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

Yang Nan

Instructor: Ravi Tandon

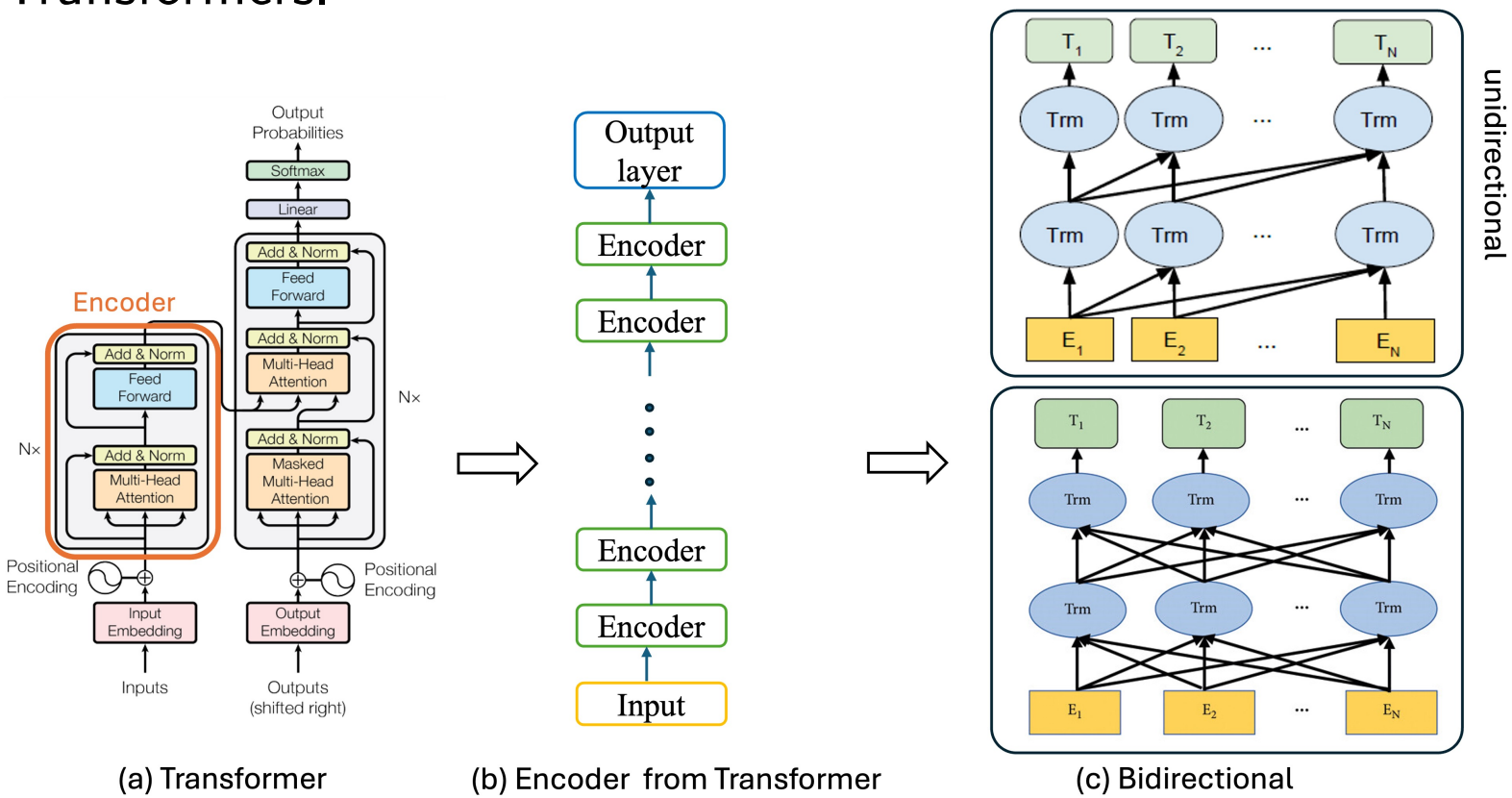
Jan. 28, 2025

What is BERT?

- BERT stands for Bidirectional Encoder Representations from Transformers.

What is BERT?

- BERT stands for Bidirectional Encoder Representations from Transformers.

