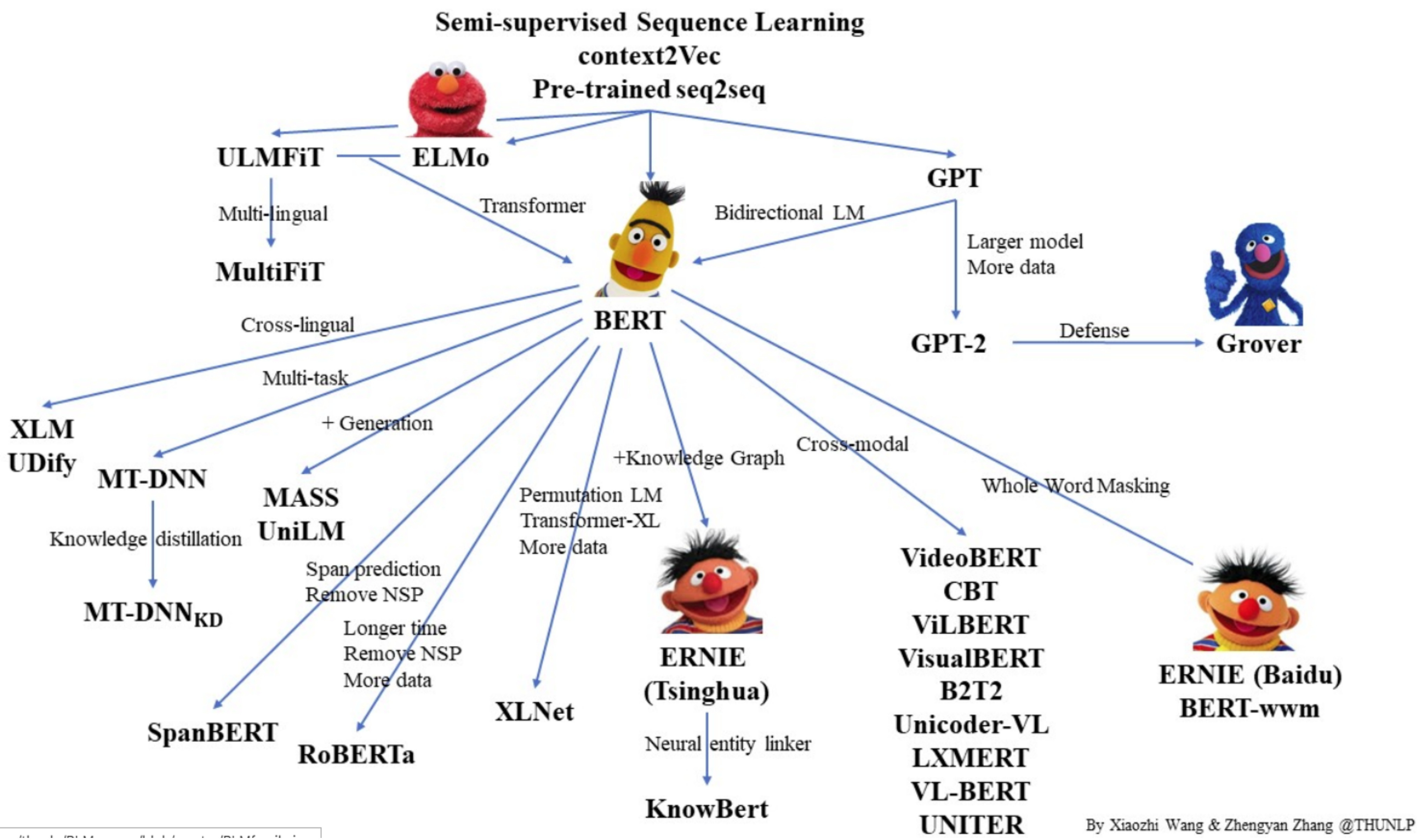
# *BERT*

Nikolay Arefyev
*CMC MSU Department of Algorithmic Languages &*
*Samsung Moscow Research Center*


Semi-supervised Sequence Learning
context2Vec
Pre-trained seq2seq
ULMFiT
ELMo
GPT
Multi-lingual
MultiFiT
Transformer
Bidirectional LM
BERT
Larger model
More data
GPT-2
Defense
Grover
Cross-lingual
Multi-task
+ Generation
XLM
UDify
MT-DNN
MASS
UniLM
Knowledge distillation
$MT\text{-}DNN_{KD}$
Span prediction
Remove NSP
SpanBERT
Longer time
Remove NSP
More data
RoBERTa
Permutation LM
Transformer-XL
More data
XLNet
+Knowledge Graph
ERNIE
(Tsinghua)
Neural entity linker
KnowBert
Cross-modal
VideoBERT
CBT
ViLBERT
VisualBERT
B2T2
Unicoder-VL
LXMERT
VL-BERT
UNITER
Whole Word Masking
ERNIE (Baidu)
BERT-wwm
By Xiaozhi Wang & Zhengyan Zhang @THUNLP

# BERT

- BERT = **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
- Presented in NAACL (June, 2019), best long paper award

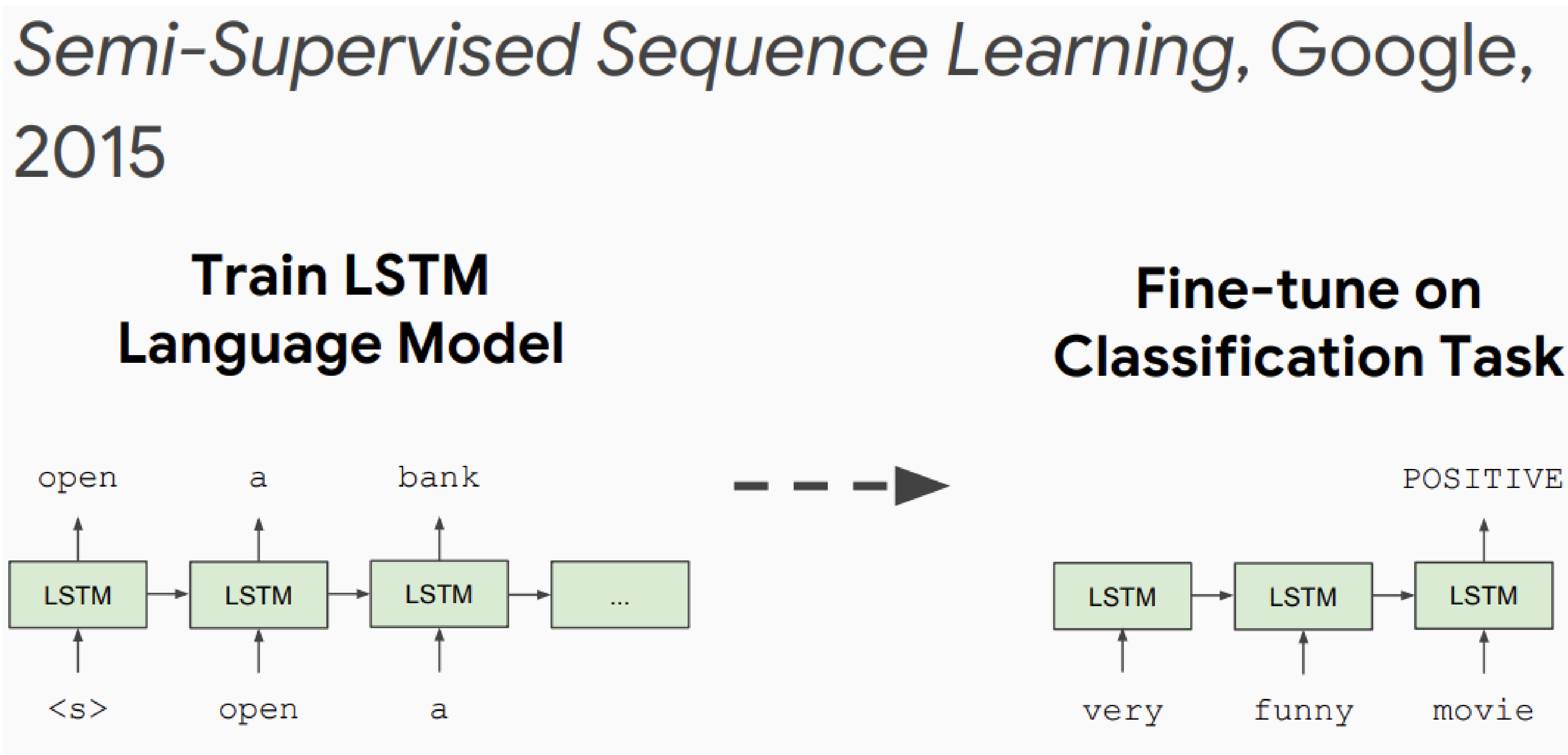
**BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**

**Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

[v1] Thu, 11 Oct 2018 00:50:01 UTC (227 KB)
**[v2]** Fri, 24 May 2019 20:37:26 UTC (309 KB)

# Contextualized representations pre-training

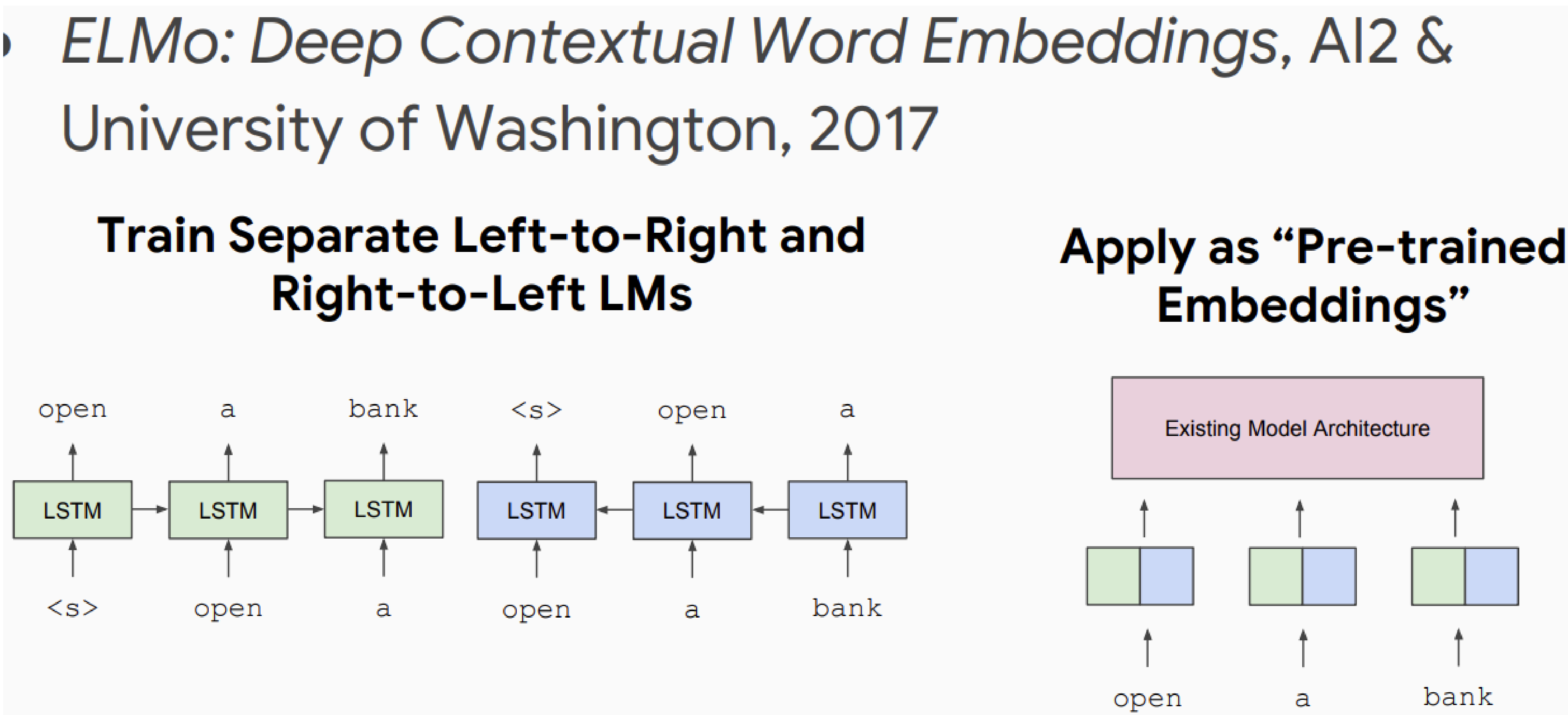
- Fwd LSTM LM => not bidirectional repr. (before fine-tuning word representations depend on left context only)
- Fine-tuned for downstream task




Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Slides. https://nlp.stanford.edu/seminar/details/jdevlin.pdf

# Contextualized representations pre-training

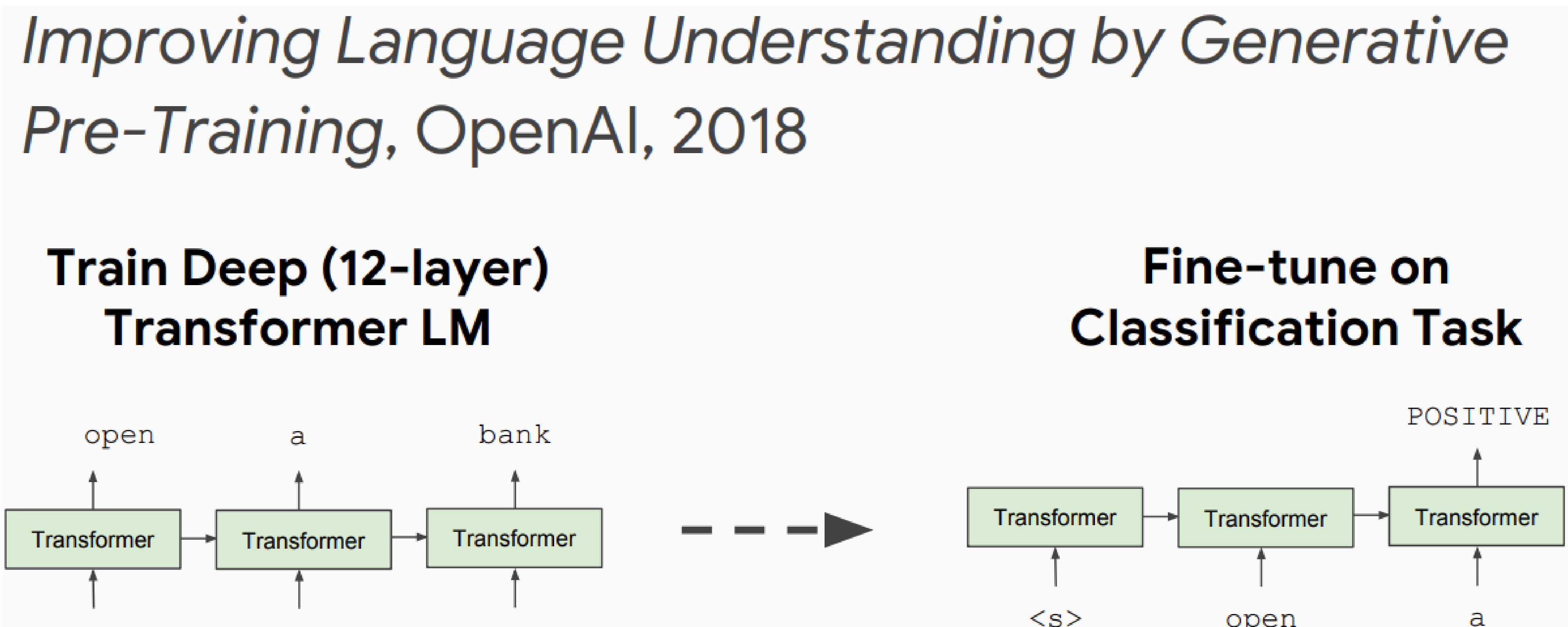
- Fwd&bwd LSTM LMs, word representations are concatenated => not “deep bidirectional repr.”
- Used as additional inputs in other NNs
- Improved results of several SOTA architectures on different tasks




Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Slides. https://nlp.stanford.edu/seminar/details/jdevlin.pdf

# Contextualized representations pre-training

- Fwd **Transformer** LM => not bidirectional repr. (before finetuning)
- Fine-tuned after minimal task-specific adaptations
- Outperformed SOTA task-specific architectures



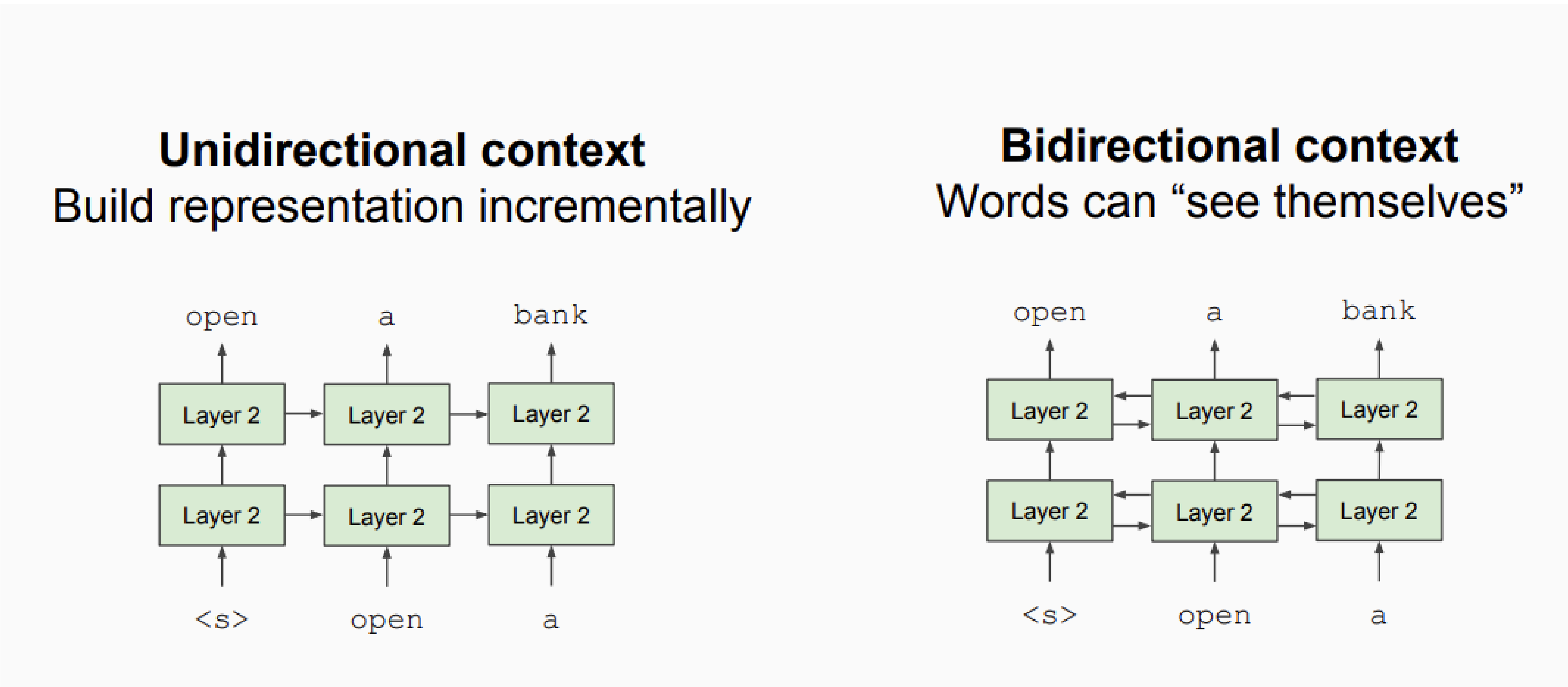

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Slides. https://nlp.stanford.edu/seminar/details/jdevlin.pdf

# Contextualized representations pre-training

- **Problem**: Language models only use left context *or* right context, but language understanding is bidirectional.
- Why are LMs unidirectional?
- Reason 1: Directionality is needed to generate a well-formed probability distribution.
  - We don't care about this.
- Reason 2: Words can "see themselves" in a bidirectional encoder.

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Slides. https://nlp.stanford.edu/seminar/details/jdevlin.pdf

# Contextualized representations pre-training




Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Slides. https://nlp.stanford.edu/seminar/details/jdevlin.pdf

# Masked LM

- **Solution**: Mask out *k*% of the input words, and then predict the masked words
    - We always use $k$ = 15%

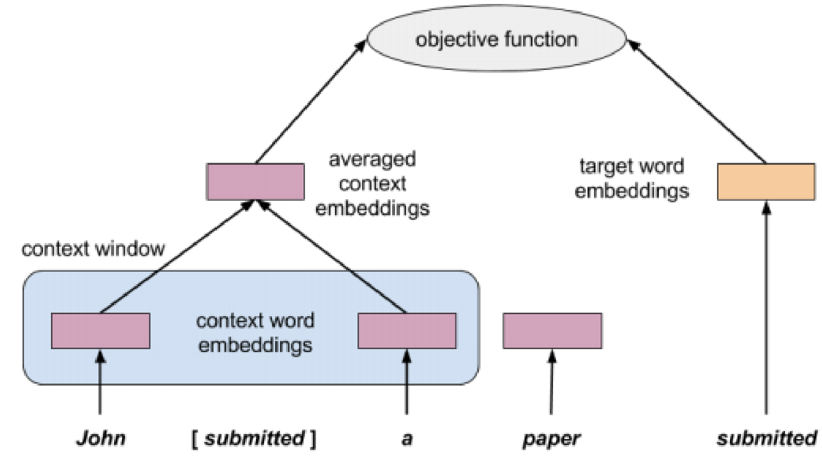
```
                                  store             gallon
                                    ↑                 ↑
the man went to the [MASK] to buy a [MASK] of milk
```

- Too little masking: Too expensive to train
- Too much masking: Not enough context

