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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova
Google AI Language

CS330 Student Presentation

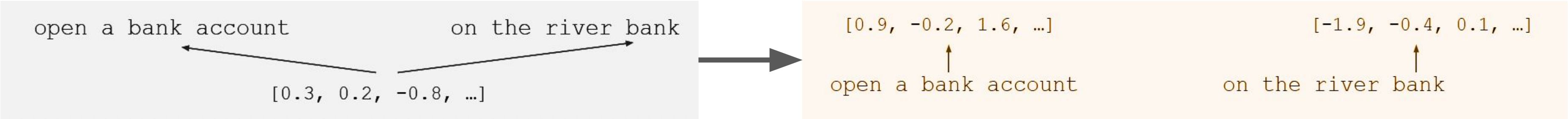
Outline

- Background & Motivation
- Method Overview
- Experiments
- Takeaways & Discussion

Background & Motivation

- Pre-training in NLP
  - Word embeddings are the basis of deep learning for NLP
  - Word embeddings (word2vec, GloVe) are often pre-trained on text corpus
  - Pre-training can effectively improve many NLP tasks

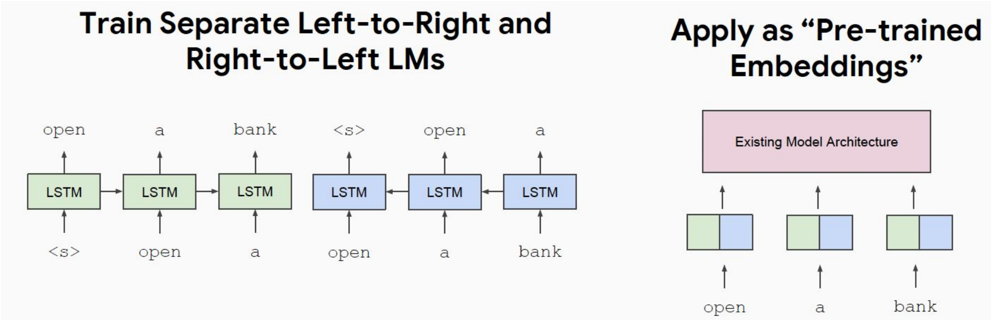
- Contextual Representations
  - Problem: Word embeddings are applied in a context free manner
  - Solution: Train contextual representations on text corpus



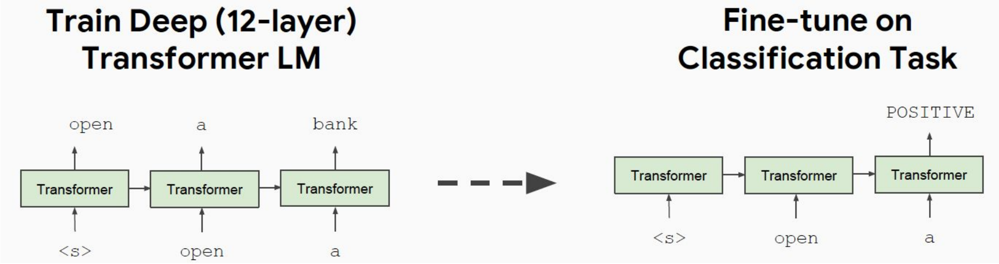
Background & Motivation - related work

Two pre-training representation strategies

- Feature-based approach, ELMo (Peters et al., 2018a)