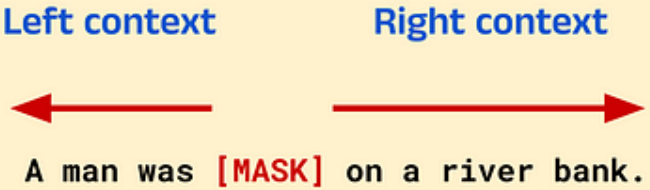


(a) Transformer (b) Encoder from Transformer (c) Bidirectional

Bidirectional vs Unidirectional

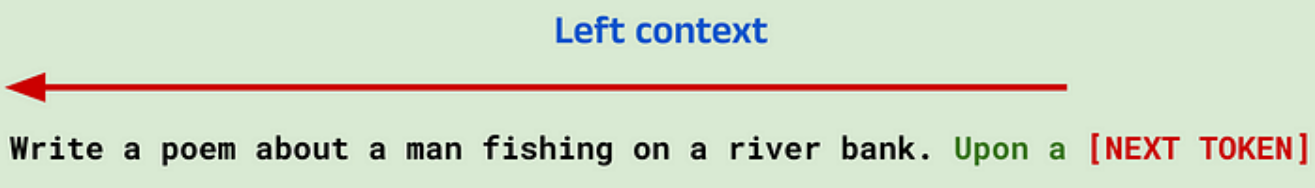
BIDIRECTIONAL CONTEXT

When predicting words within a sequence, all of the surrounding words can be used to gain contextual information.



UNIDIRECTIONAL CONTEXT

When predicting future words, only the previous words can be used to gain contextual information.



Why BERT?

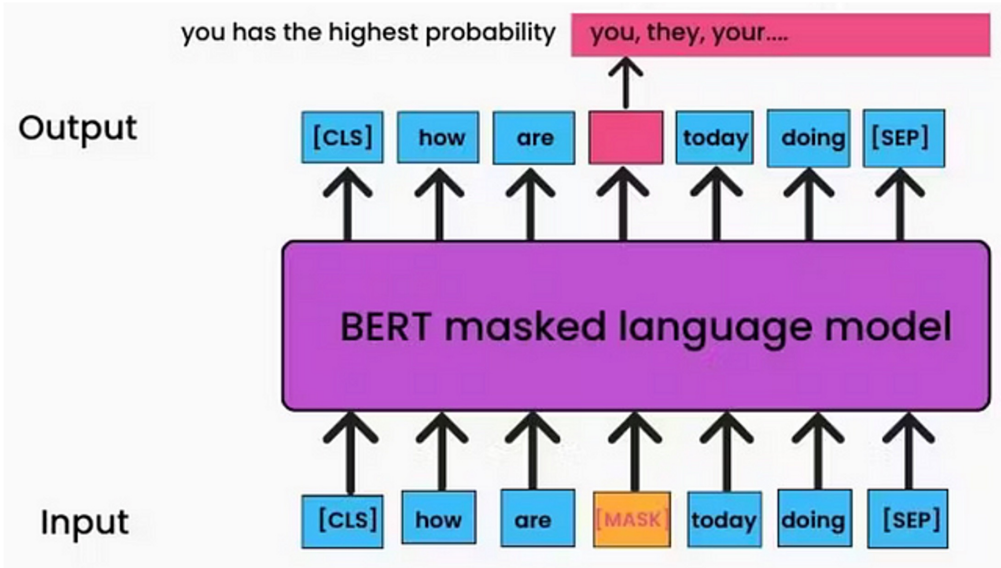
- It provides a pre-trained deep learning model designed for natural language understanding (NLU).
  - Vision Domain: Pre-trained models (e.g., ResNet) $\rightarrow$ Fine-tuned for tasks (e.g., classification, detection).
  - Pre-BERT NLP: Task-specific models $\rightarrow$ No universal pre-trained architecture.
  - With BERT: Universal pre-trained model $\rightarrow$ Fine-tuned for NLP tasks.

Why BERT?

