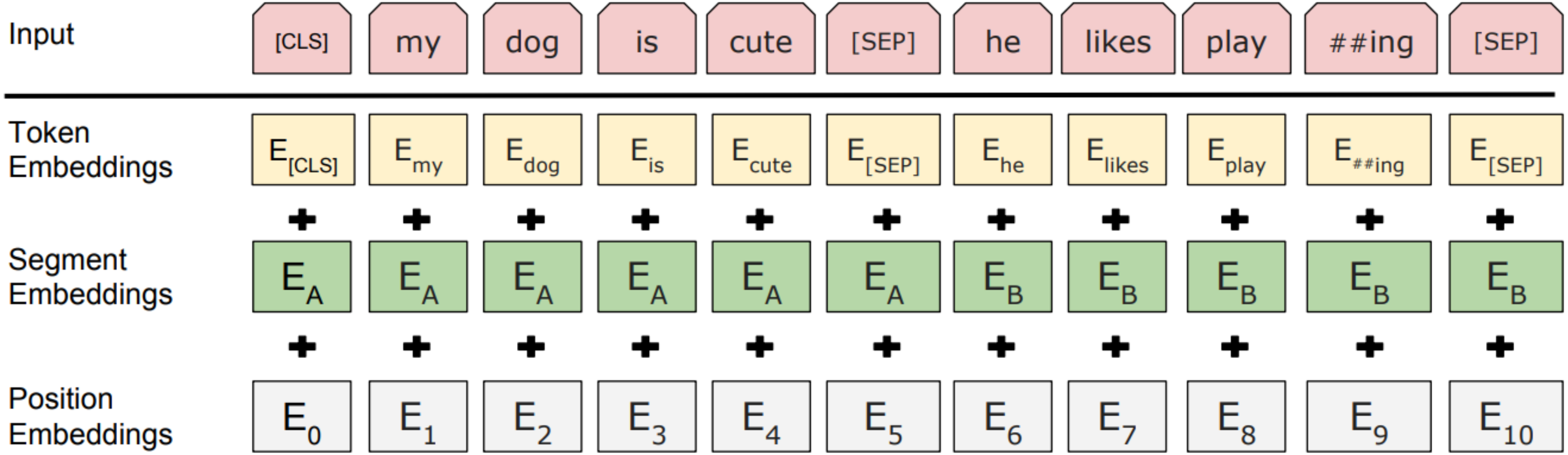
- Idea: modeling relationship between sentences



