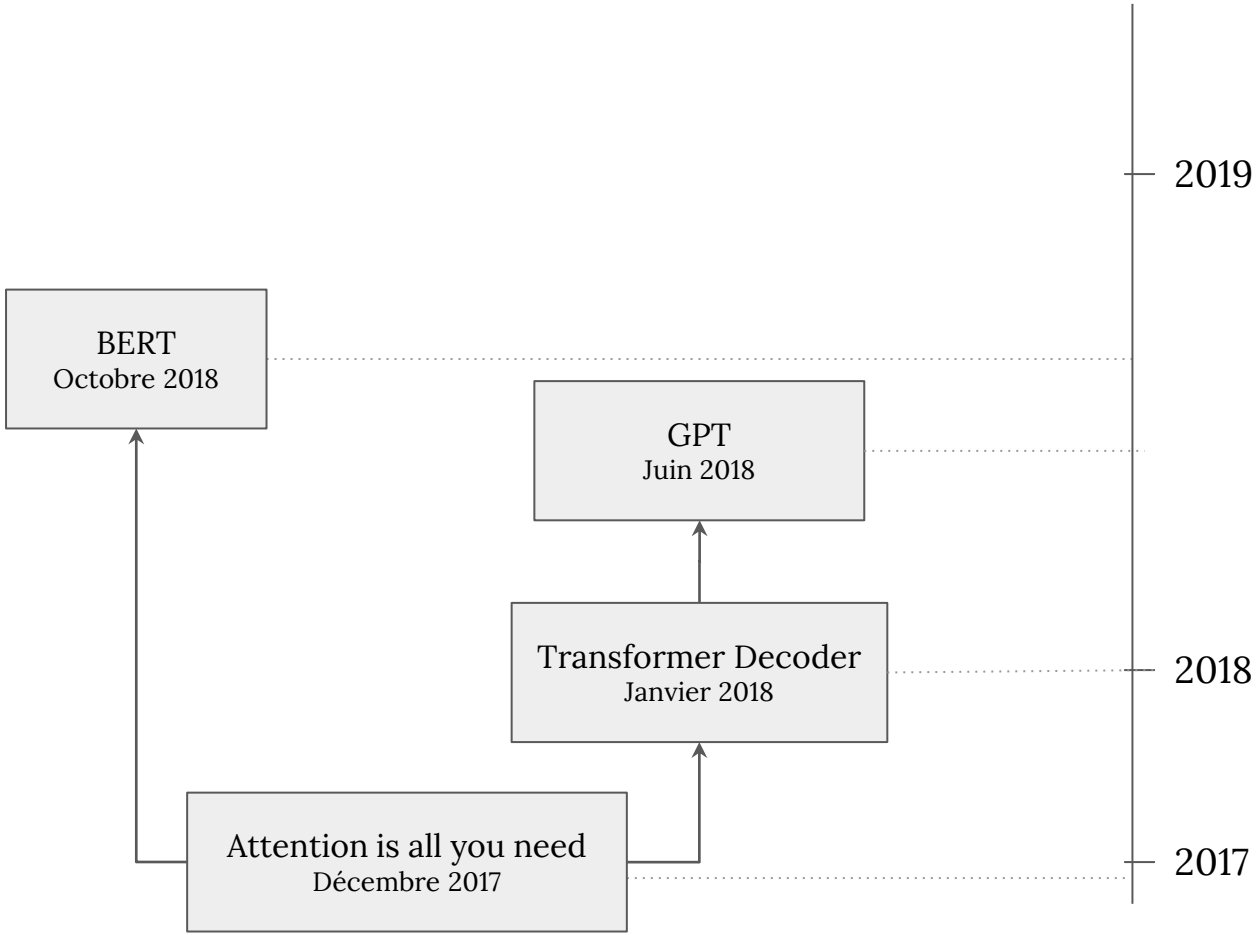
# BERT

Pre-training of Deep Bidirectional Transformers for language understanding


2019
BERT
Octobre 2018
GPT
Juin 2018
Transformer Decoder
Janvier 2018
2018
Attention is all you need
Décembre 2017
2017

# Outline

- Introduction of 2 new pre-training tasks
- Switches from Transformer-Decoder to *Transformer-**Encoder***