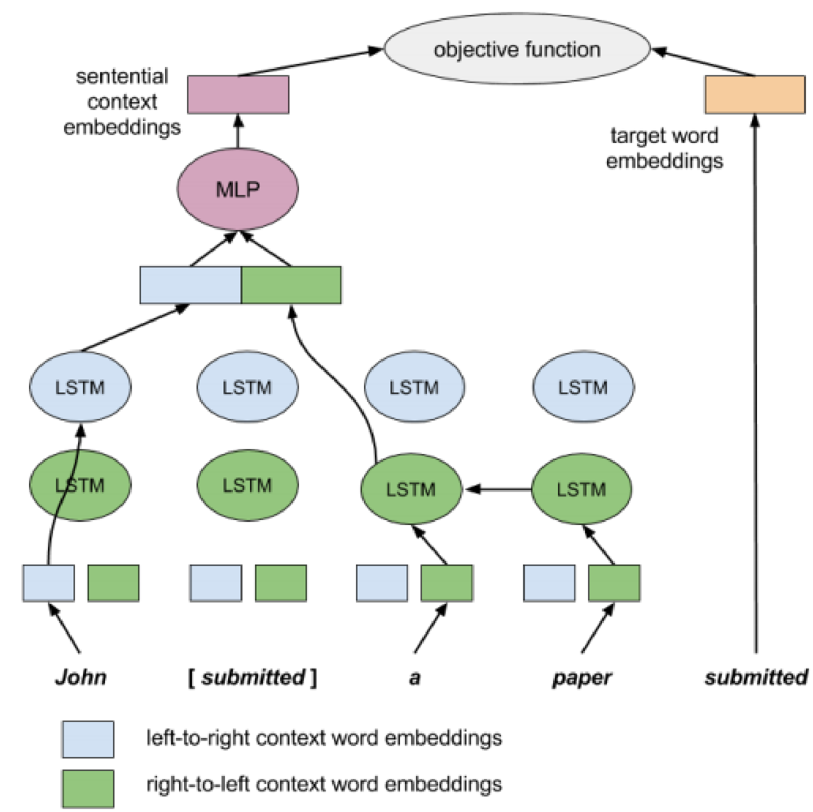
Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Slides. https://nlp.stanford.edu/seminar/details/jdevlin.pdf

# Context2vec

- Similar to word2vec CBOW
  - Predicts **w** given its context **l**,**r**
  - Negative sampling objective
- Instead of averaging context embs:
  - concat($\text{LSTM}_{\text{fwd}}$(l), $\text{LSTM}_{\text{bwd}}$(r))
  - Linear → ReLU → Linear
- "deep bidirectional" word repr.?
  - combines **l** and **r** using FFNN
  - but doesn't see target word
- 2B words ukWaC (12GB), removed long sentences (>64 words)



(a) word2vec *CBOW*




Melamud et al. context2vec: Learning Generic Context Embedding with Bidirectional LSTM, 2016

# Transformer (MLM) pretrain for WSI

Oleg Struyanskiy and Nikolay Arefyev. Neural Networks with Attention for Word Sense Induction // AIST’2018 (July 2018, BERT paper appeared in October 2018)

- Pre-train Transformer to restore words hidden from its input (later called Masked LM pre-train)
  - Replace ambiguous words with CENTERWORD, require it at the output

    ***They were standing on the CENTERWORD of the Volga river → bank***
  - Idea: since there are several possible answers, the model shall learn to predict possible substitutes for the target word, hence, internally disambiguate it
  - Only 1 word was hidden in each example (too few?)
- For 341 ambiguous words found 12M text fragments (4.5GB)
  - Wanted many examples for the words of interest, while keeping training time reasonable (a few days on 1GPU)
  - Examples per word IQR: 25K-65K
  - Length: 20 words on average (following WSI dataset)
- WSI is unsupervised task, so no finetuning

# BERT secrets

- Very large model & lots of compute
- Moderately large corpora (3.2B words)
- Deep bidirectional contextualized word repr. (seeing all words when encoding each word in each layer)

| ULMfit | GPT | BERT | GPT-2 |
|---|---|---|---|
| Jan 2018 | June 2018 | Oct 2018 | Feb 2019 |
| Training: | Training | Training | Training |
| 1 GPU day | 240 GPU days | 256 TPU days<br>~320–560 GPU days | ~2048 TPU v3 days according to a reddit thread |

Christopher Manning. Cs224n/Ling284 slides.Lecture 13.

# Masked LM objective

- Problem: don’t want masking for downstream tasks, but only outputs from masked timesteps affect MLM loss

  => bad representations for all other timesteps?

  Solution: CE on (some) non-masked timesteps also
- Problem: can learn simply to copy non-masked timesteps without even looking at context

  Solution: replace non-masked timesteps with random words sometimes

This is **denoising autoencoder**: replace some tokens from a text fragment with random tokens / [MASK] token and try to reconstruct initial fragment!

# Masked LM objective

- Solution: 15% of the words to predict, but don't replace with `[MASK]` 100% of the time. Instead:
- 80% of the time, replace with `[MASK]`

  `went to the store → went to the [MASK]`
- 10% of the time, replace random word

  `went to the store → went to the running`
- 10% of the time, keep same

  `went to the store → went to the store`

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Slides. https://nlp.stanford.edu/seminar/details/jdevlin.pdf

# Next Sentence Prediction (NSP) objective

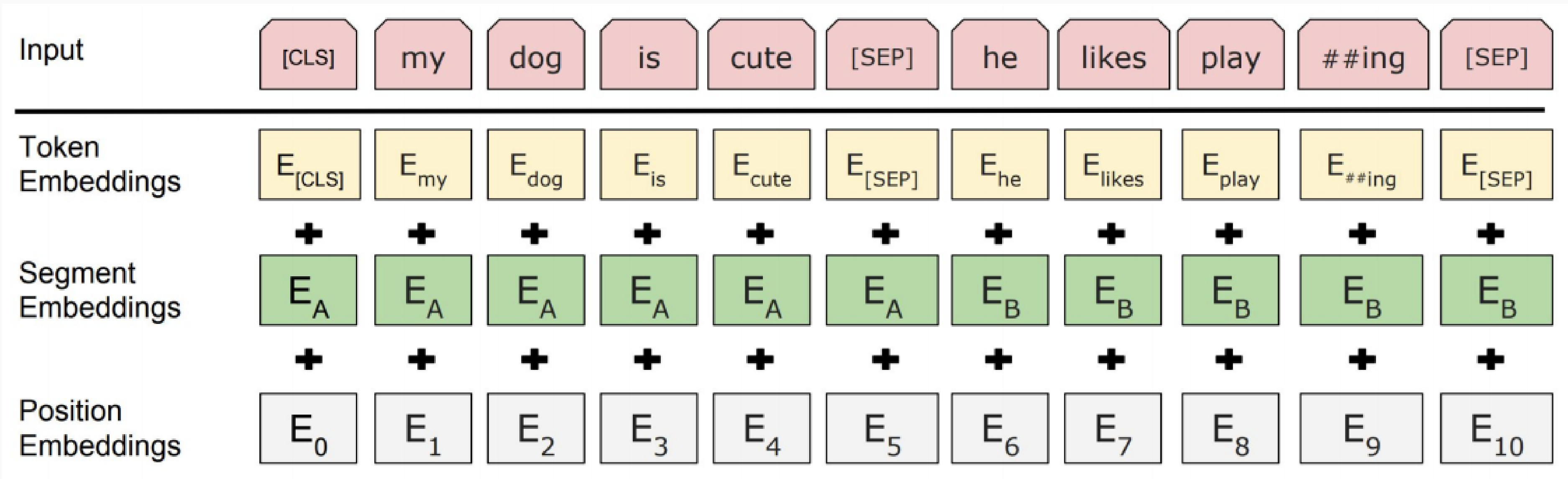
- Input 2 sentences: A,B; predict if B is next sentence after A
  - Sample B following A (with p=0.5) or random sentence from the corpus
  - Learn to extract relations between 2 sentences (for tasks like paraphrase detection, NLI)
- NSP was shown to worsen results in the following research (too simple objective?)
  - More difficult alternative: always sample 2 adjacent sentences and predict their order