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

BERT – Input Representation

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

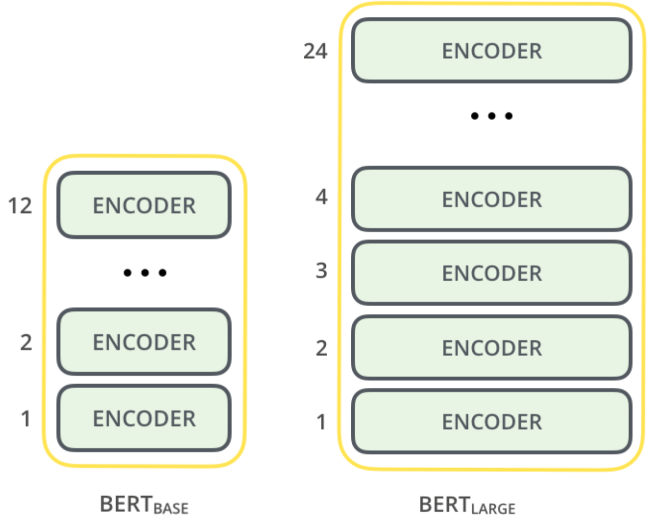


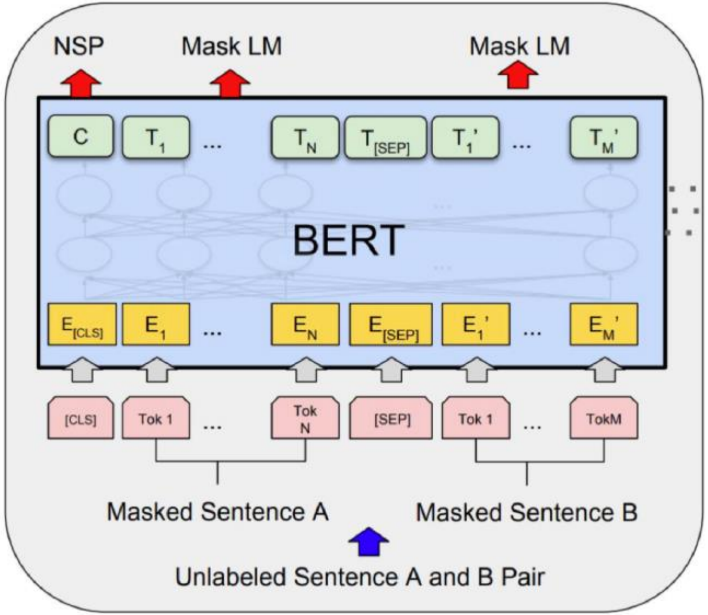
Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{\#\#ing}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Position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

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

BERT Training

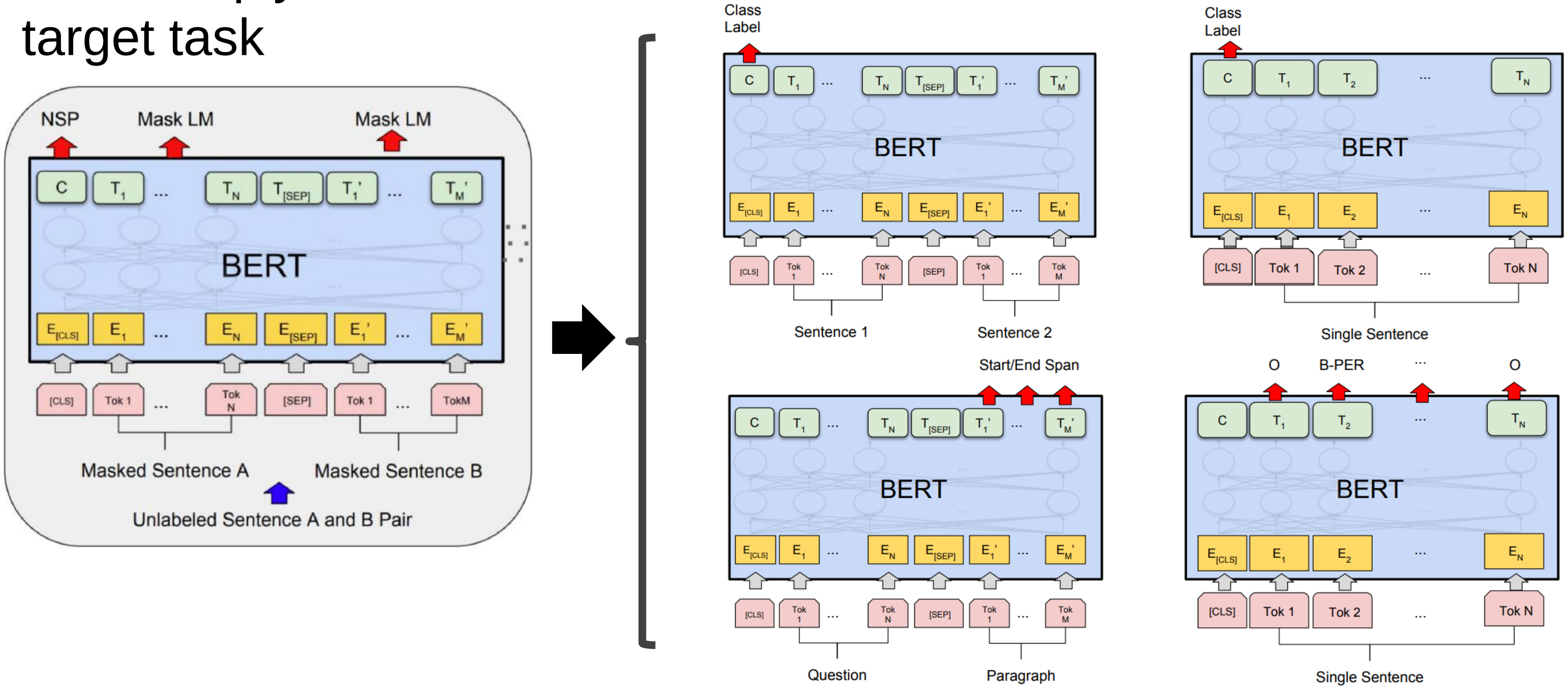
- Training data: Wikipedia + BookCorpus
- 2 BERT models
  - BERT-Base: 12-layer, 768-hidden, 12-head
  - BERT-Large: 24-layer, 1024-hidden, 16-head





