BERT
Bidirectional Transformers for Language Understanding

Dr Mehreen Alam

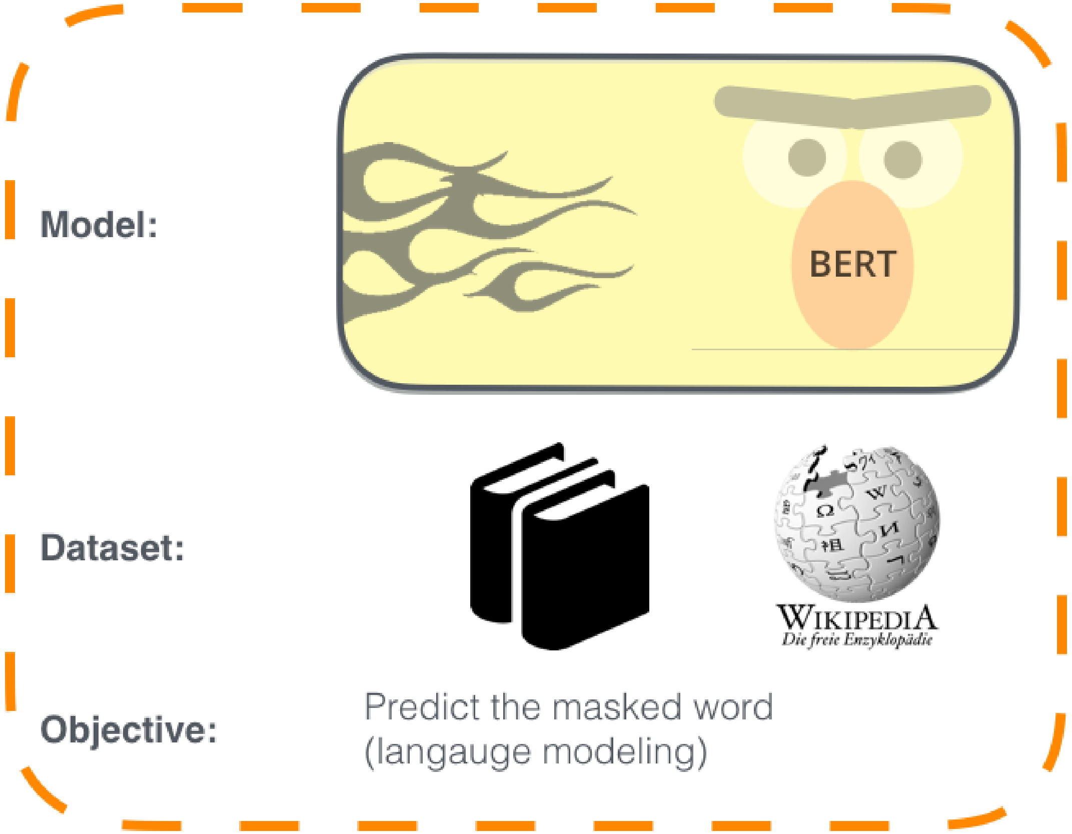
Contents

- Limitations of Sequential Techniques
- Intro
- Architecture
- Results
- Useful Links

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



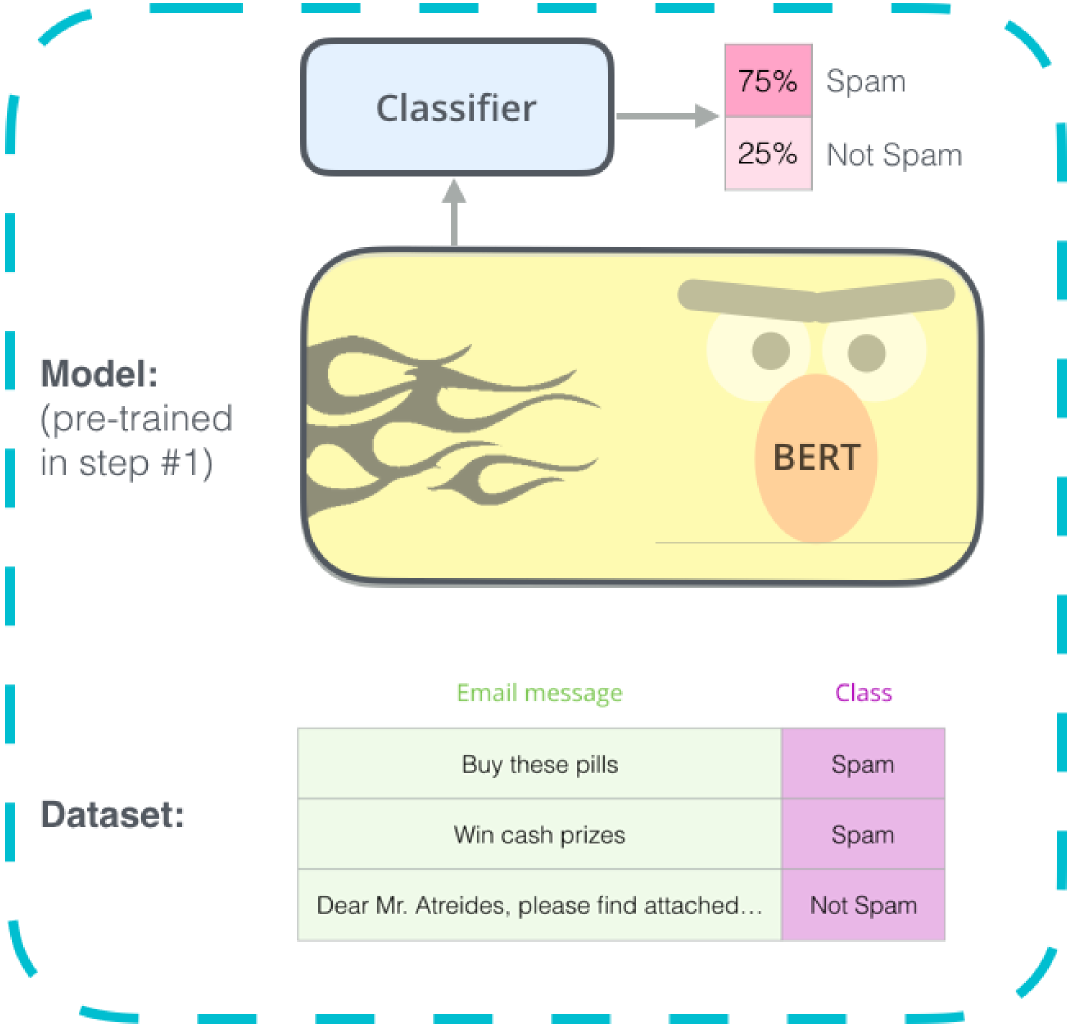
Model:

Dataset:

Objective: Predict the masked word (langauge modeling)

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

Supervised Learning Step



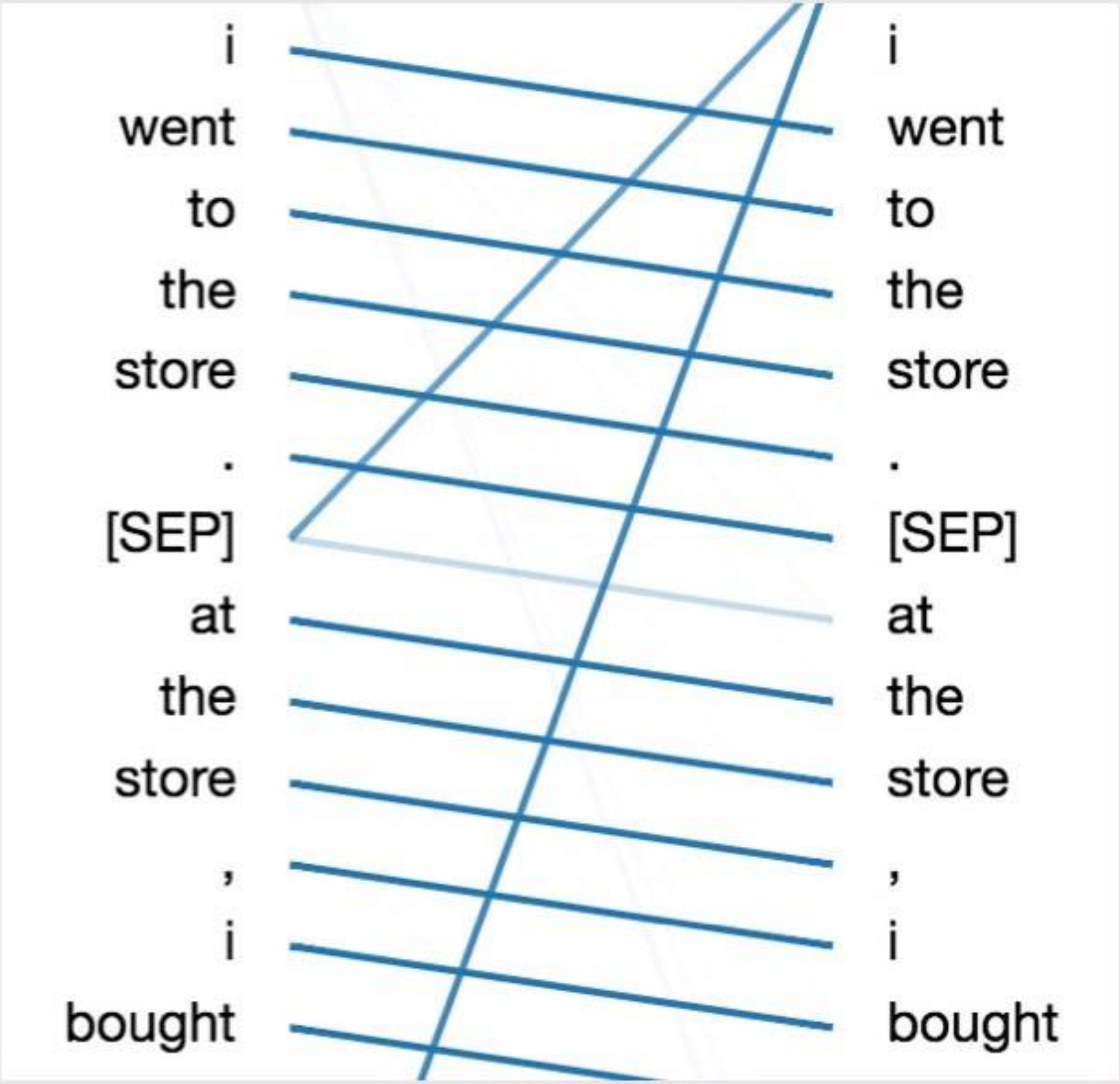
Model: (pre-trained in step #1)

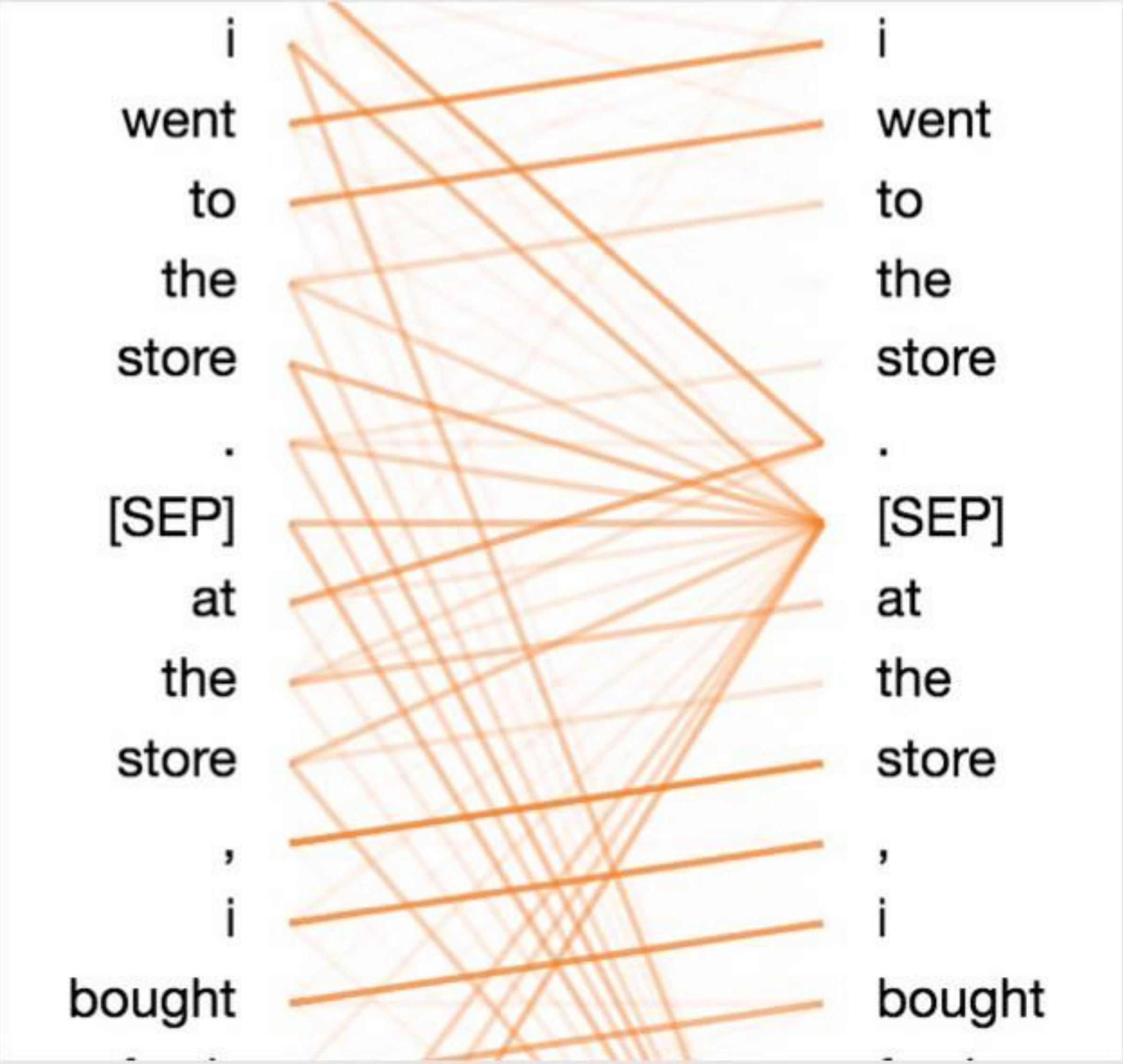
Dataset:

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

Uses:

- Sentiment analysis
  - Input: Movie/Product review. Output: is the review positive or negative?
  - Example dataset: SST (Stanford Sentiment Treebank)
- Fact-checking
  - Input: sentence. Output: “Claim” or “Not Claim”
- Sequence-to-sequence based language generation tasks such as:
  - Question answering
  - Abstract summarization
  - Sentence prediction
  - Conversational response generation
- Natural language understanding tasks such as:
  - Polysemy
  - Word sense disambiguation
  - Natural language inference


i
went
to
the
store
.
[SEP]
at
the
store
,
i
bought
i
went
to
the
store
.
[SEP]
at
the
store
,
i
bought


i
went
to
the
store
.
[SEP]
at
the
store
,
i
bought
i
went
to
the
store
.
[SEP]
at
the
store
,
i
bought

Examples

- Sentence Classification

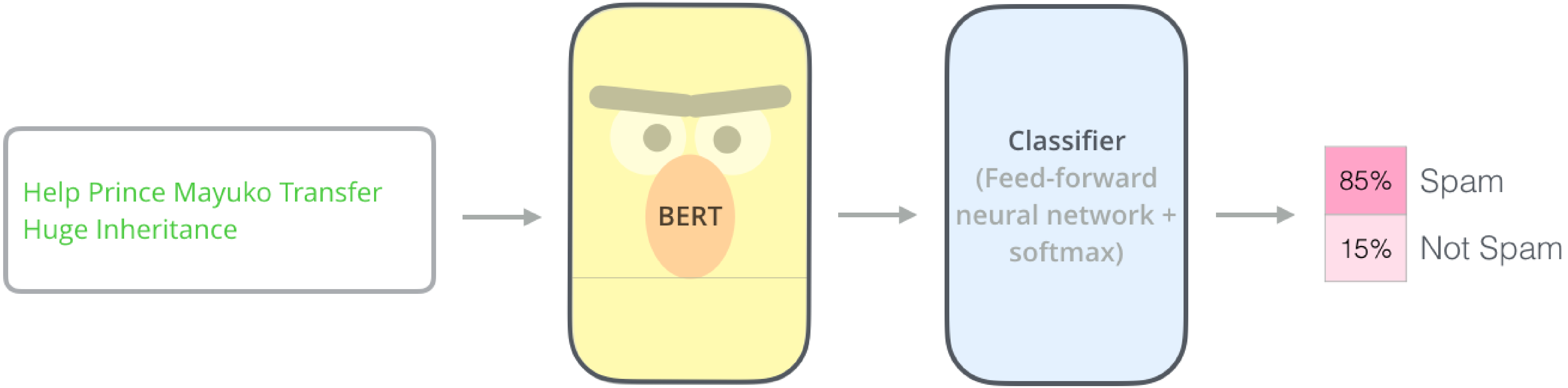
Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

Input

Features

Output

Prediction



Help Prince Mayuko Transfer Huge Inheritance

BERT

Classifier (Feed-forward neural network + softmax)