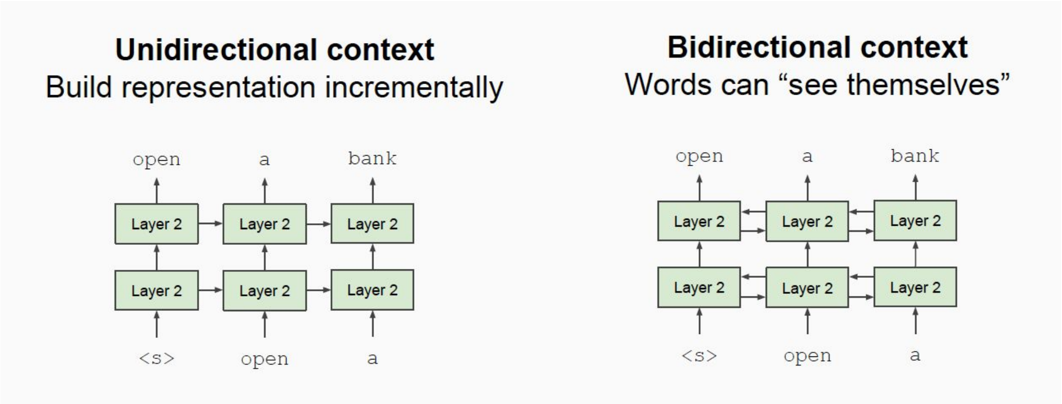
- Fine-tuning approach, OpenAI GPT (Radford et al., 2018)



Background & Motivation

- Problem with previous methods
  - Unidirectional LMs have limited expressive power
  - Can only see left context or right context

- Solution: **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
  - Bidirectional: the word can see both side at the same time
  - Empirically, improved the fine-tuning based approaches



Method Overview

BERT = Bidirectional Encoder Representations from Transformers

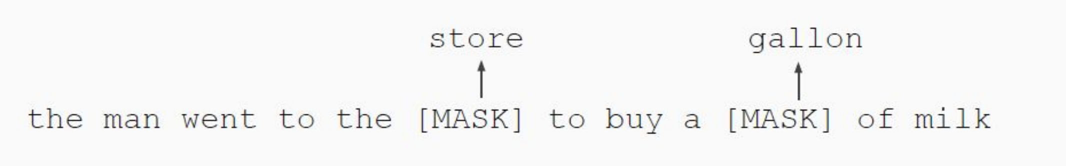
Two steps:

- Pre-training on unlabeled text corpus
  - Masked LM
  - Next sentence prediction
- Fine-tuning on specific task
  - Plug in the task specific inputs and outputs
  - Fine-tune all the parameters end-to-end

Method Overview

Pre-training Task #1: Masked LM → Solve the problem: how to train bidirectional?

- Mask out 15% of the input words, and then predict the masked words



- To reduce bias, among 15% words to predict
  - 80% of the time, replace with [MASK]
  - 10% of the time, replace random word
  - 10% of the time, keep same

Method Overview

Pre-training Task #2: Next Sentence Prediction → learn relationships between sentences

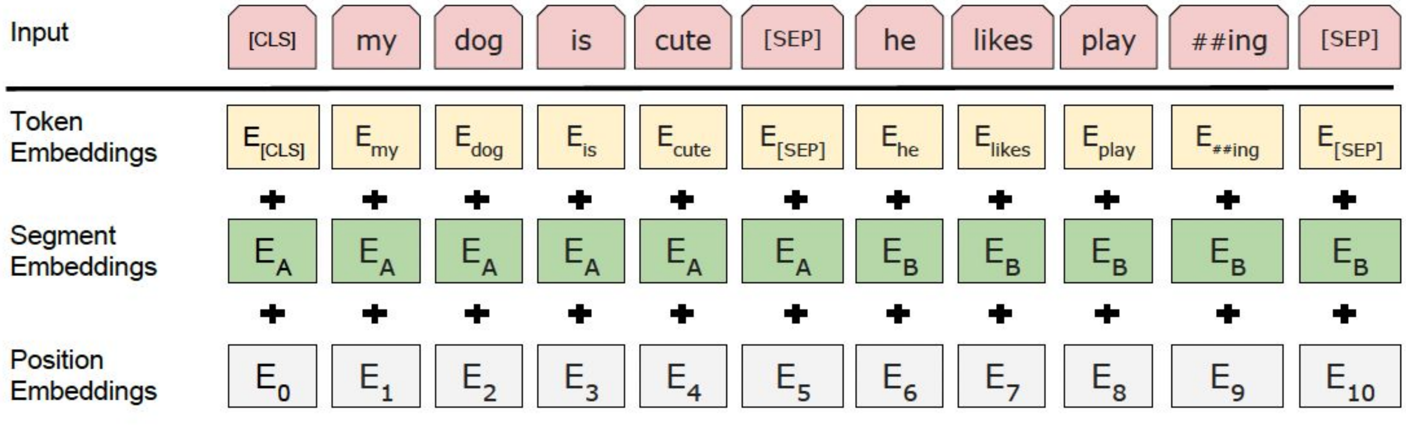
- Classification task
- Predict whether Sentence B is actual sentence that proceeds Sentence A, or a random sentence