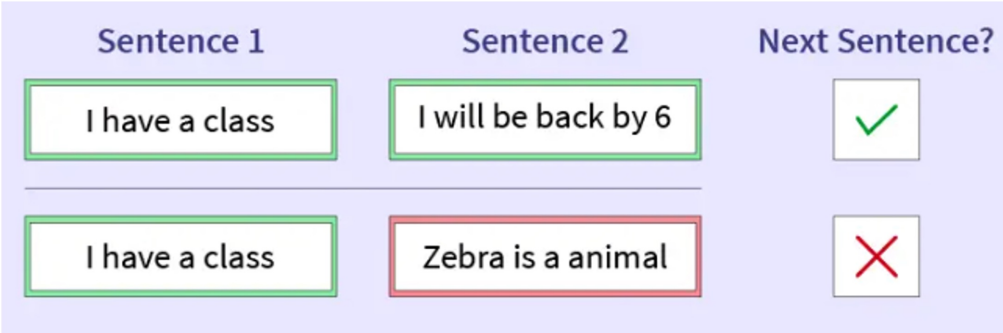
- It provides a pre-trained deep learning model designed for natural language understanding (NLU).
  - Vision Domain: Pre-trained models (e.g., ResNet) $\rightarrow$ Fine-tuned for tasks (e.g., classification, detection).
  - Pre-BERT NLP: Task-specific models $\rightarrow$ No universal pre-trained architecture.
  - With BERT: Universal pre-trained model $\rightarrow$ Fine-tuned for NLP tasks.
- Why BERT has natural language understanding?

How to pre-train BERT?

- Pre-train BERT using two unsupervised tasks:
  - Masked LM (MLM)
  - Next Sentence Prediction (NSP)



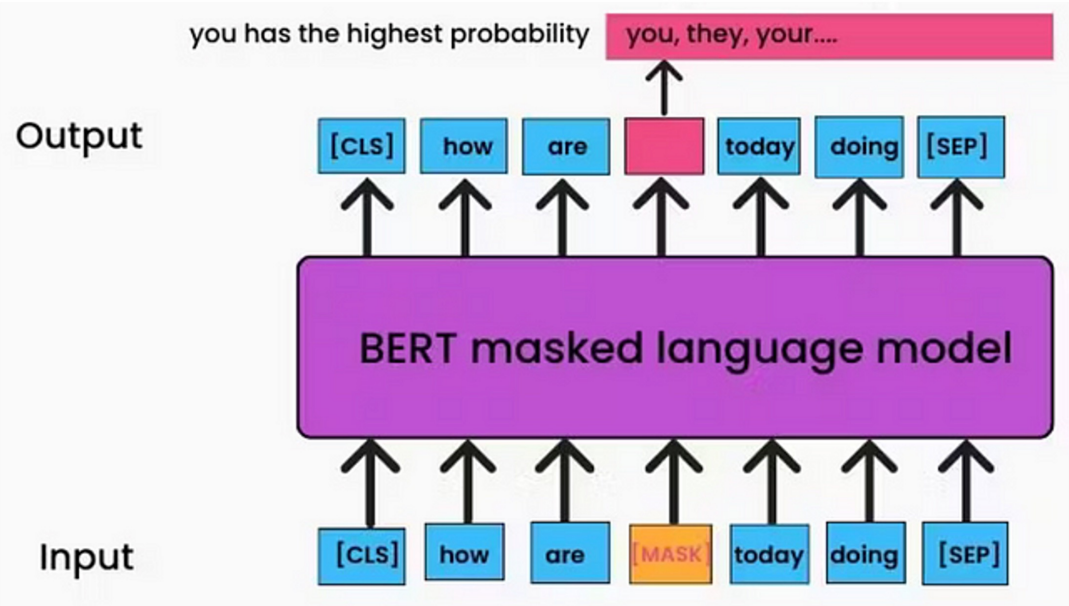
(a) MLM



(b) NSP

MLM

- MLM randomly masks 15% of the words in the input text and asks the model to predict these masked words based on the surrounding context, thereby learning deep semantic and contextual dependencies of words.

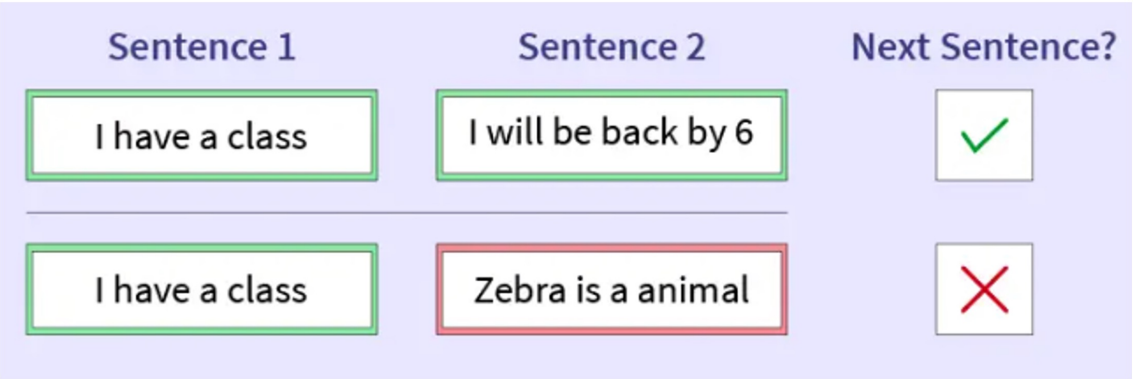


Masked LM and the Masking Procedure Assuming the unlabeled sentence is `my dog is hairy`, and during the random masking procedure we chose the 4-th token (which corresponding to `hairy`), our masking procedure can be further illustrated by