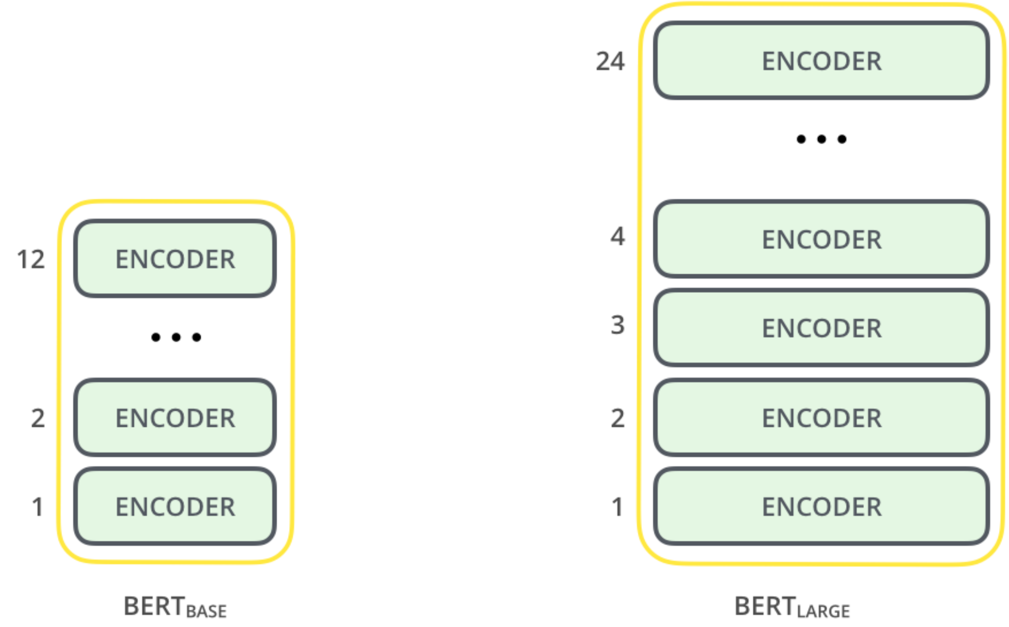
85%	Spam
15%	Not Spam

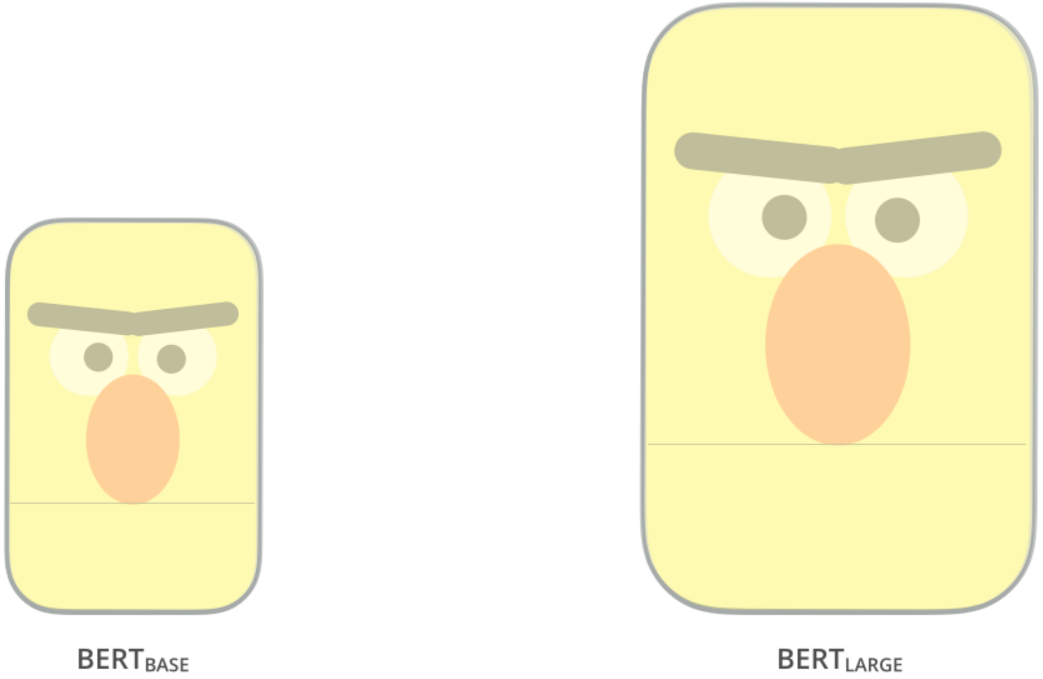
Bert base vs large

- Just Encoder:
  - Transformer Blocks: 12 vs 24
  - feedforward-networks: 768 and 1024 hidden units
  - attention heads: 12 and 16 respectively) than the default configuration in the reference implementation of the Transformer in the initial paper (6 encoder layers, 512 hidden units, and 8 attention heads).
  - 110 M vs 345M

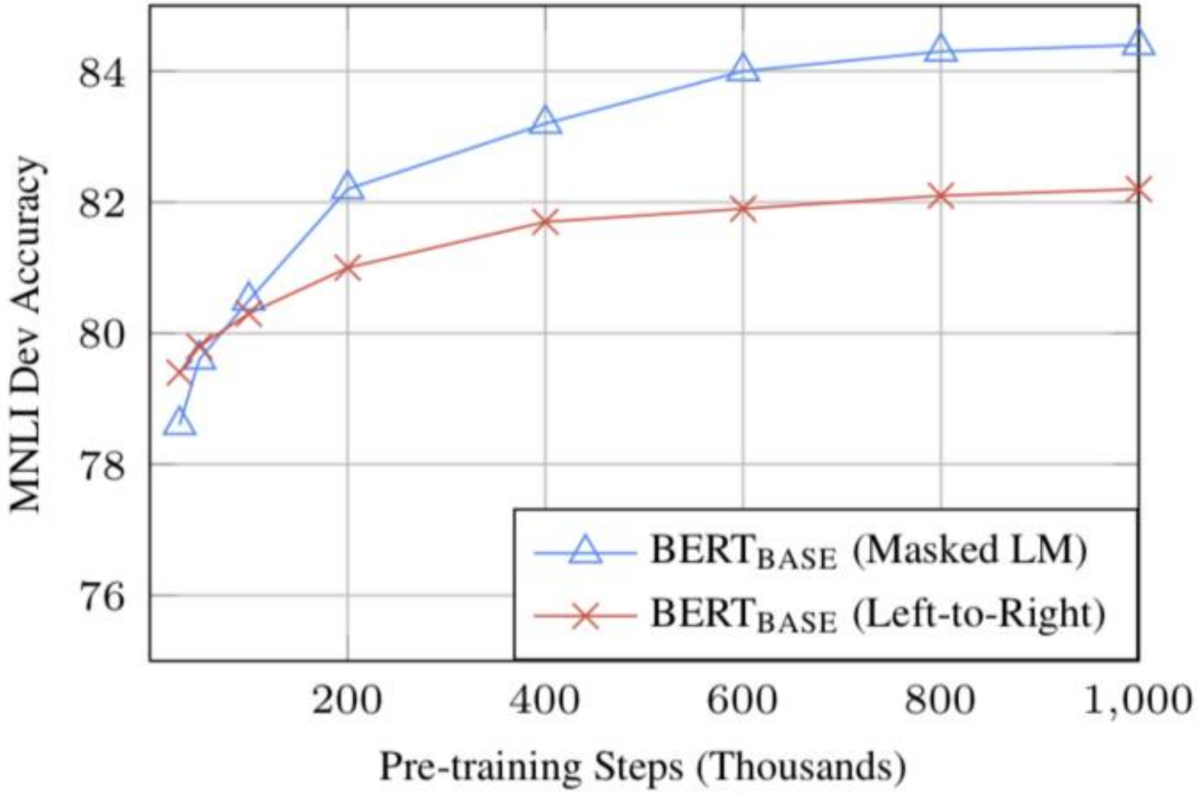
Architecture

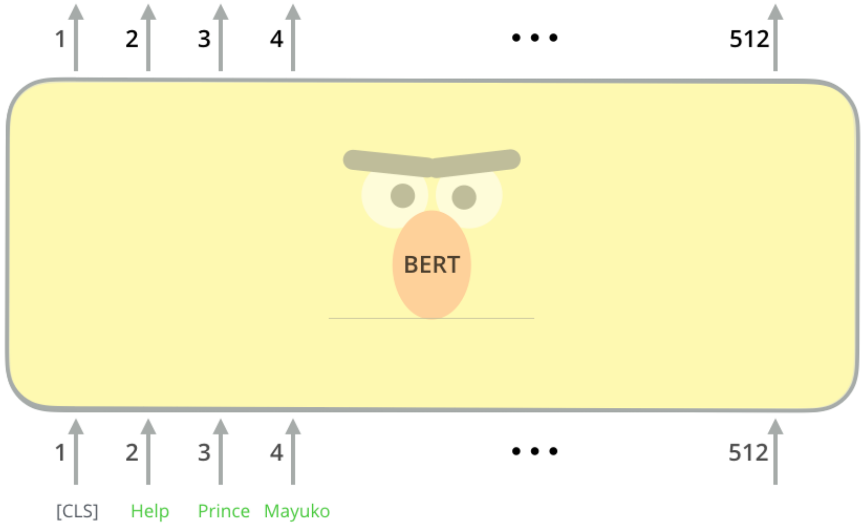
- BERT BASE – Comparable in size to the OpenAI Transformer in order to compare performance
- BERT LARGE – A ridiculously huge model which achieved the state-of-the-art results reported in the paper