**Sentence A** = `The man went to the store.`
**Sentence B** = `He bought a gallon of milk.`
**Label** = `IsNextSentence`

**Sentence A** = `The man went to the store.`
**Sentence B** = `Penguins are flightless.`
**Label** = `NotNextSentence`

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Slides. https://nlp.stanford.edu/seminar/details/jdevlin.pdf

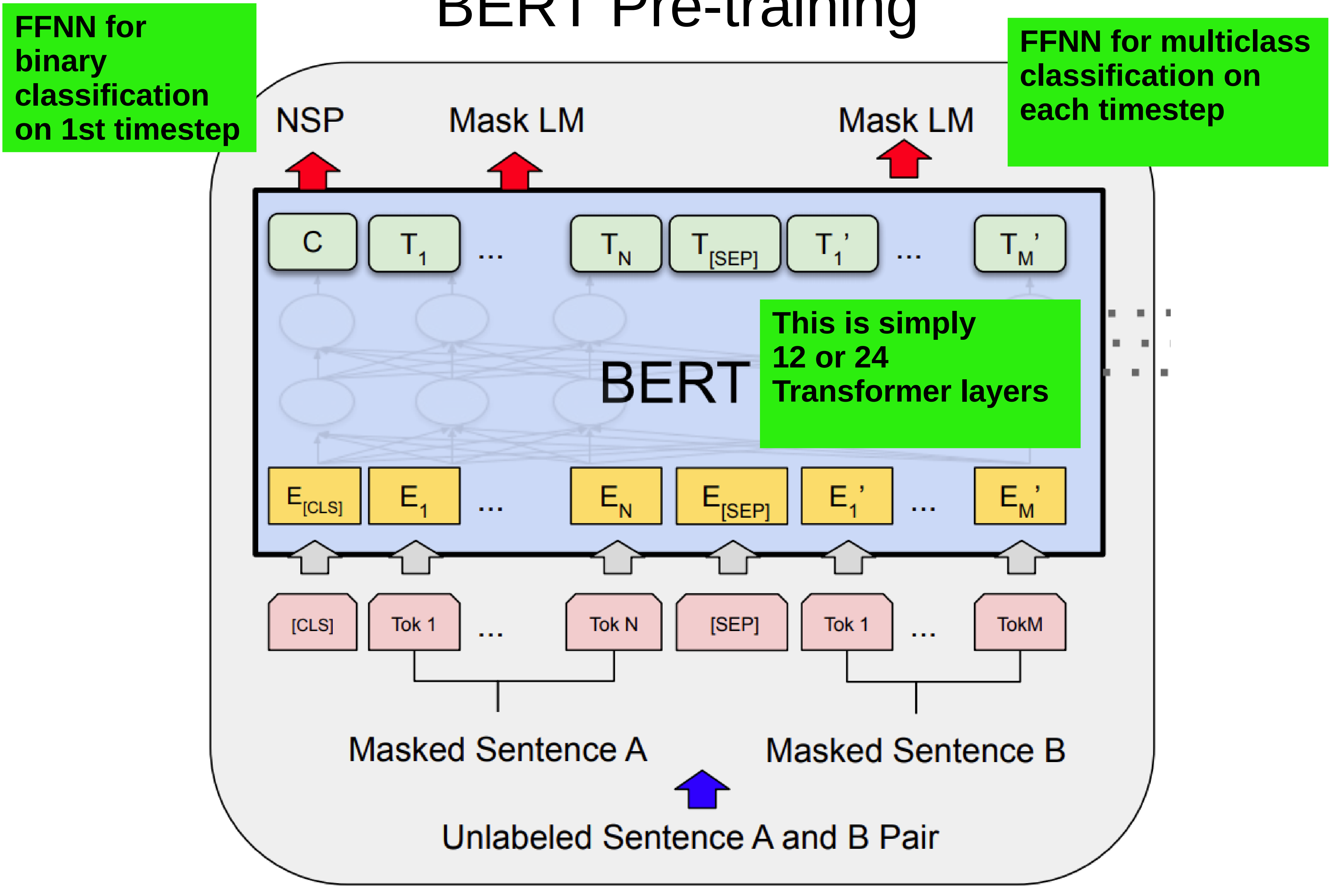
# Inputs




- Use 30,000 WordPiece vocabulary on input.
- Each token is sum of three embeddings
- Single sequence is much more efficient.

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Slides. https://nlp.stanford.edu/seminar/details/jdevlin.pdf