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`

Method Overview

Input Representation

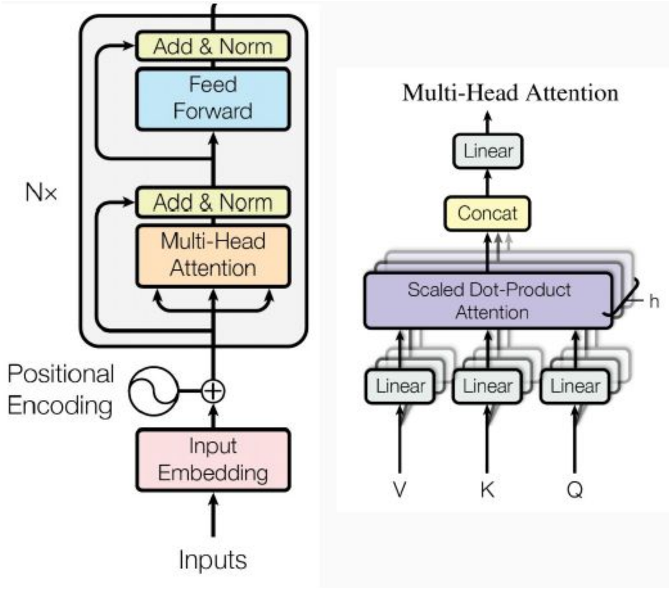


- Use 30,000 WordPiece vocabulary on input
- Each input embedding is sum of three embeddings

Method Overview

Transformer Encoder

- Multi-headed self attention
    - Models context
- Feed-forward layers
    - Computes non-linear hierarchical features
- Layer norm and residuals
    - Makes training deep networks healthy
- Positional encoding
    - Allows model to learn relative positioning



Method Overview

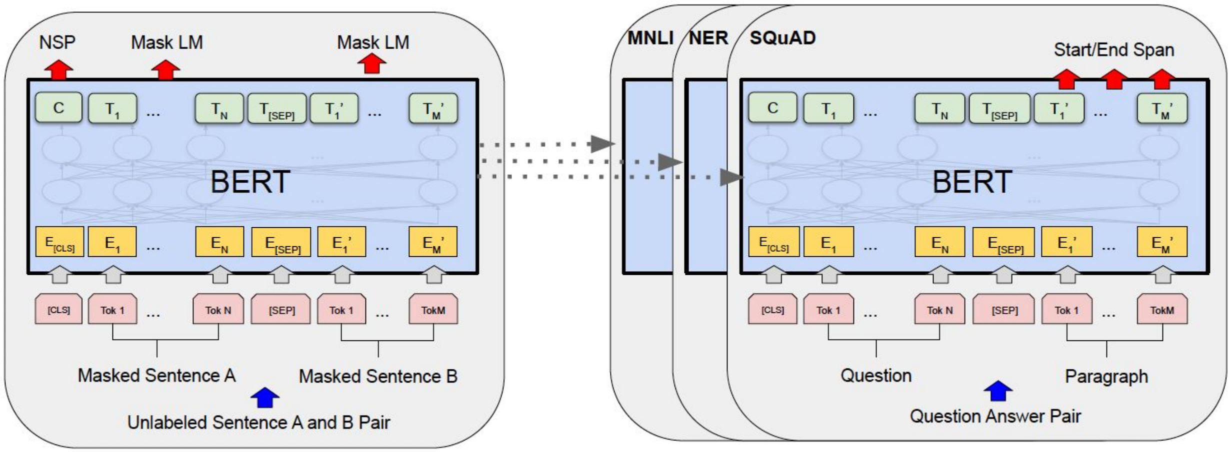
Model Details

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

Method Overview

Fine-tuning Procedure

- Apart from output layers, the same architecture are used in both pre-training and fine-tuning.