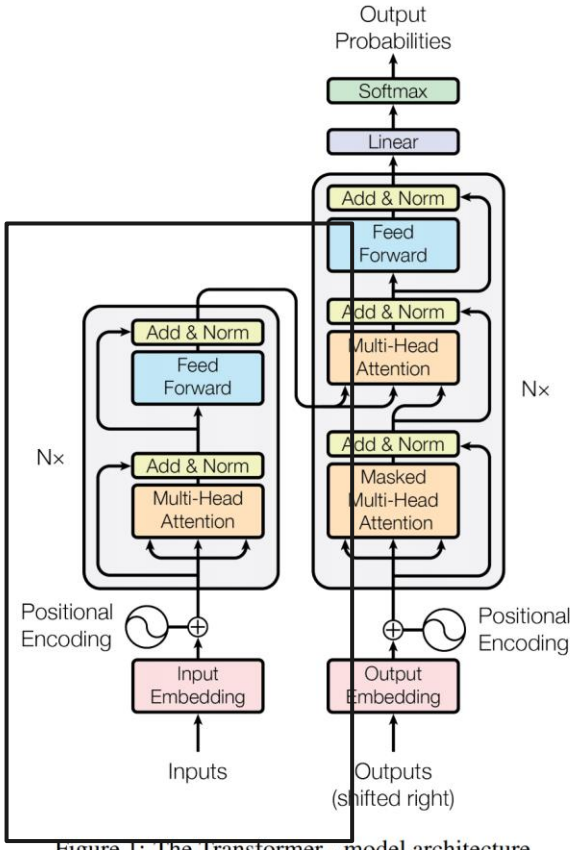



Figure 1: The Transformer - model architecture.

BERT$_{BASE}$

12X

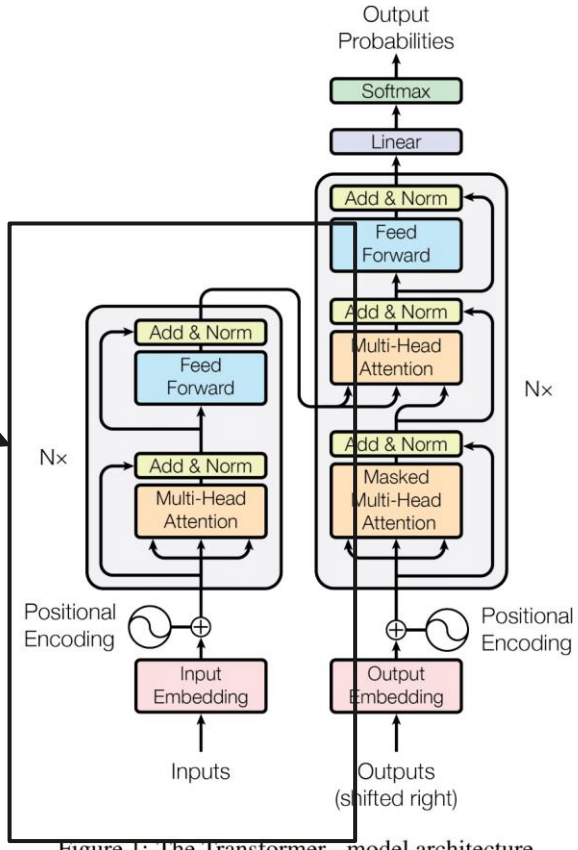



Figure 1: The Transformer - model architecture.