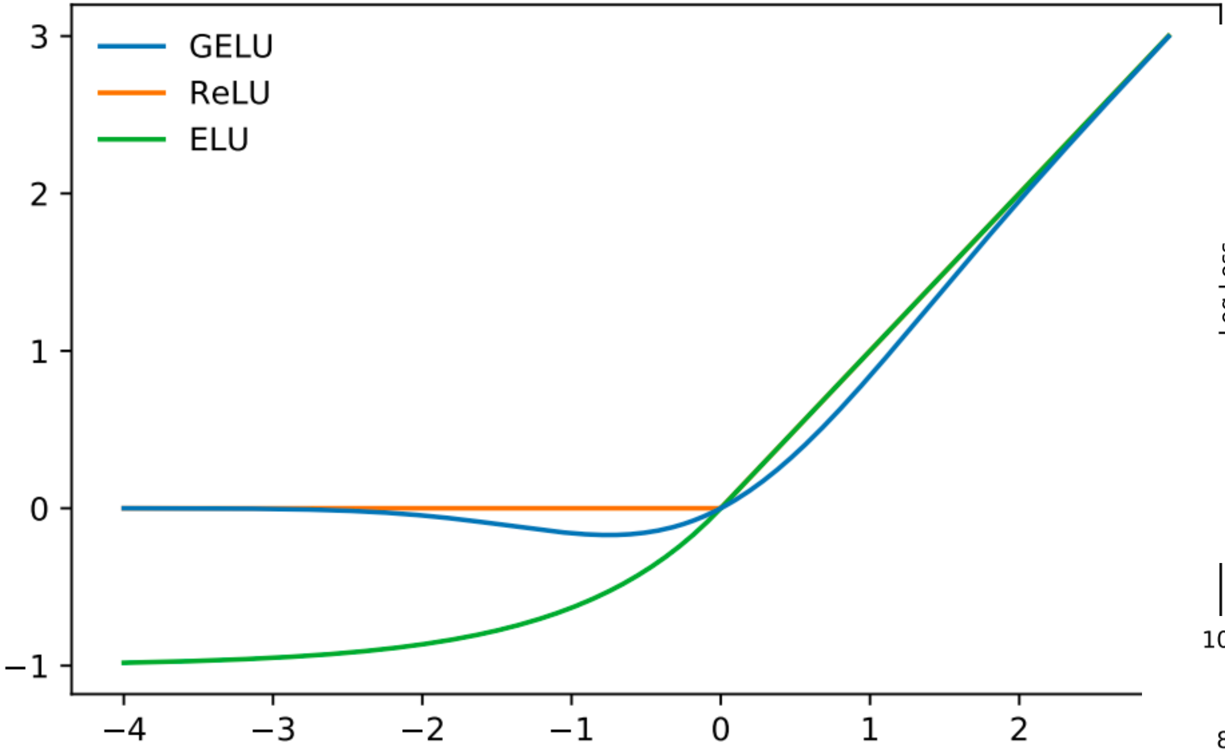
# BERT Pre-training

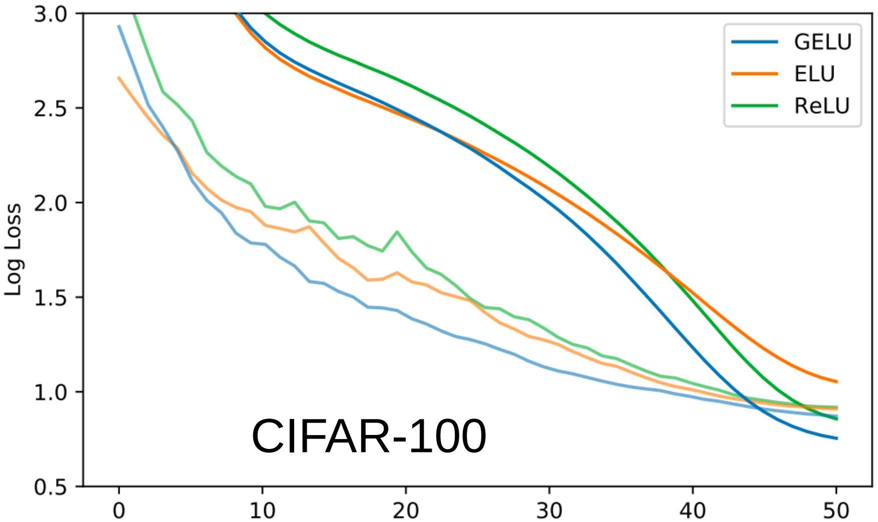


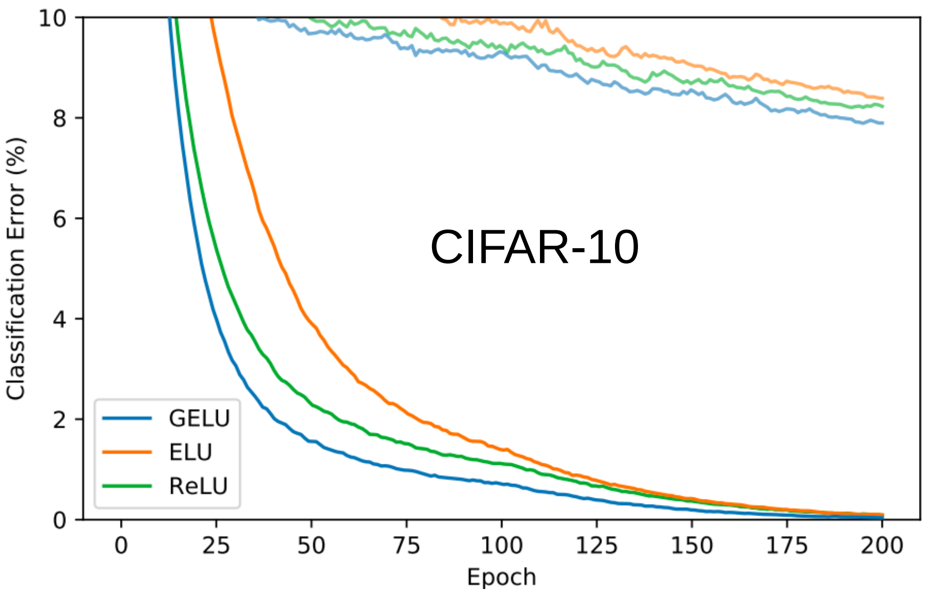

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

# GELU

- both GPT and BERT changed ReLU to GELU




Hendrycks and Gimpel. Gaussian Error Linear Units, 2016

# BERT Pre-training

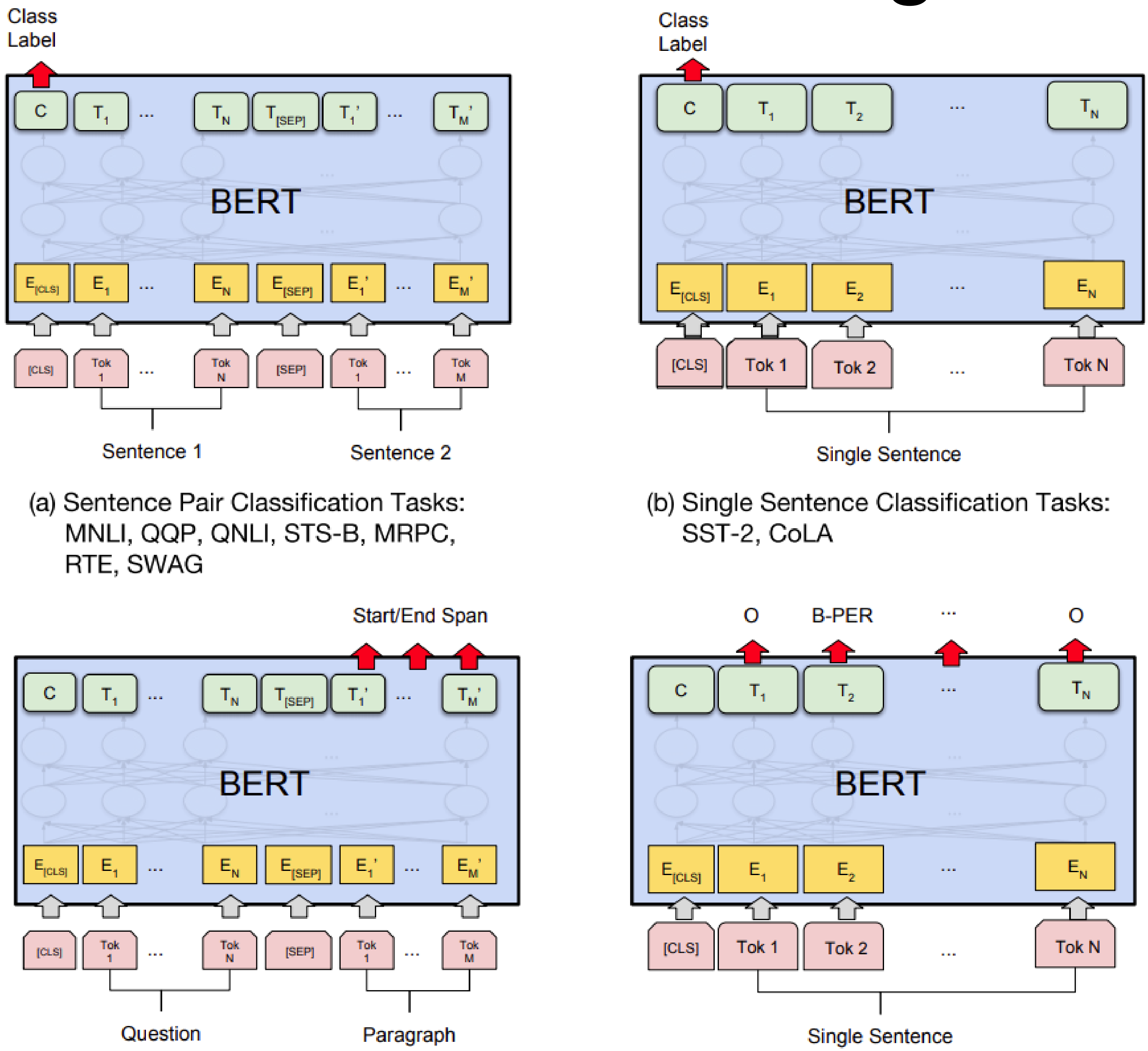
- Data: Wikipedia (2.5B words) + BookCorpus (800M words)
- Batch Size: 131,072 words (1024 sequences * 128 length or 256 sequences * 512 length)
- Training Time: 1M steps (~40 epochs)