Pre-training

Fine-Tuning

Experiments

GLUE (General Language Understanding Evaluation)

- Two types of tasks
    - Sentence pair classification tasks
    - Single sentence classification tasks

MultiNLI

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

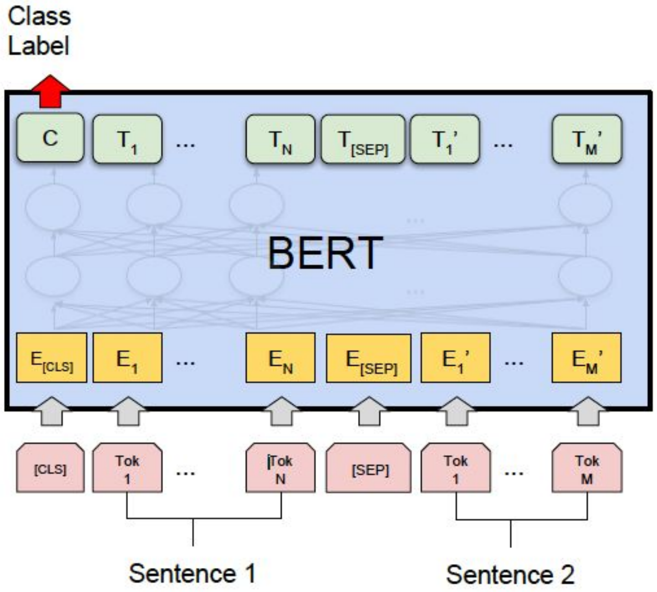
CoLa

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

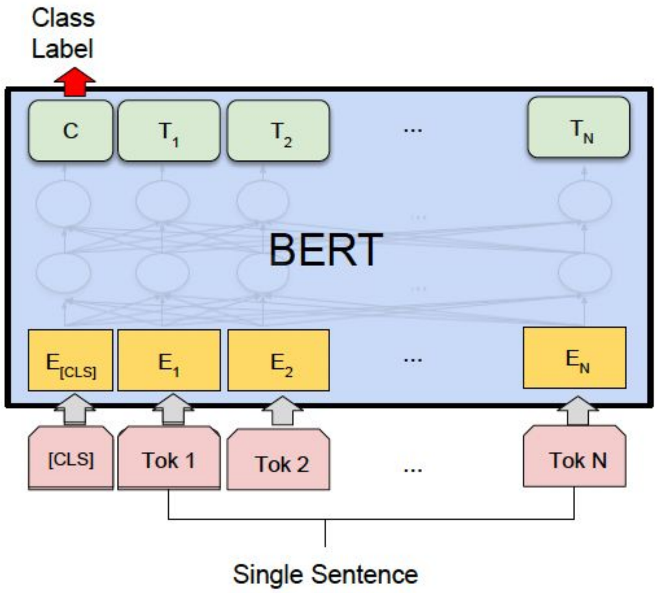
Sentence: The car honked down the road.
Label: Unacceptable