BERT$_{LARGE}$

24X

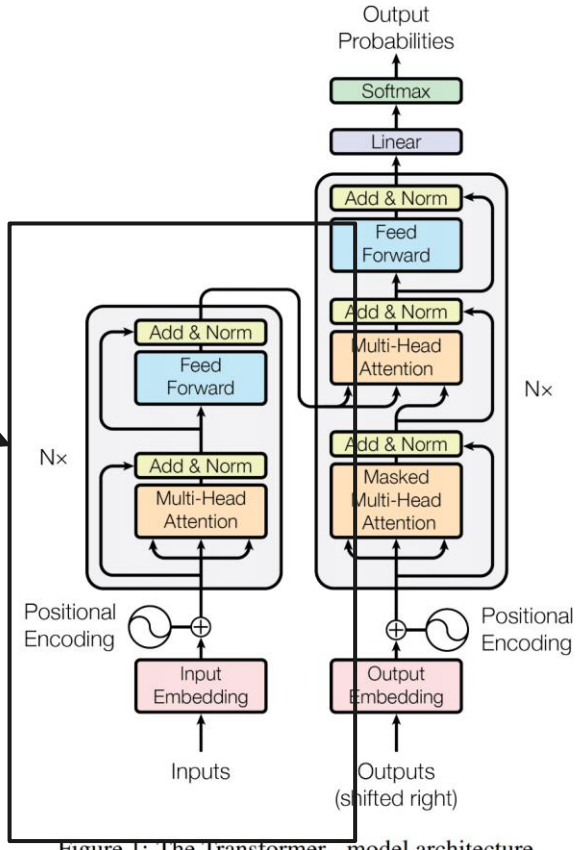
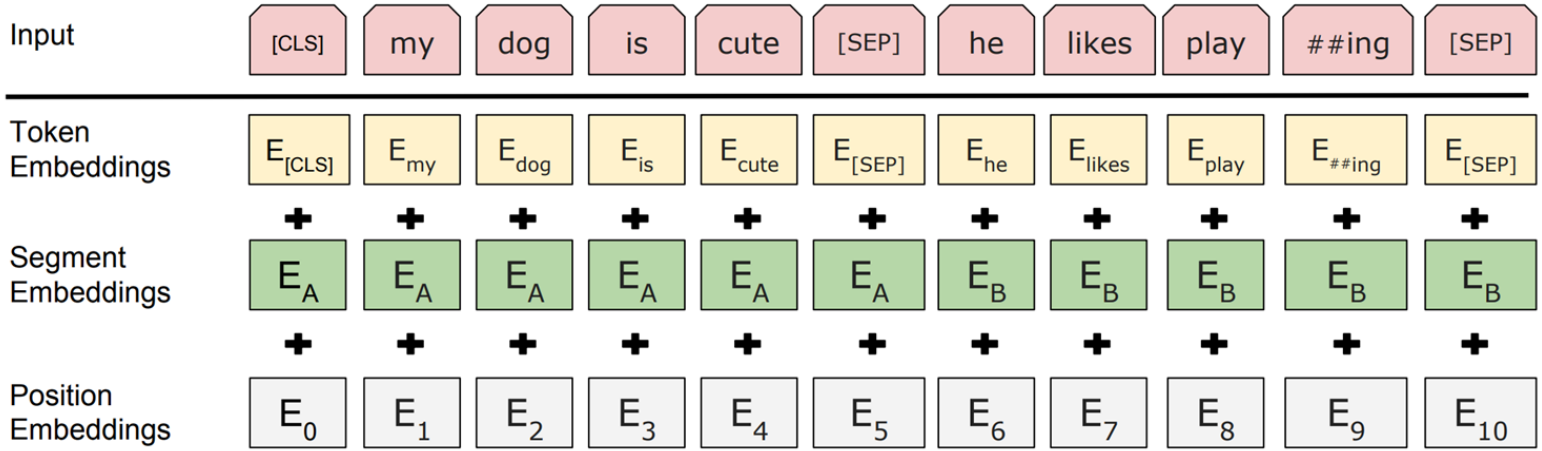


