BERT

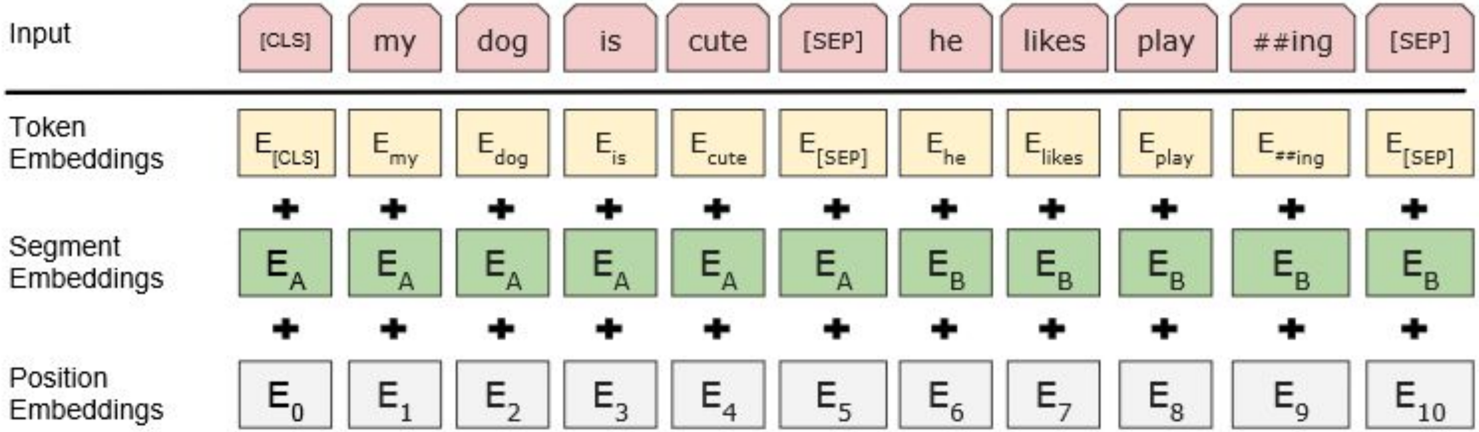
Data Analysis

Dasha Popova

BERT

- = Bidirectional Encoder Representations from Transformers
- Devlin et al. 2018 (https://arxiv.org/pdf/1810.04805.pdf)
- The source: https://github.com/google-research/bert
- Jacob Devlin (jacobdevlin@google.com), Ming-Wei Chang (mingweichang@google.com), Kenton Lee (kentonl@google.com)
- BERT is the first *unsupervised*, *deeply bidirectional* system for pre-training NLP
- bidirectional = учитывает левый и правый контекст
- unsupervised = обучен на неразмеченном корпусе (BooksCorpus, Wikipedia)

BERT: architecture (Devlin et al. 2018)

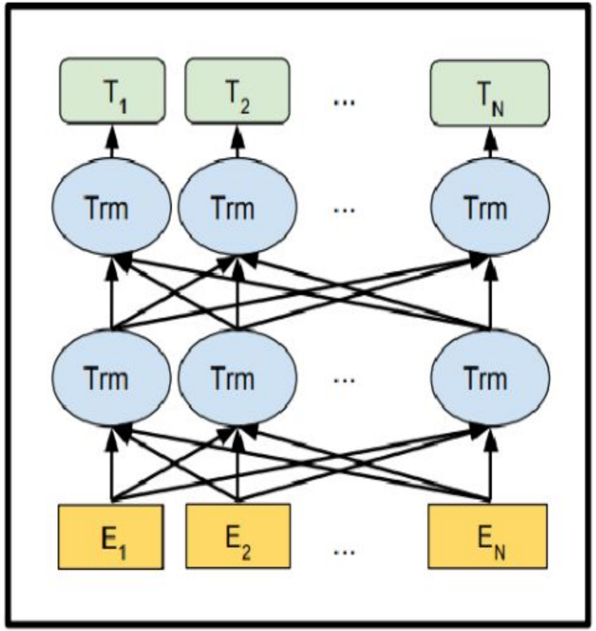


Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{\#\#ing}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Position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

Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

BERT: architecture (https://github.com/google-research/bert)

★ BERT represents "bank" using both its left and right context — I made a ... deposit — starting from the very bottom of a deep neural network, so it is *deeply bidirectional*



BERT: architecture (https://github.com/google-research/bert)

★ Masked Language Models (MLM): 15% токенов на входе маскируются, потом последовательность прогоняется через глубокий двунаправленный трансформер для предсказания замаскированных слов:

  ○ Input: the man went to the [MASK1] . he bought a [MASK2] of milk.