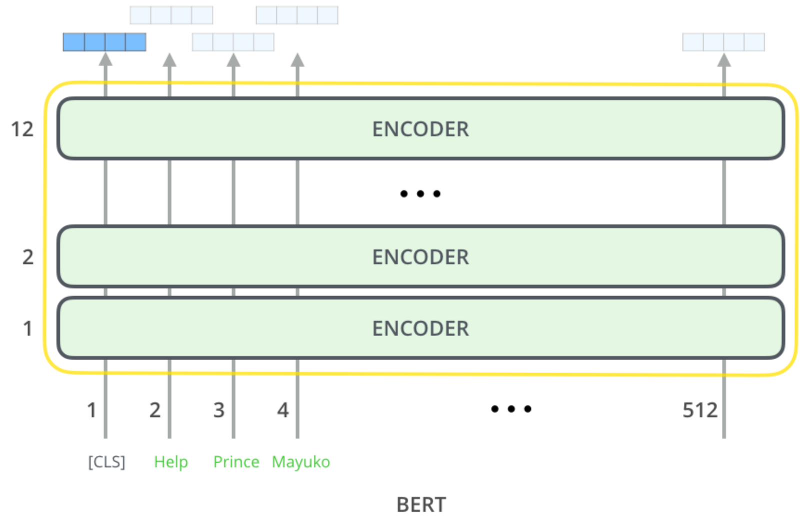


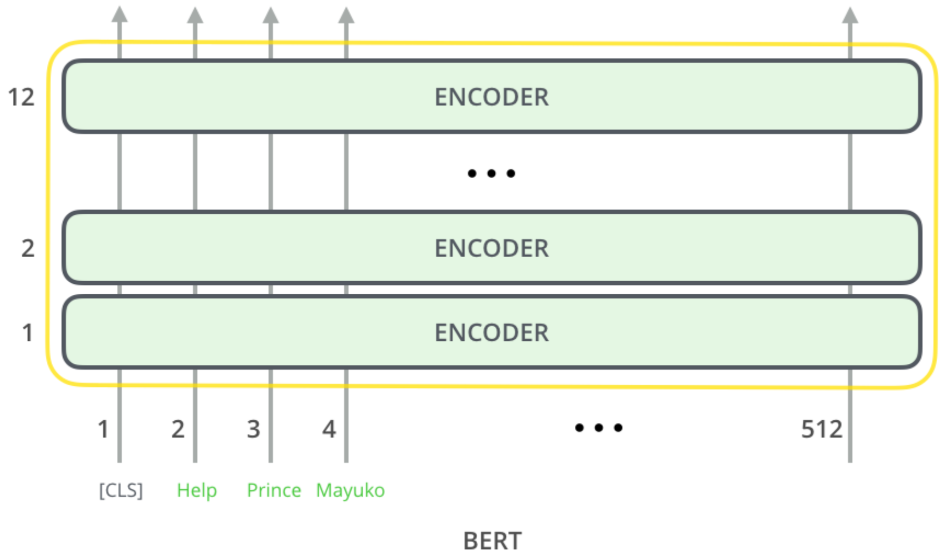


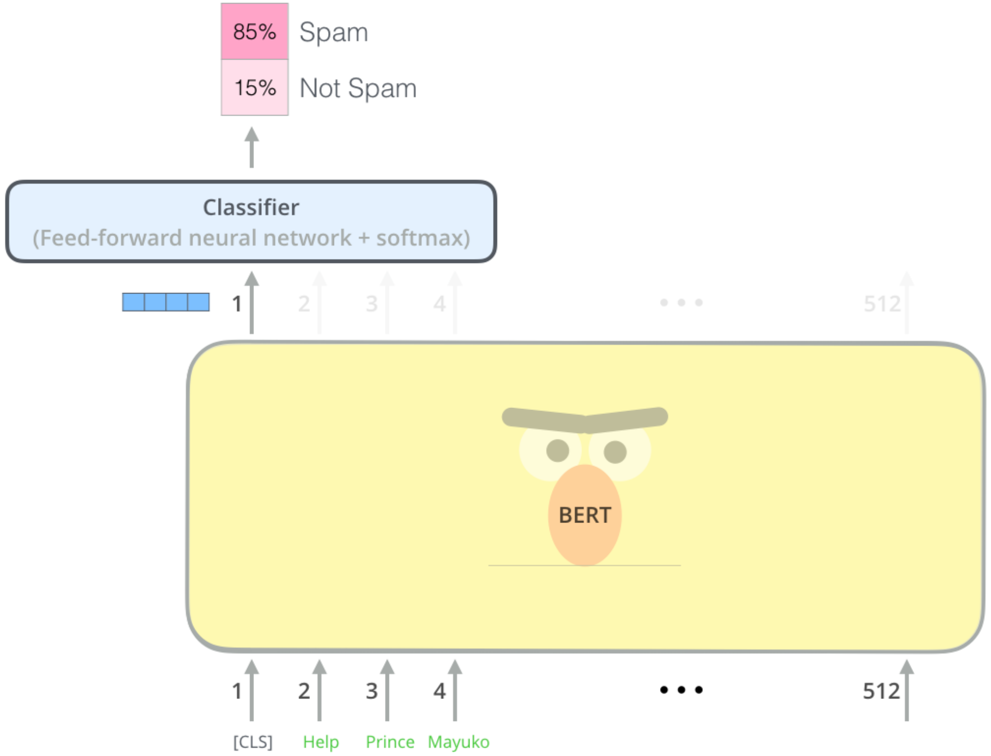
	Training Compute + Time	Usage Compute
BERT$_{BASE}$	4 Cloud TPUs, 4 days	1 GPU
BERT$_{LARGE}$	16 Cloud TPUs, 4 days	1 TPU