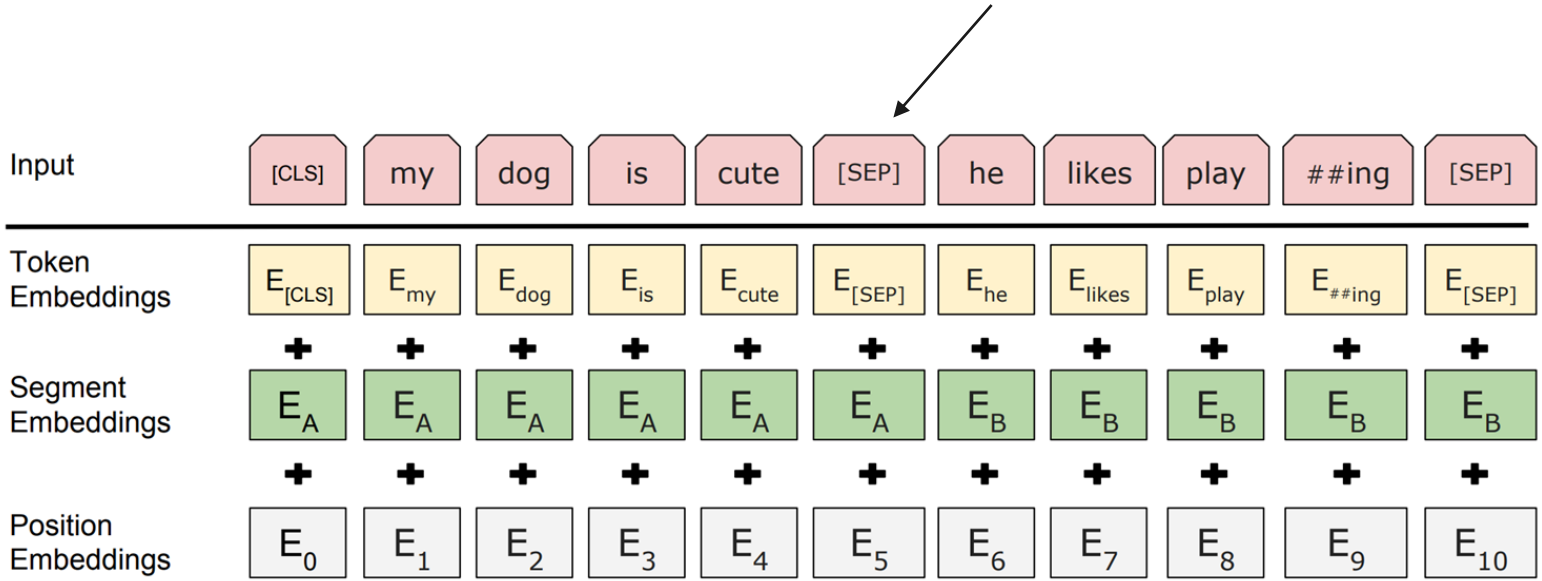

Figure 1: The Transformer - model architecture.