- 80% of the time: Replace the word with the `[MASK]` token, e.g., `my dog is hairy` → `my dog is [MASK]`
- 10% of the time: Replace the word with a random word, e.g., `my dog is hairy` → `my dog is apple`
- 10% of the time: Keep the word unchanged, e.g., `my dog is hairy` → `my dog is hairy`. The purpose of this is to bias the representation towards the actual observed word.

NSP

- NSP tasks the model with predicting whether a given pair of sentences is logically ordered, with 50% of the time the second sentence being randomly selected from other sentences, helping the model understand sentence-level relationships and context.



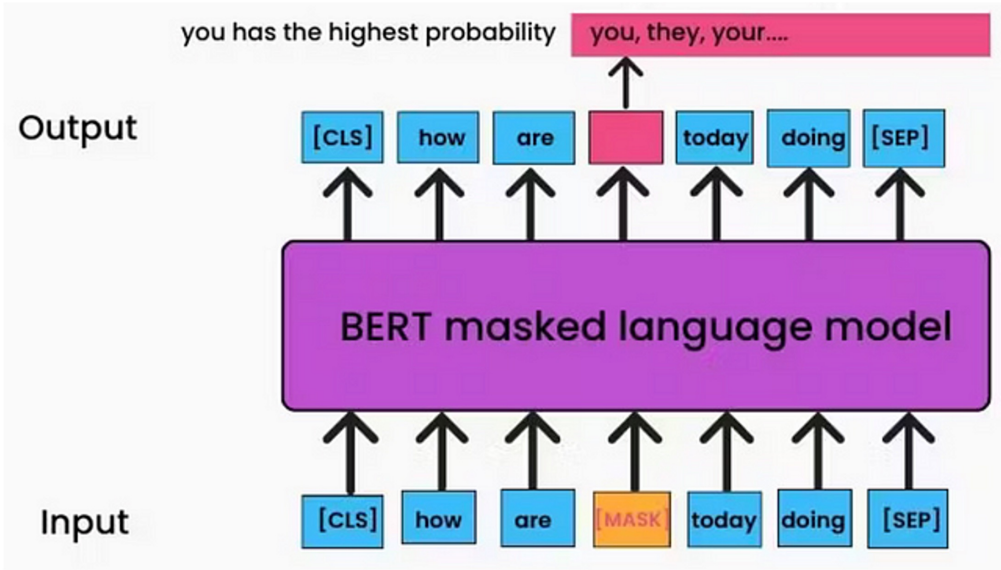
Next Sentence Prediction The next sentence prediction task can be illustrated in the following examples.

Input = `[CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]`
Label = `IsNext`

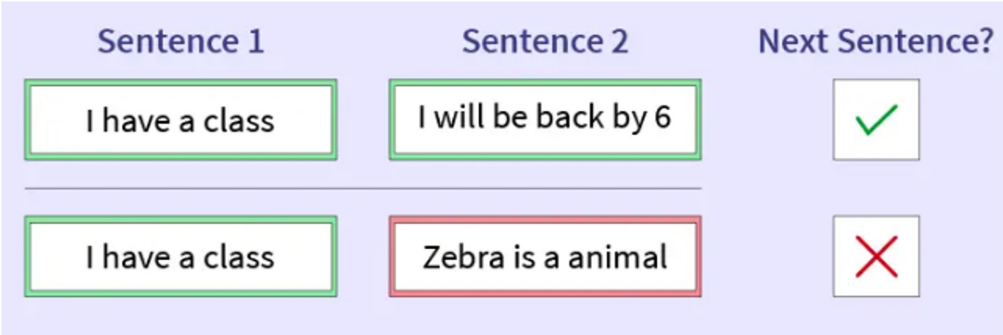
Input = `[CLS] the man [MASK] to the store [SEP] penguin [MASK] are flight ##less birds [SEP]`
Label = `NotNext`

How to pre-train BERT?