1
2
3
4
• • •
512
BERT
1
2
3
4
• • •
512
[CLS]
Help
Prince
Mayuko

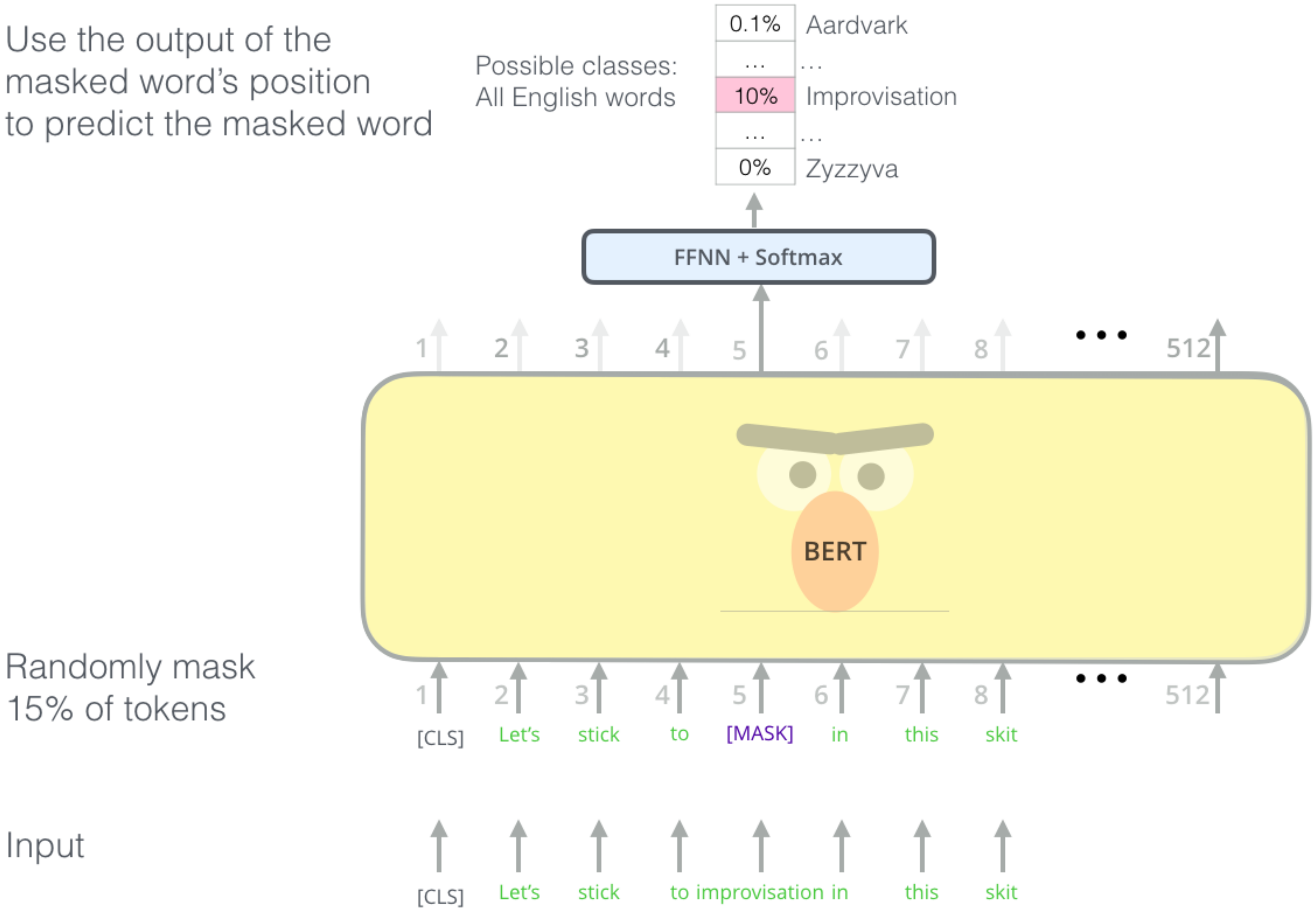

12
ENCODER
• • •
2
ENCODER
1
ENCODER
1
2
3
4
• • •
512
[CLS]
Help
Prince
Mayuko
BERT


12
ENCODER
• • •
2
ENCODER
1
ENCODER
1
2
3
4
• • •
512
[CLS]
Help
Prince
Mayuko
BERT


85%
Spam
15%
Not Spam
Classifier
(Feed-forward neural network + softmax)
1
2
3
4
• • •
512
BERT
1
2
3
4
• • •
512
[CLS]
Help
Prince
Mayuko