# Input of BERT

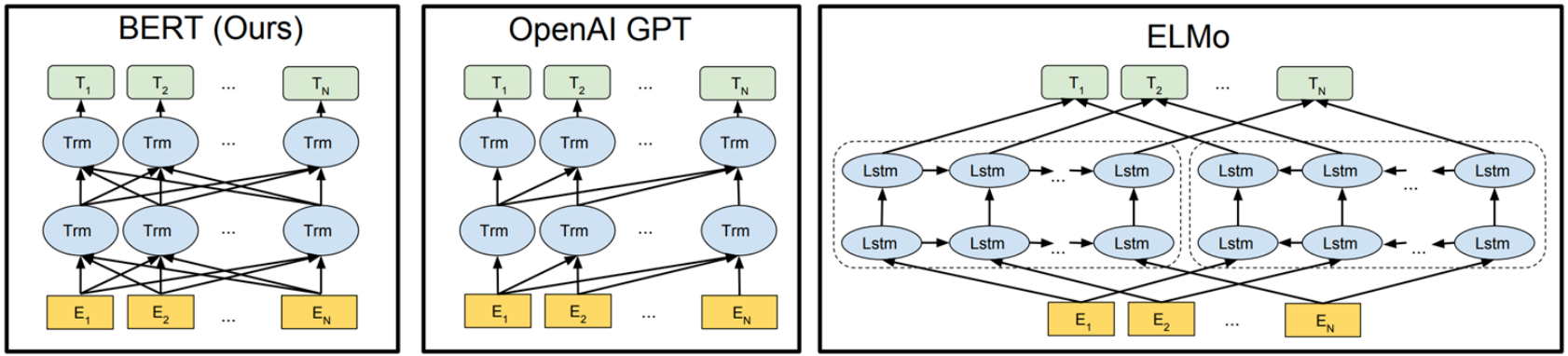


# Input of BERT

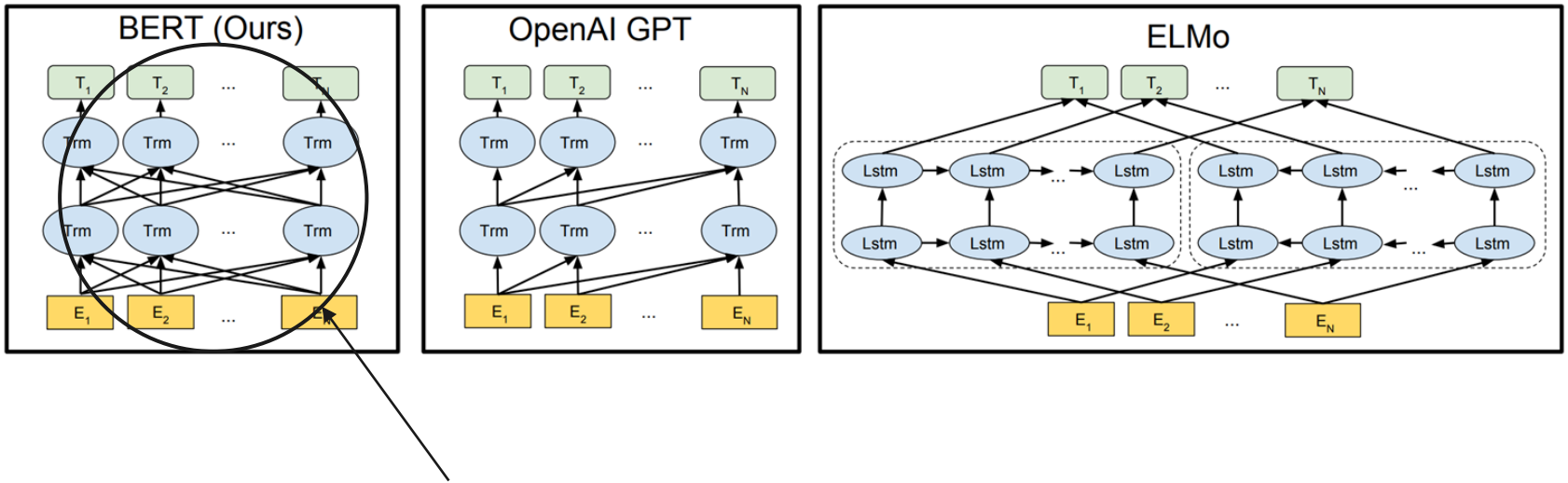
Introduces separators at unsupervised training time
(GPT introduced it only at fine-tuning)



# Architecture v.s. Others



# Architecture v.s. Others




What can go wrong in there...?

# Unsupervised task #1: Masked LM

- Mask 15% of the input tokens in each sequence at random by introducing as `[MASK]` token
- The `[MASK]` token is never seen during fine-tuning
  - 80% of the time, replace `[MASK]` token with the `[MASK]` token
  - 10% of the time, replace `[MASK]` token with a random word
  - 10% of the time, keep the word unchanged
- The T-E does not know in advance which token it will need to predict (15% of the masked ones)
- Converges slower than traditional T-D which predicts every tokens on each batch

# Unsupervised task #2: Next Sentence Prediction

- Capture the relation between 2 sentences (good for QA)
- Binarized next sentence prediction (different from Skip-Thought that we saw)

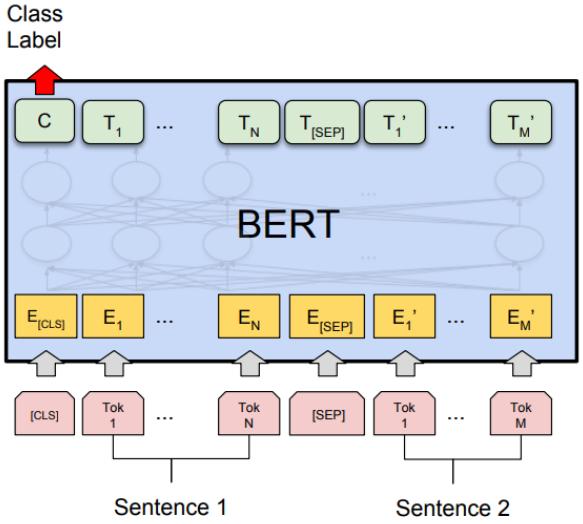
$$\text{Input} = \texttt{[CLS] the man went to [MASK] store [SEP]}$$
$$\texttt{he bought a gallon [MASK] milk [SEP]}$$
$$\text{Label} = \texttt{IsNext}$$

$$\text{Input} = \texttt{[CLS] the man [MASK] to the store [SEP]}$$
$$\texttt{penguin [MASK] are flight ##less birds [SEP]}$$
$$\text{Label} = \texttt{NotNext}$$