  ○ Labels: [MASK1] = store; [MASK2] = gallon

★ Предсказание следующего предложения (Next Sentence Prediction, NSP):

  ○ Sentence A: the man went to the store.

  ○ Sentence B: he bought a gallon of milk.

  ○ Label: IsNextSentence/NotNextSentence

BERT: application

BERT показывает отличные результаты при решении различных задач, например:

- GLUE (General Language Understanding Evaluation)
- SQuAD (Stanford Question Answering Dataset) -- span-level task
- NER (Named-entity Recognition) -- word-level task
- NLI (Natural Language Inference) -- sentence-pair level task
- ...

Решение различных задач NLU не требует перетренировки модели!

BERT: application

Devlin et al. 2018

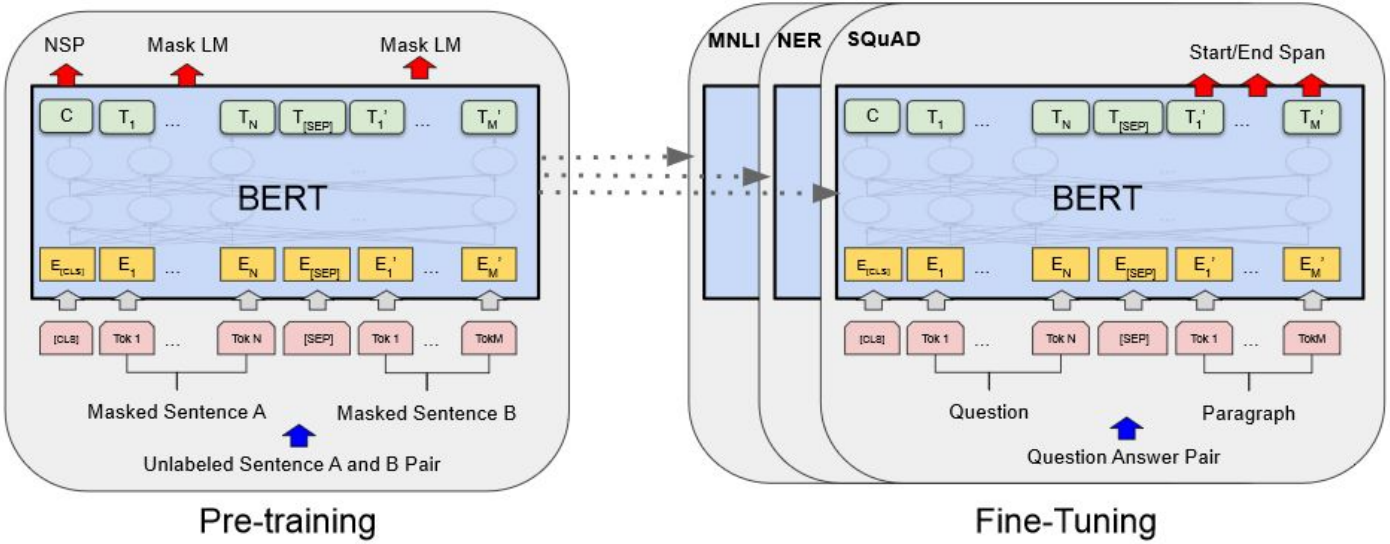


Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers).

BERT: versions

- Модификации BERTa:
  - RoBERTa: нет NSP, динамическая маскировка, чтобы замаскированный токен изменялся в эпоху обучения, обучена на большем объёме данных https://ai.facebook.com/blog/roberta-an-optimized-method-for-pretraining-self-supervised-nlp-systems/
  - Albert (Lite BERT): размерность вложений скрытых слоев (например, 768) больше размерности словарных вложений (например, 128) https://ai.googleblog.com/2019/12/albert-lite-bert-for-self-supervised.html
  - …
  - BERT-Tiny, BERT-Mini, BERT-Small, BERT-Medium, BERT-Base
- Тренировка модели для различных языков