Experiments

GLUE



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



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

Experiments

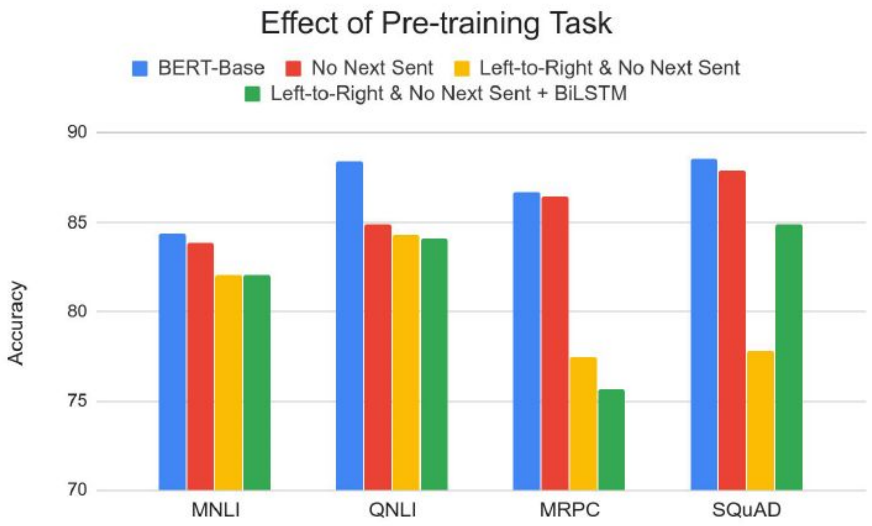
GLUE

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	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

Ablation Study

Effect of Pre-training Task

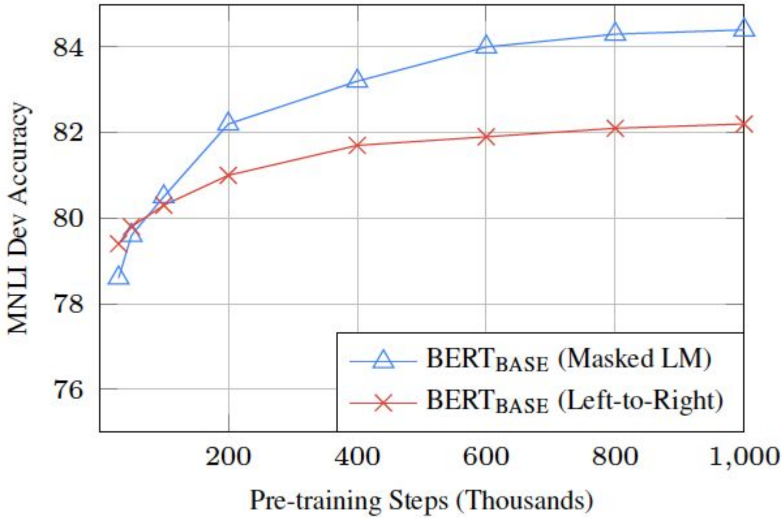
- Masked LM (compared to left-to-right LM) is very important on some tasks, Next Sentence Prediction is important on other tasks.
- Left-to-right model doesn’t work well on word-level task (SQuAD), although this is mitigated by BiLSTM.



Ablation Study

Effect of Directionality and Training Time

- Masked LM takes slightly longer to converge
- But absolute results are much better almost immediately



Ablation Study

Effect of Model Size

- Big models help a lot
- Going from 110M -> 340M params helps even on datasets with 3,600 labeled examples (MRPC)

Hyperparams				Dev Set Accuracy		
#L	#H	#A	LM (ppl)	MNLI-m	MRPC	SST-2
3	768	12	5.84	77.9	79.8	88.4
6	768	3	5.24	80.6	82.2	90.7
6	768	12	4.68	81.9	84.8	91.3
12	768	12	3.99	84.4	86.7	92.9
12	1024	16	3.54	85.7	86.9	93.3
24	1024	16	3.23	86.6	87.8	93.7

Ablation Study

Effect of Model Size

- Big models help a lot
- Going from 110M -> 340M params helps even on datasets with 3,600 labeled examples (MRPC)

Hyperparams				Dev Set Accuracy		
#L	#H	#A	LM (ppl)	MNLI-m	MRPC	SST-2
3	768	12	5.84	77.9	79.8	88.4
6	768	3	5.24	80.6	82.2	90.7
6	768	12	4.68	81.9	84.8	91.3
12	768	12	3.99	84.4	86.7	92.9
12	1024	16	3.54	85.7	86.9	93.3
24	1024	16	3.23	86.6	87.8	93.7