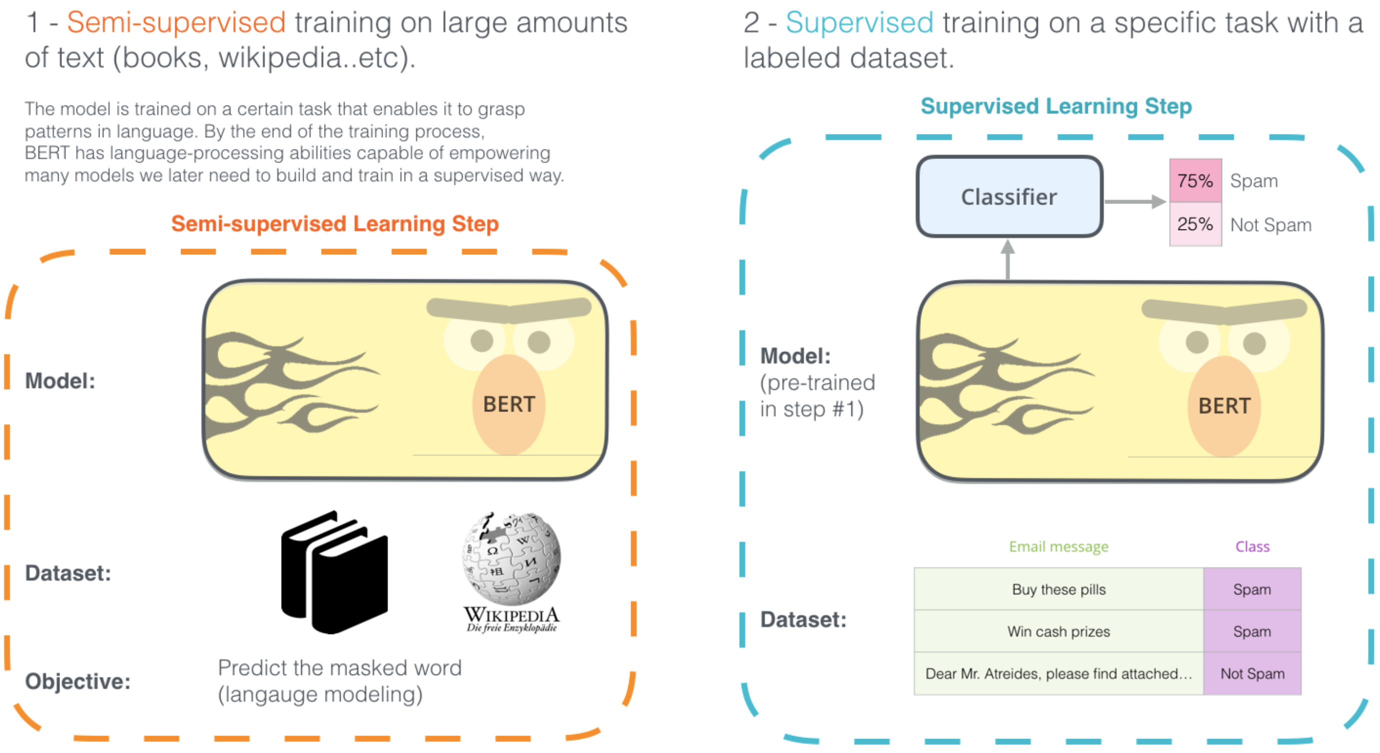
BERT Fine-Tuning for Understanding Tasks

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



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

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

Model:

Dataset:

Objective: Predict the masked word (langauge modeling)