**+10k steps lr warmup**

- Optimizer: AdamW, 1e-4 learning rate, linear decay
- `BERT-Base`: 12-layer, 768-hidden, 12-head **110M params**
- `BERT-Large`: 24-layer, 1024-hidden, 16-head **340M params**
- Trained on 4x4 or 8x8 TPU slice for 4 days

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Slides. https://nlp.stanford.edu/seminar/details/jdevlin.pdf

# BERT Fine-tuning




(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

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

# GLUE Results

| System | MNLI-(m/mm) | QQP | QNLI | SST-2 | CoLA | STS-B | MRPC | RTE | **Average** |
|---|---|---|---|---|---|---|---|---|---|
| | 392k | 363k | 108k | 67k | 8.5k | 5.7k | 3.5k | 2.5k | - |
| Pre-OpenAI SOTA | 80.6/80.1 | 66.1 | 82.3 | 93.2 | 35.0 | 81.0 | 86.0 | 61.7 | 74.0 |
| BiLSTM+ELMo+Attn | 76.4/76.1 | 64.8 | 79.9 | 90.4 | 36.0 | 73.3 | 84.9 | 56.8 | 71.0 |
| OpenAI GPT | 82.1/81.4 | 70.3 | 88.1 | 91.3 | 45.4 | 80.0 | 82.3 | 56.0 | 75.2 |
| $\text{BERT}_{\text{BASE}}$ | 84.6/83.4 | 71.2 | 90.1 | 93.5 | 52.1 | 85.8 | 88.9 | 66.4 | 79.6 |
| $\text{BERT}_{\text{LARGE}}$ | **86.7/85.9** | **72.1** | **91.1** | **94.9** | **60.5** | **86.5** | **89.3** | **70.1** | **81.9** |

**MultiNLI**

Premise: Hills and mountains are especially sanctified in Jainism.
Hypothesis: Jainism hates nature.
Label: Contradiction

**CoLa**

Sentence: The wagon rumbled down the road.
Label: Acceptable

Sentence: The car honked down the road.
Label: Unacceptable

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Slides. https://nlp.stanford.edu/seminar/details/jdevlin.pdf

# SQuAD 1.1 Results

**What was another term used for the oil crisis?**
*Ground Truth Answers:* first oil shock | shock | shock | first oil shock | shock
*Prediction:* shock

The 1973 oil crisis began in October 1973 when the members of the Organization of Arab Petroleum Exporting Countries (OAPEC, consisting of the Arab members of OPEC plus Egypt and Syria) proclaimed an oil embargo. By the end of the embargo in March 1974, the price of oil had risen from US$3 per barrel to nearly $12 globally; US prices were significantly higher. The embargo caused an oil crisis, or "shock", with many short- and long-term effects on global politics and the global economy. It was later called the "first oil shock", followed by the 1979 oil crisis, termed the "second oil shock."

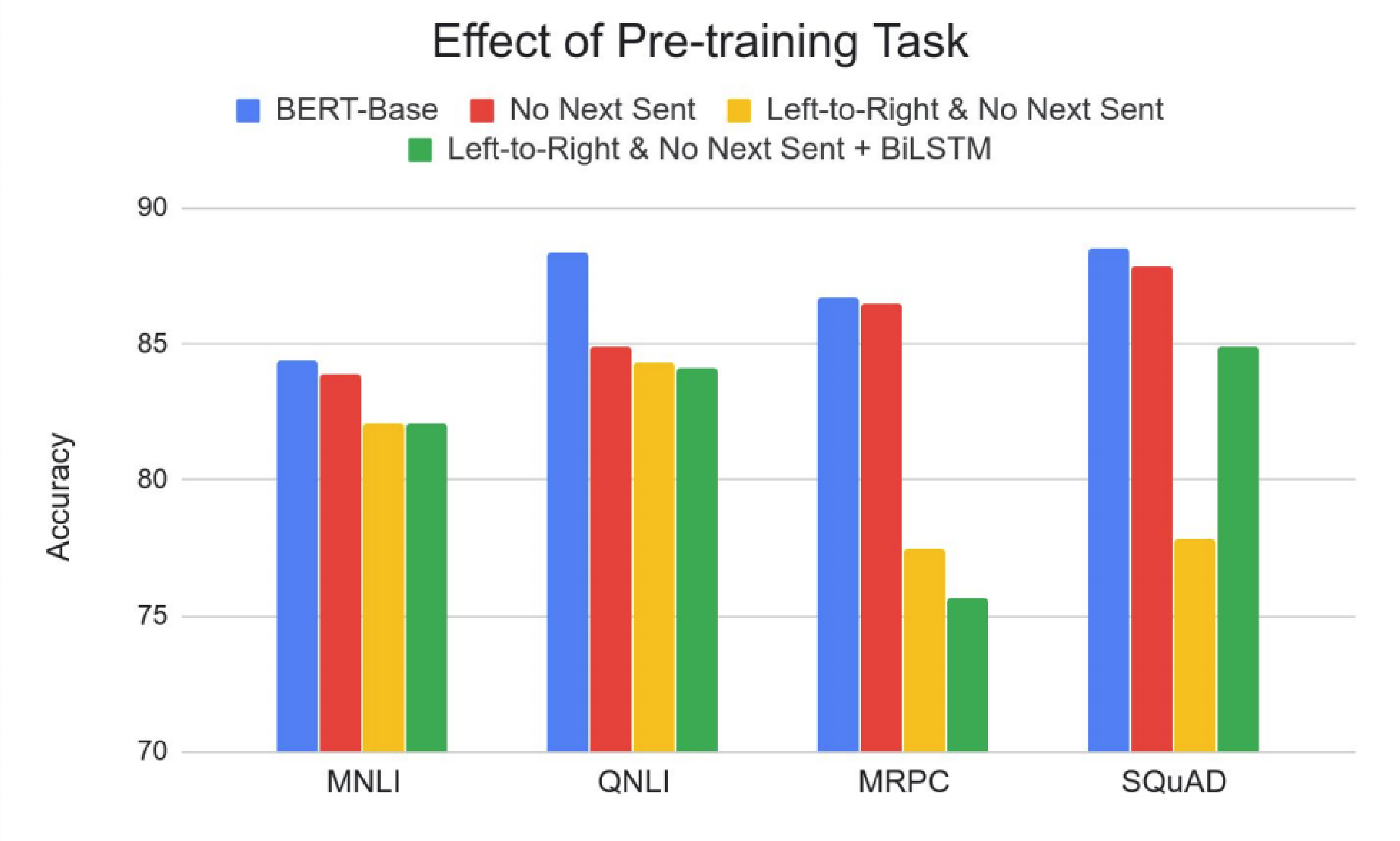
- Only new parameters: Start vector and end vector.
- Softr … itions.