Bert: From Decoders to Encoders

- MLM



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

Possible classes: All English words

0.1%	Aardvark
…	…
10%	Improvisation
…	…
0%	Zyzzyva

FFNN + Softmax

Randomly mask 15% of tokens

[CLS] Let's stick to [MASK] in this skit

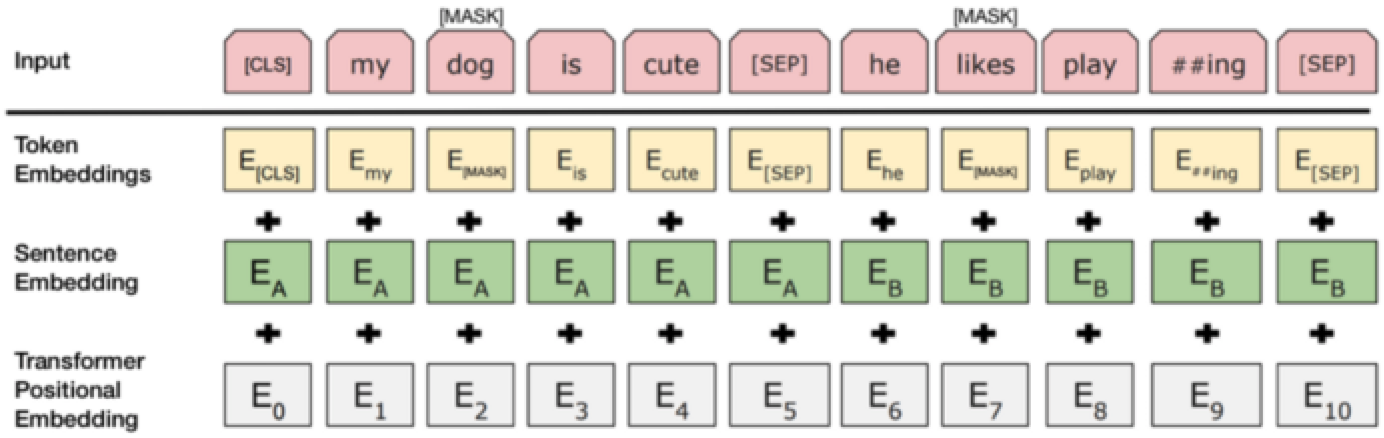
Input

[CLS] Let's stick to improvisation in this skit

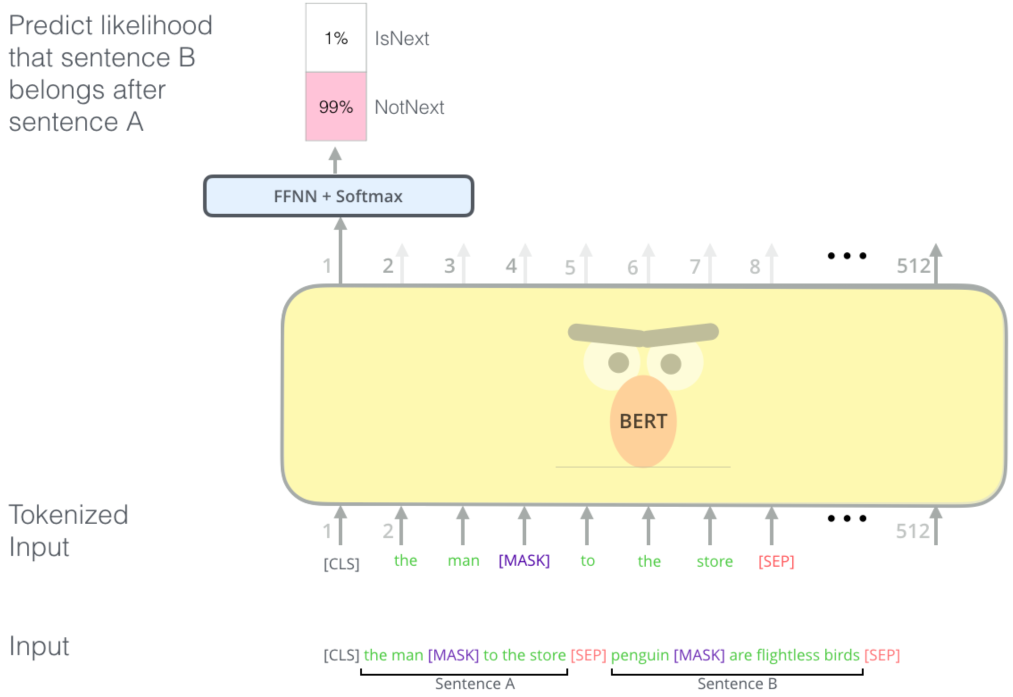
- MLM
    - Training the language model in BERT is done by predicting 15% of the tokens in the input, that were randomly picked.
        - 80% are replaced with a “[MASK]” token, 10% with a random word, and 10% use the original word.
- Other Masking Options:
    - [MASK] 100%:
        - model wouldn’t necessarily produce good token representations for non-masked words. The non-masked tokens were still used for context, but the model was optimized for predicting masked words.
    - [MASK] 90% and random words 10% of the time:
        - this would teach the model that the observed word is never correct.
    - [MASK] 90% and kept the same word 10% of the time:
        - then the model could just trivially copy the non-contextual embedding.

Bert: From Decoders to Encoders

- Two Sentence Task



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



Predict likelihood that sentence B belongs after sentence A

1%	IsNext
99%	NotNext

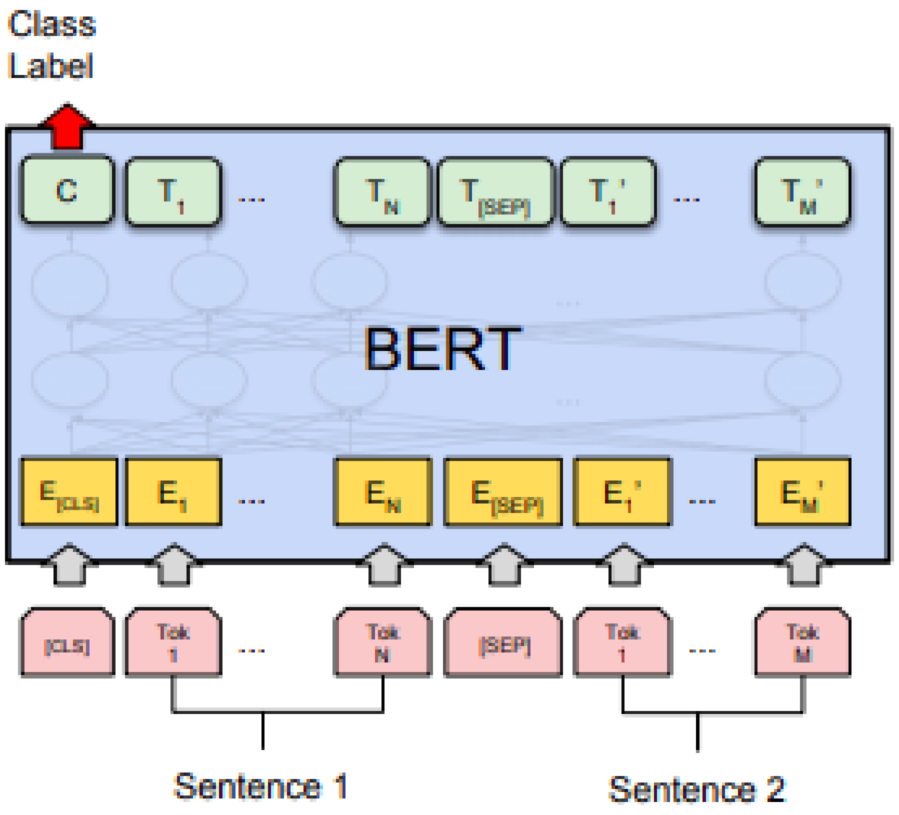
FFNN + Softmax

Tokenized Input

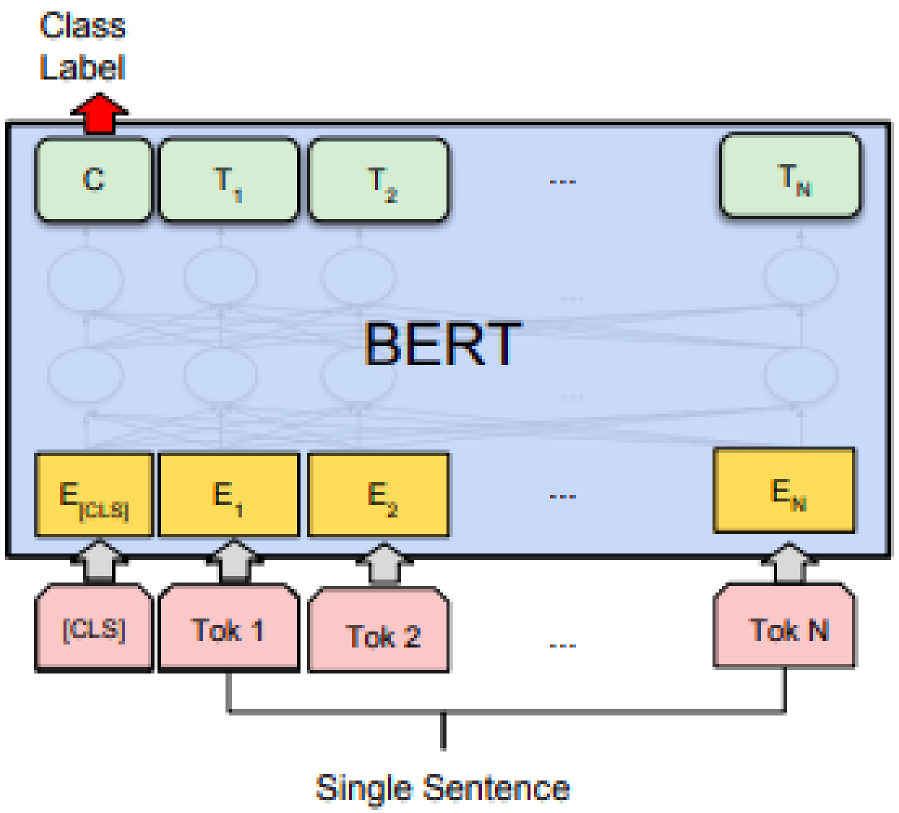
Input

[CLS] the man [MASK] to the store [SEP] penguin [MASK] are flightless birds [SEP]