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

Supervised Learning Step

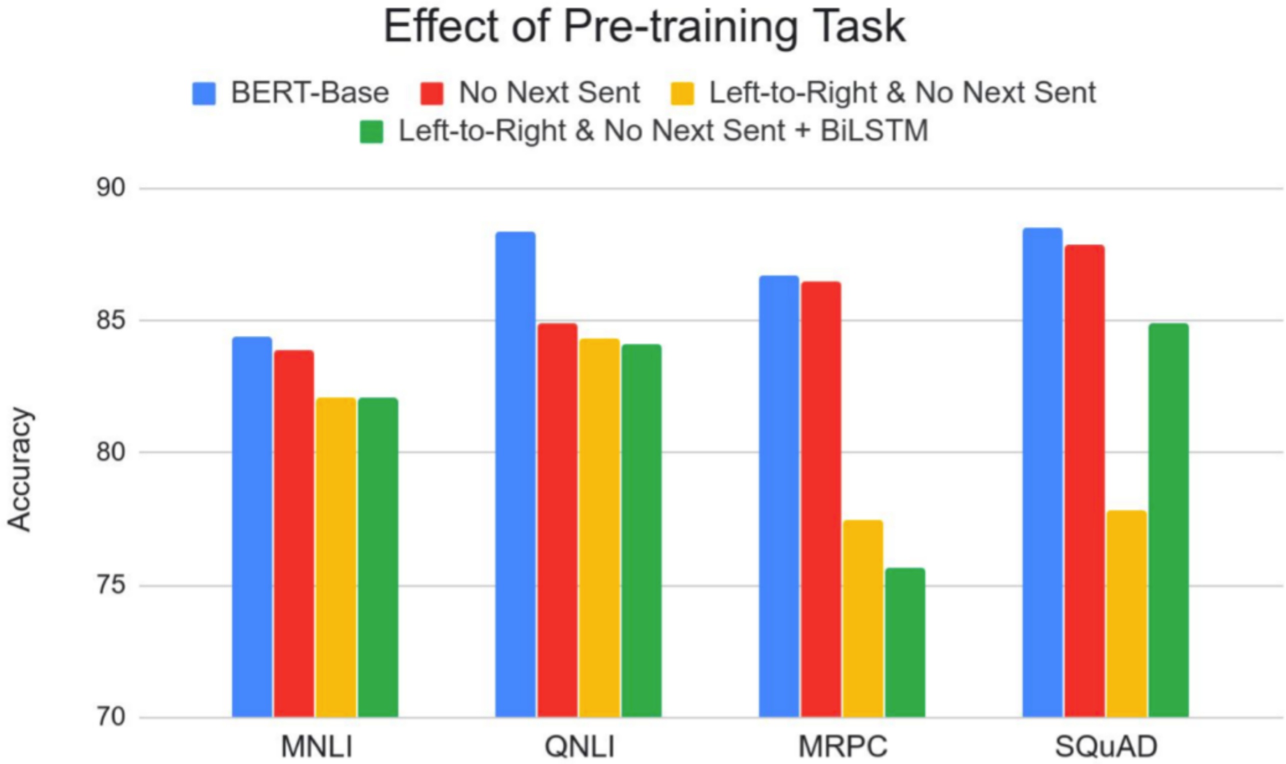
Model: (pre-trained in step #1)

Dataset:

Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

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

BERT Fine-Tuning Results



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

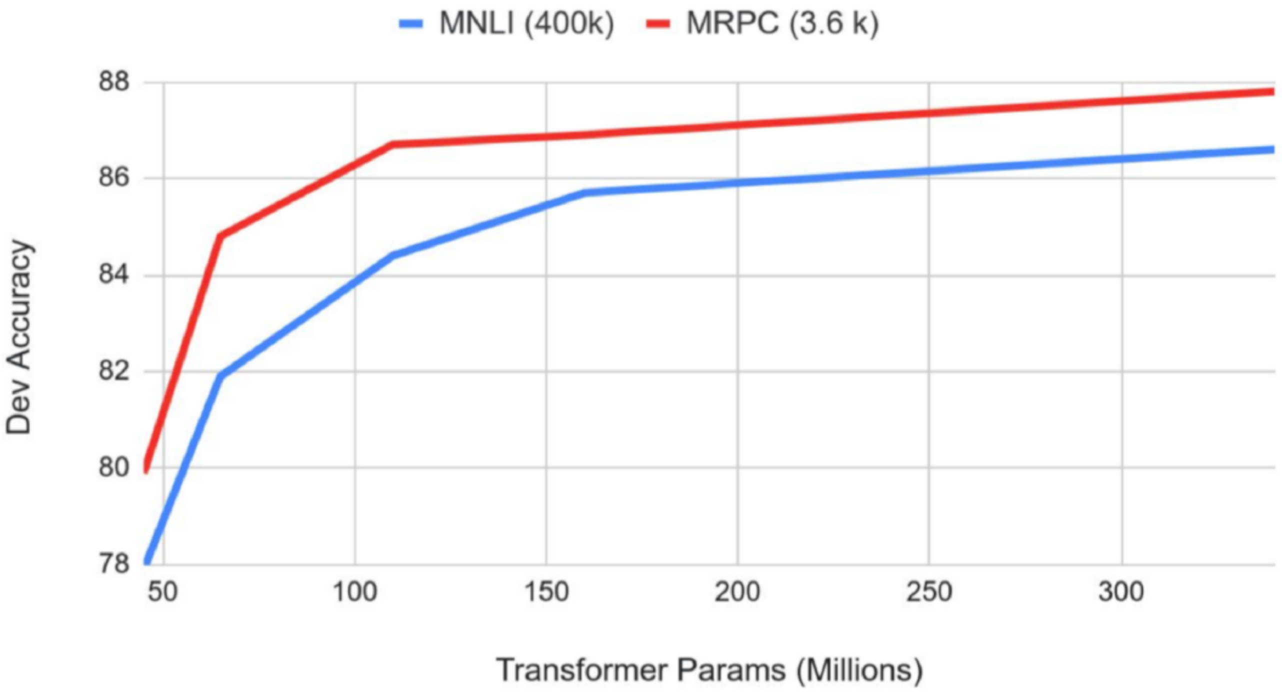
BERT Results on NER

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 LM + fine-tune	92.4
CVT Clark	Cross-view training + multitask learn	92.61
BERT-Large (Devlin+, 2019)	Transformer 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

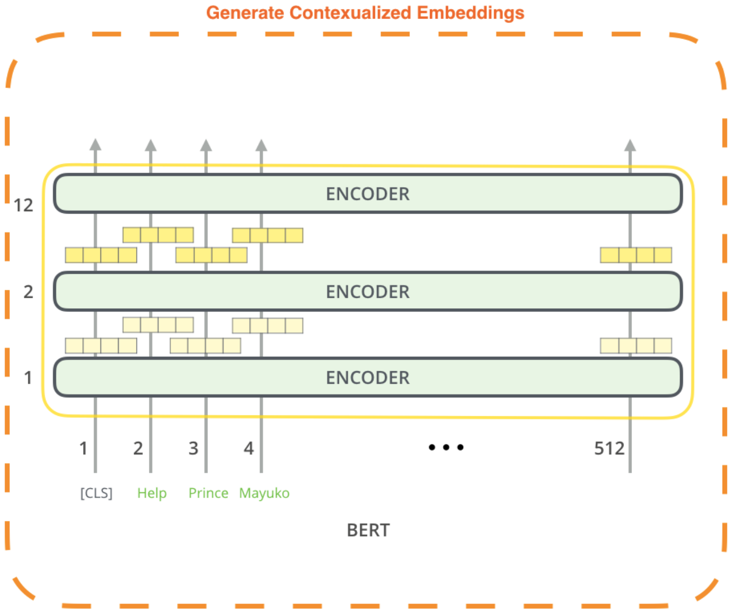
- Improving performance by increasing model size



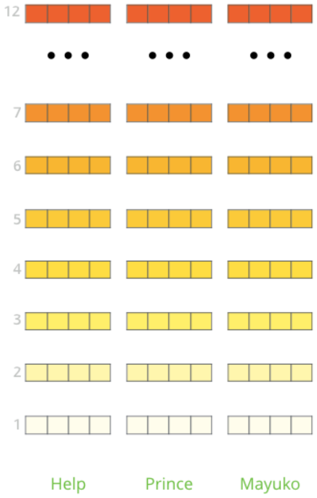
Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, in *NAACL-HLT*, 2019.