$$P_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}}$$

| Rank | Model | EM | F1 |
|---|---|---|---|
| | Human Performance<br>*Stanford University*<br>(Rajpurkar et al. '16) | 82.304 | 91.221 |
| 1<br>Oct 05, 2018 | BERT (ensemble)<br>*Google AI Language*<br>https://arxiv.org/abs/1810.04805 | **87.433** | **93.160** |
| 2<br>Oct 05, 2018 | BERT (single model)<br>*Google AI Language*<br>https://arxiv.org/abs/1810.04805 | 85.083 | 91.835 |
| 2<br>Sep 26, 2018 | nlnet (ensemble)<br>*Microsoft Research Asia* | 85.954 | 91.677 |
| 5<br>Sep 09, 2018 | nlnet (single model)<br>*Microsoft Research Asia* | 83.468 | 90.133 |
| 3<br>Jul 11, 2018 | QANet (ensemble)<br>*Google Brain & CMU* | 84.454 | 90.490 |

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Slides. https://nlp.stanford.edu/seminar/details/jdevlin.pdf

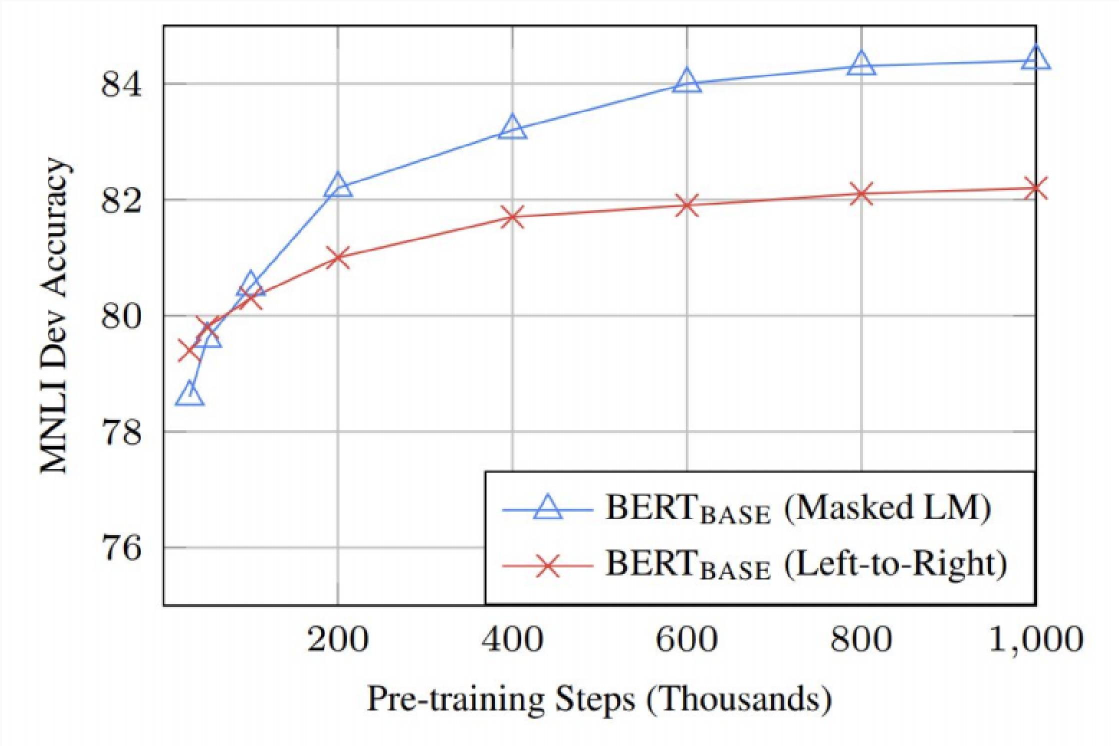
# Pre-training objectives affect




Masked LM (compared to left-to-right LM) is very important on some tasks, Next Sentence Prediction is important on other tasks. Left-to-right model does very poorly on word-level task (SQuAD), although this is mitigated by BiLSTM

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Slides. https://nlp.stanford.edu/seminar/details/jdevlin.pdf

# MLM vs. LM pretraining

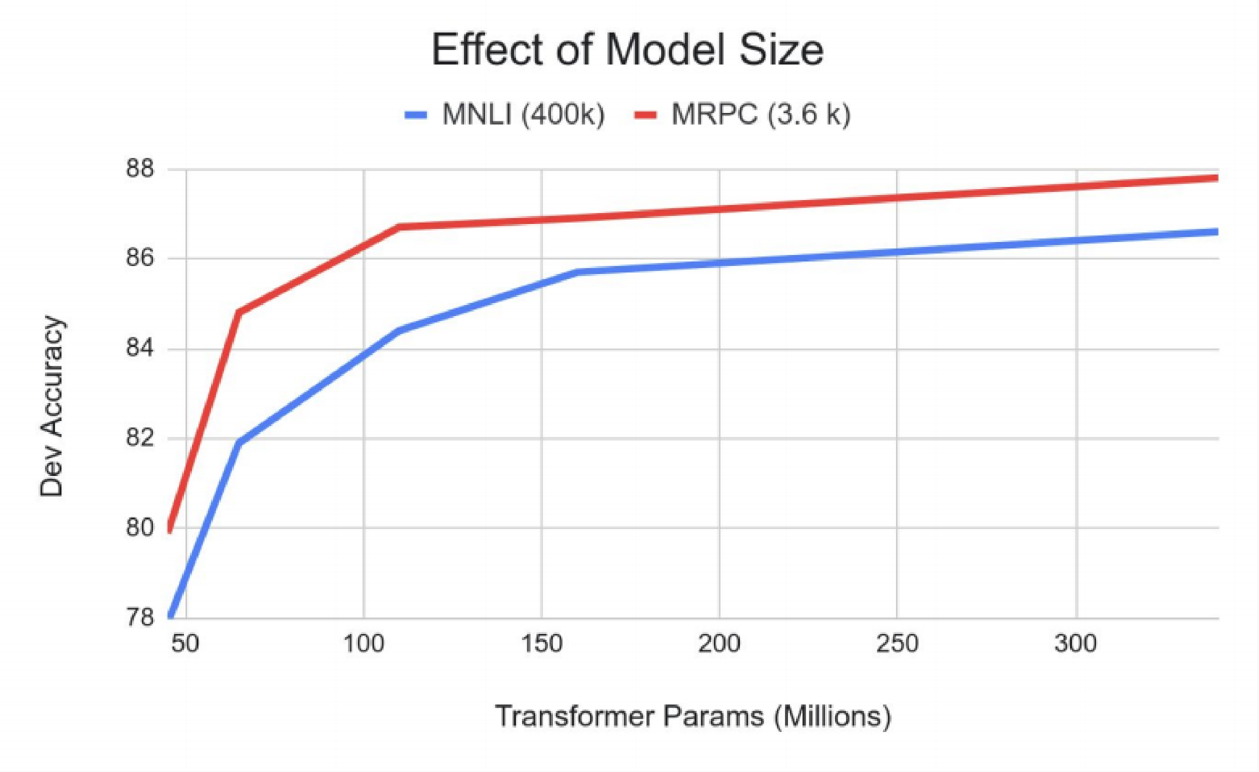



Masked LM takes slightly longer to converge because we only predict 15% instead of 100%

But absolute results are much better almost immediately

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Slides. https://nlp.stanford.edu/seminar/details/jdevlin.pdf

# Do we need huge models?




Big models help *a lot*

Going from 110M -> 340M params helps even on datasets with 3,600 labeled examples

Improvements have *not* asymptoted

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Slides. https://nlp.stanford.edu/seminar/details/jdevlin.pdf