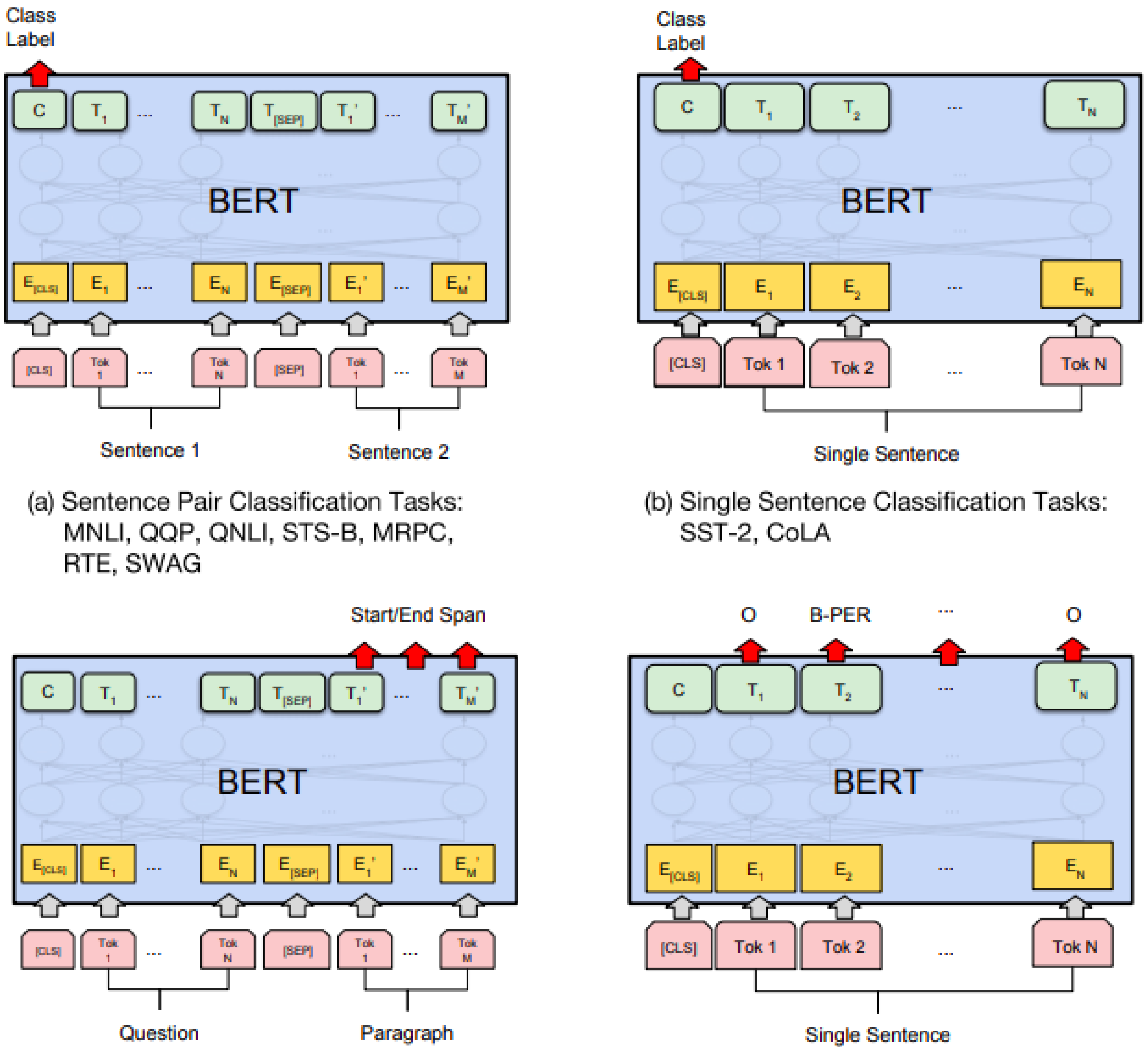
Task Specific Models



(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG



(b) Single Sentence Classification Tasks:
SST-2, CoLA



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Bert Variants

- patentBERT: perform patent classification.
- docBERT: 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: for scientific text
- 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.

Results

- state-of-the-art results on eleven natural language processing tasks:
    - GLUE score to 80.5% (7.7% point absolute improvement),
    - MultiNLI accuracy to 86.7% (4.6% absolute improvement),
    - SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement), and
    - SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

THE
TRANSFORMER

OpenAI
Transformer

ULM-FiT

ELMo

BERT

To be noted

- BERT is a model with absolute position embeddings, so it’s usually advised to pad the inputs on the right rather than the left.
- BERT was trained with the masked language modeling (MLM) and next sentence prediction (NSP) objectives. It is efficient at predicting masked tokens and at NLU in general, but is not optimal for text generation.
- Model size matters, even at huge scale. BERT_large (345 million parameters) demonstrably superior, on small-scale tasks to BERT_base (“only” 110 million parameters).
- With enough training data, more training steps == higher accuracy.
  - on the MNLI task, the BERT_base accuracy improves by 1.0% when trained on 1M steps (128,000 words batch size) compared to 500K steps with the same batch size.
- BERT’s bidirectional approach (MLM) converges slower than left-to-right approaches:
  - (because only 15% of words are predicted in each batch) but bidirectional training still outperforms left-to-right training after a small number of pre-training steps.

Useful Links

- https://jalammar.github.io/illustrated-bert/
- https://searchenterpriseai.techtarget.com/definition/BERT-language-model
- https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270
- https://www.youtube.com/watch?v=xI0HHN5XKDo
- https://towardsml.com/2019/09/17/bert-explained-a-complete-guide-with-theory-and-tutorial/
- https://towardsml.com/2019/09/17/bert-explained-a-complete-guide-with-theory-and-tutorial/
- https://huggingface.co/transformers/model_doc/bert.html
- https://github.com/google-research/bert
- Link to Original Paper:
  - https://arxiv.org/pdf/1810.04805.pdf