- Pre-train BERT using two unsupervised tasks:
  - Masked LM (MLM)
  - Next Sentence Prediction (NSP)
- Why BERT has natural language understanding?



(a) MLM

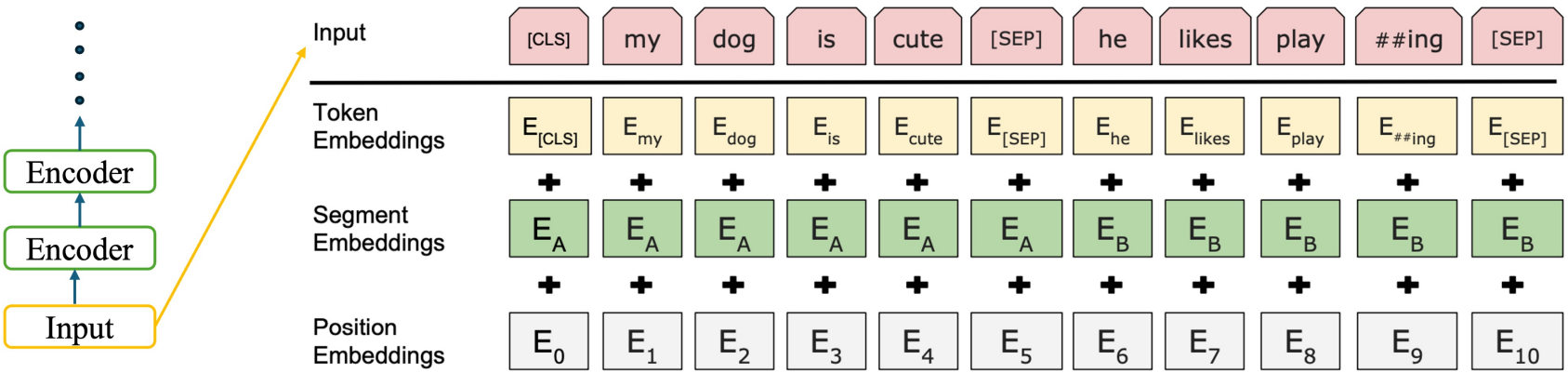


(b) NSP

Input Embeddings

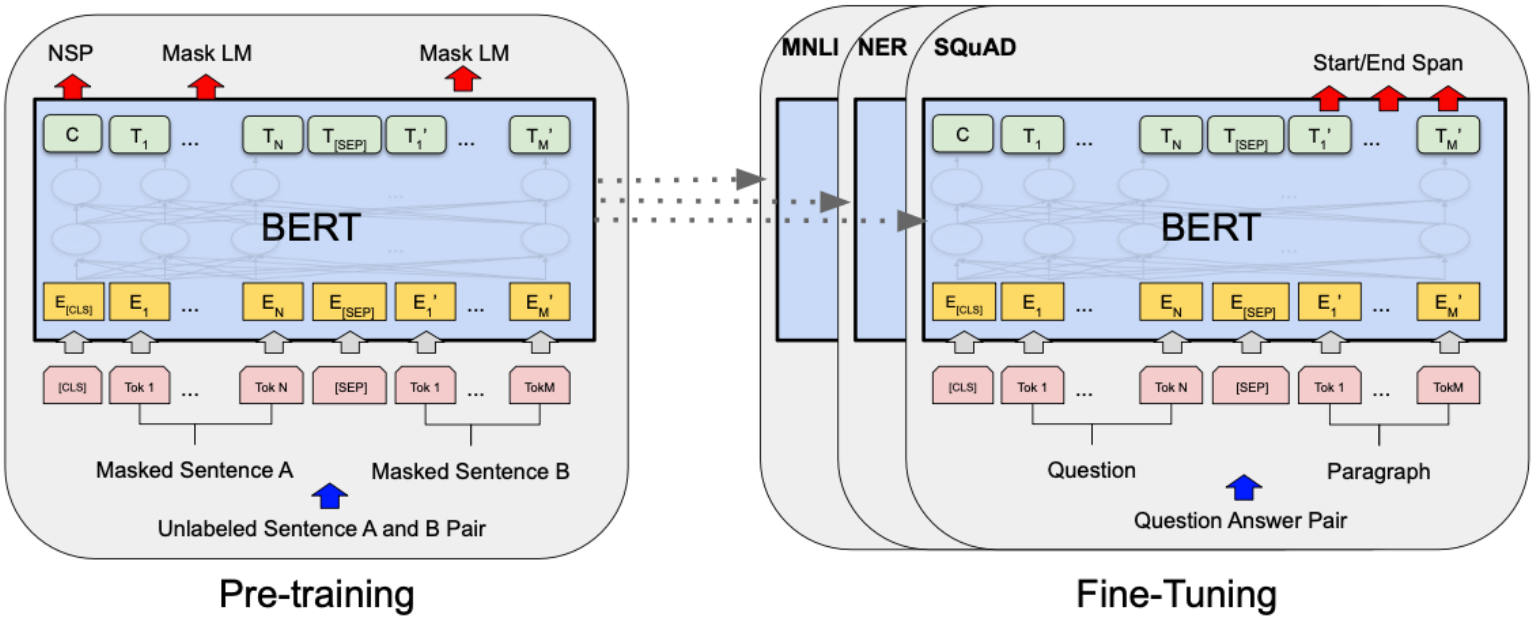
- The [CLS] token is added at the beginning of the input sequence to aggregate information for classification tasks, while the [SEP] token is used to separate different sentences.