Takeaways & Discussion

Contributions

- Demonstrate the importance of bidirectional pre-training for language representations
- The first **fine-tuning** based model that achieves state-of-the-art on a large suite of tasks, outperforming many **task-specific architectures**
- Advances the state of the art for 11 NLP tasks

Takeaways & Discussion

Critiques

- Bias: Mask token only seen at pre-training, never seen at fine-tuning
- High computation cost
- Not end-to-end
- Doesn’t work for language generation task

Takeaways & Discussion

BERT v.s. MAML

- Two stages
    - Learning the initial weights through pre-training / outer loop updates
    - Fine-tuning / inner loop updates
    - 2-step vs end-to-end
- Shared architecture across different tasks

Thank You!

Ablation Study

Effect of Masking Strategy

- Feature-based Approach with BERT (NER)
- Masking 100% of the time hurts on the feature-based approach
- Using random word 100% of time hurts slightly

Masking Rates			Dev Set Results		
MASK	SAME	RND	MNLI Fine-tune	NER Fine-tune	NER Feature-based
80%	10%	10%	84.2	95.4	$\boxed{94.9}$
100%	0%	0%	84.3	94.9	$\boxed{94.0}$
80%	0%	20%	84.1	95.2	94.6
80%	20%	0%	84.4	95.2	94.7
0%	20%	80%	83.7	94.8	94.6
0%	0%	100%	83.6	94.9	94.6

Ablation Study

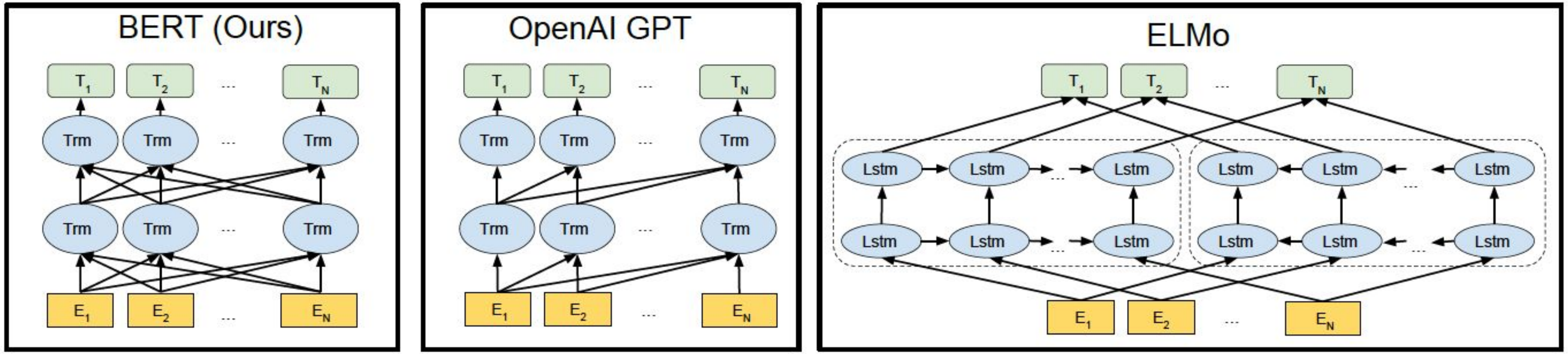
Effect of Masking Strategy

- Feature-based Approach with BERT (NER)
- Masking 100% of the time hurts on the feature-based approach
- Using random word 100% of time hurts slightly

Masking Rates			Dev Set Results		
MASK	SAME	RND	MNLI Fine-tune	NER Fine-tune	NER Feature-based
80%	10%	10%	84.2	95.4	94.9
100%	0%	0%	84.3	94.9	94.0
80%	0%	20%	84.1	95.2	94.6
80%	20%	0%	84.4	95.2	94.7
0%	20%	80%	83.7	94.8	94.6
0%	0%	100%	83.6	94.9	94.6

Method Overview

Compared with OpenAI GPT and ELMo



Ablation Study

Effect if Pre-training Task

- 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.

	Dev Set				
Tasks	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