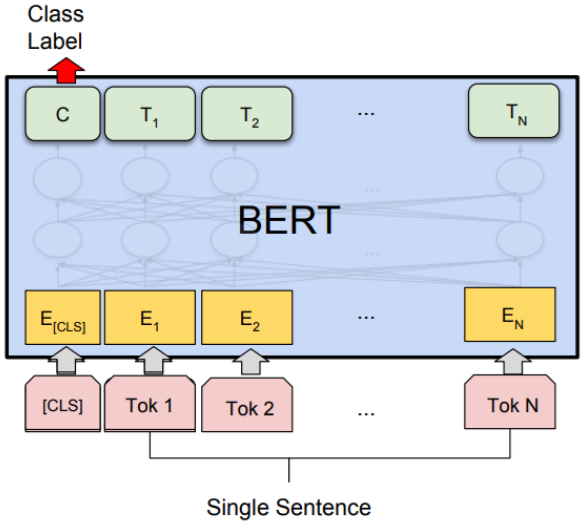
# Pre-training dataset

- BookCorpus as in GPT (800M words
- **and** English Wikipedia (2,500M words)

# Fine-tuning

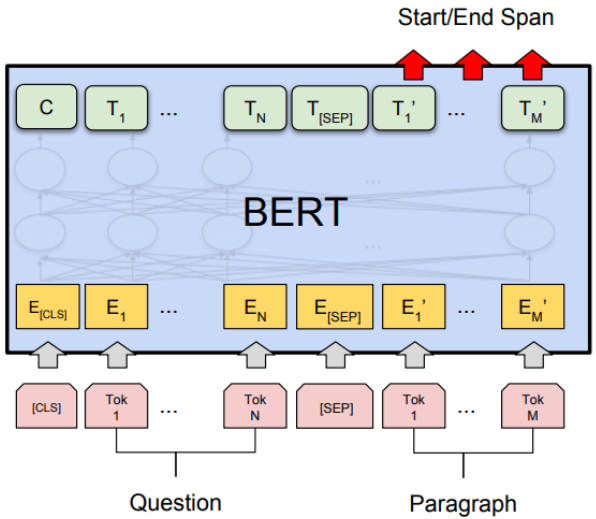



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

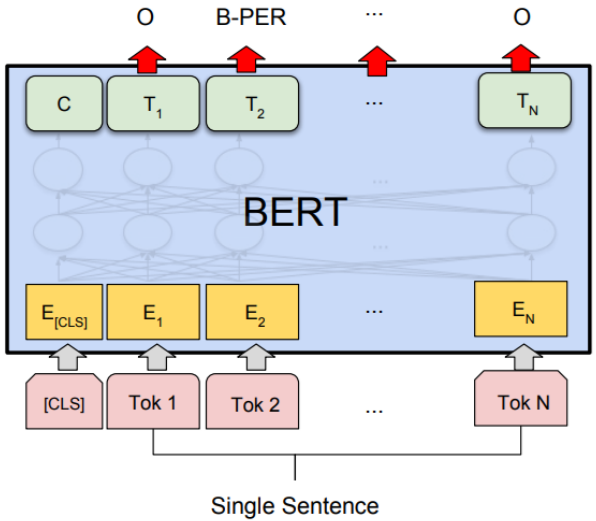



(b) Single Sentence Classification Tasks: SST-2, CoLA

# Fine-tuning




(c) Question Answering Tasks:
SQuAD v1.1




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

# Results

Too many, see the paper!

# Ablation study

Trained w/o next sentence prediction

Trained left-to-right language model

| Tasks | Dev Set MNLI-m (Acc) | QNLI (Acc) | MRPC (Acc) | SST-2 (Acc) | SQuAD (F1) |
|---|---|---|---|---|---|
| $\text{BERT}_{\text{BASE}}$ | 84.4 | 88.4 | 86.7 | 92.7 | 88.5 |
| No NSP | 83.9 | 84.9 | 86.5 | 92.6 | 87.9 |
| LTR & No NSP | 82.1 | 84.3 | 77.5 | 92.1 | 77.8 |
| + BiLSTM | 82.1 | 84.1 | 75.7 | 91.6 | 84.9 |

Table 5: Ablation over the pre-training tasks using the $\text{BERT}_{\text{BASE}}$ architecture. "No NSP" is trained without the next sentence prediction task. "LTR & No NSP" is trained as a left-to-right LM without the next sentence prediction, like OpenAI GPT. "+ BiLSTM" adds a randomly initialized BiLSTM on top of the "LTR + No NSP" model during fine-tuning.

# Conclusion

- Uses a bidirectional LM (compared to unidirectional LM of GPT) which is the single most important contribution
- They assume that for **sentence level tasks** the model should attend not only to previous words but also to futur words
- First fine-tuning based representation to achieve state-of-the art results on both sentence-level ***and token-level*** tasks