BERT for Contextual Embeddings

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



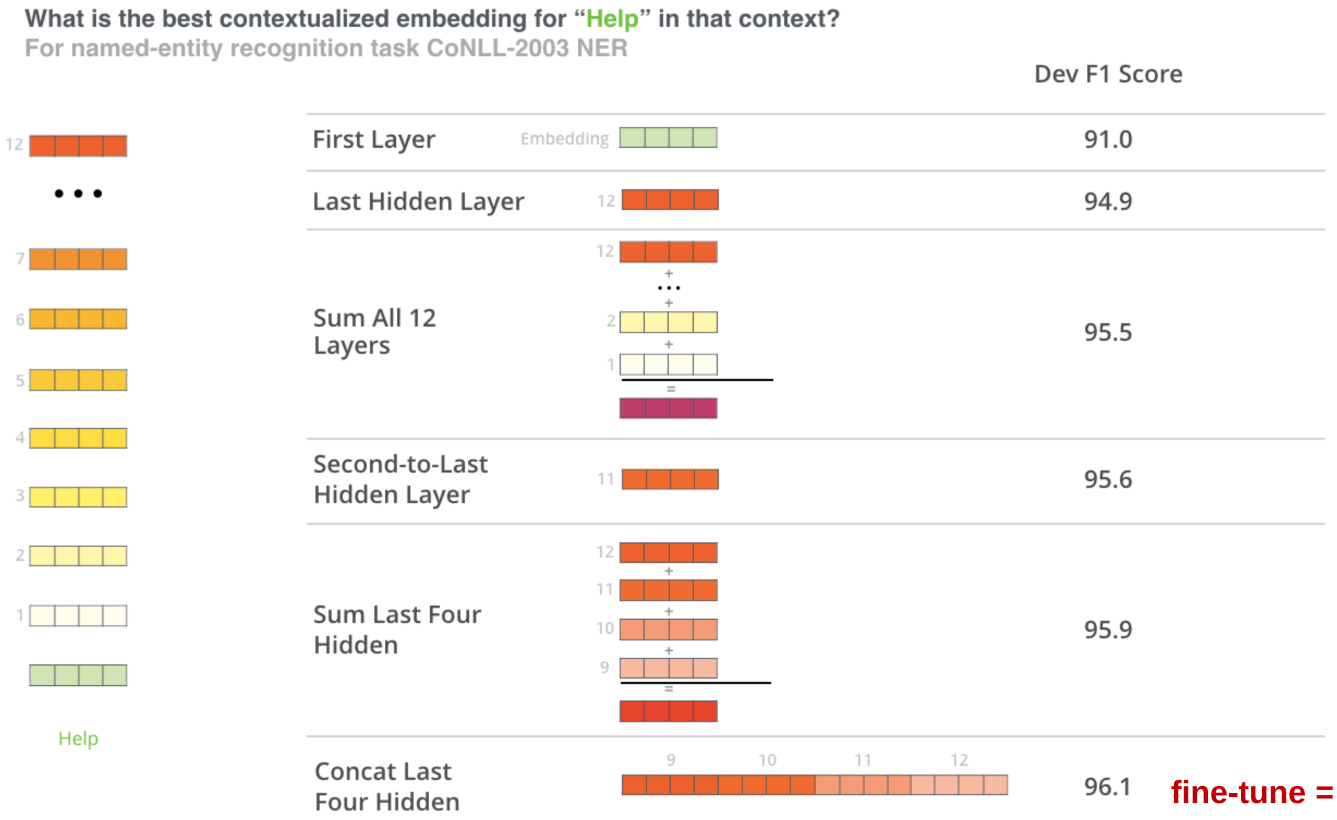
The output of each encoder layer along each token's path can be used as a feature representing that token.



But which one should we use?

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

BERT Contextual Embeddings Results on NER



What is the best contextualized embedding for “Help” in that context?

For named-entity recognition task CoNLL-2003 NER

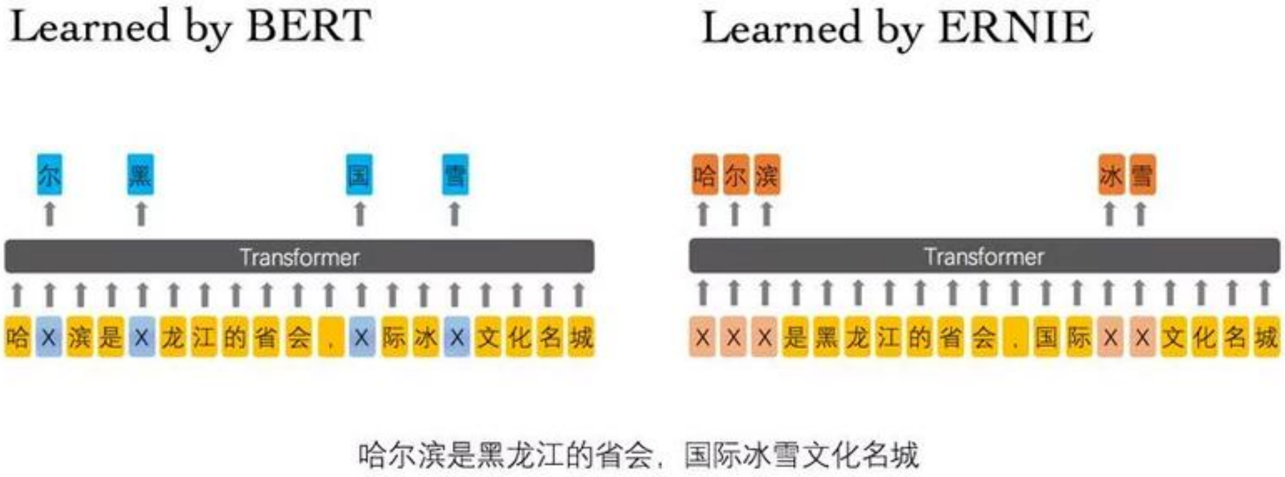
	Dev F1 Score
First Layer	91.0
Last Hidden Layer	94.9
Sum All 12 Layers	95.5
Second-to-Last Hidden Layer	95.6
Sum Last Four Hidden	95.9
Concat Last Four Hidden	96.1

fine-tune = 96.4

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

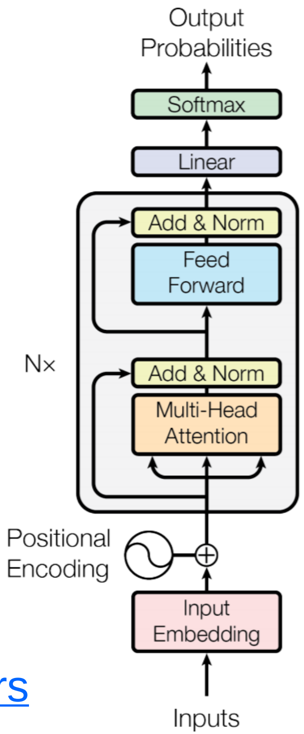
ERNIE: **E**nhanced **R**epresentation through k**N**owledge **I**nt**E**gration

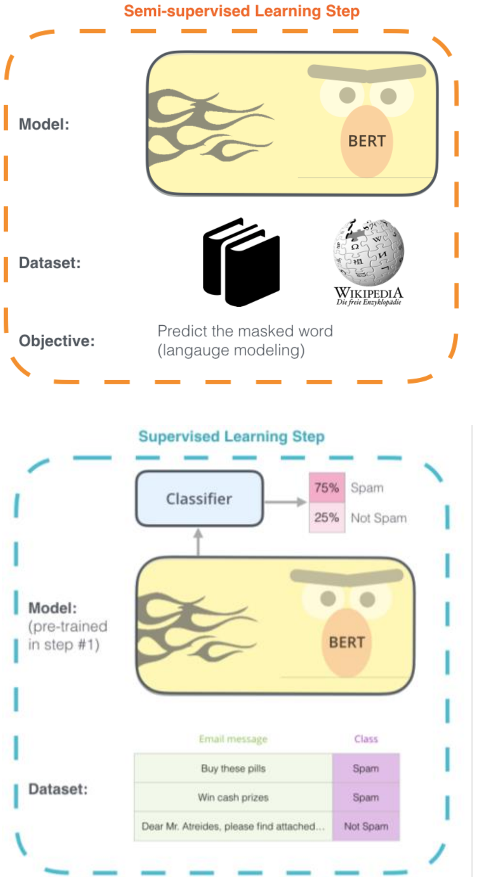
- BERT models local cooccurrence between tokens, while characters are modeled independently
  - 哈(ha), 爾(er), 濱(bin) instead 哈爾濱(Harbin)
- ERNIE incorporates knowledge by masking semantic units/entities


Learned by BERT
尔 黑 国 雪
Transformer
哈 X 滨 是 X 龙 江 的 省 会 ， X 际 冰 X 文 化 名 城
Learned by ERNIE
哈 尔 滨 冰 雪
Transformer
X X X 是 黑 龙 江 的 省 会 ， 国 际 X X 文 化 名 城
哈尔滨是黑龙江的省会，国际冰雪文化名城

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
    https://github.com/huggingface/transformers





Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached...	Not Spam