Pre-training data The pre-training procedure largely follows the existing literature on language model pre-training. For the pre-training corpus we use the BooksCorpus (800M words) (Zhu et al., 2015) and English Wikipedia (2,500M words). For Wikipedia we extract only the text passages and ignore lists, tables, and headers. It is critical to use a document-level corpus rather than a shuffled sentence-level corpus such as the Billion Word Benchmark (Chelba et al., 2013) in order to extract long contiguous sequences.



How to fine-tune BERT?

- Fine-tuning BERT involves training the pre-trained BERT model on a specific downstream task (e.g., classification, question answering) by adding a task-specific layer on top and updating the entire model using labeled task data. This allows the model to adapt its general language understanding to the specific task at hand.



Experiments - GLUE

- The GLUE (General Language Understanding Evaluation) benchmark is a collection of diverse NLP tasks designed to evaluate the general understanding of language of a model. It includes tasks like textual entailment, sentiment analysis, and sentence similarity.

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
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**

Experiments - SQuAD v1.1

- SQuAD v1.1 (Stanford Question Answering Dataset) is a reading comprehension benchmark consisting of over 100,000 questions, where each question has an answer that is a text span extracted from the corresponding passage. It evaluates the ability of a model to understand and extract relevant information from context.

System	Dev		Test	
	EM	F1	EM	F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
Published				
BiDAF+ELMo (Single)	-	85.6	-	85.8
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

Experiments - SQuAD v2.0

- SQuAD v2.0: This dataset extends SQuAD v1.1 by introducing questions without answers in the passage, challenging models to identify when no answer exists while maintaining high performance on answerable questions.
- SWAG: The Situations With Adversarial Generations (SWAG) dataset tests the ability of a model to perform commonsense reasoning by selecting the most plausible continuation of a given scenario from multiple choices.

System	Dev		Test	
	EM	F1	EM	F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	86.3	89.0	86.9	89.5
#1 Single - MIR-MRC (F-Net)	-	-	74.8	78.0
#2 Single - nlnet	-	-	74.2	77.1
Published				
unet (Ensemble)	-	-	71.4	74.9
SLQA+ (Single)	-		71.4	74.4
Ours				
$\text{BERT}_{\text{LARGE}}$ (Single)	78.7	81.9	80.0	83.1

(a) SQuAD 2.0

System	Dev	Test
ESIM+GloVe	51.9	52.7
ESIM+ELMo	59.1	59.2
OpenAI GPT	-	78.0
$\text{BERT}_{\text{BASE}}$	81.6	-
$\text{BERT}_{\text{LARGE}}$	**86.6**	**86.3**
Human (expert)†	-	85.0
Human (5 annotations)†	-	88.0

(b) SWAG

Ablation Studies

- Ablation over the pre-training tasks using the BERTBASE 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 + NoNSP" model during fine-tuning.

	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

Conclusion

- Contributions:
  - BERT introduced a novel bidirectional Transformer-based pre-training framework, enabling deeper context understanding compared to previous unidirectional models.
  - It demonstrated state-of-the-art performance on a wide range of NLP tasks, such as GLUE, SQuAD, and SWAG, without task-specific architectures.
  - The use of Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) as pre-training objectives effectively improved natural language understanding
  - It leverages large-scale pre-training data, leading to significant performance gains across tasks..

Conclusion

- Limitations:
  - The bidirectional nature of BERT, while beneficial for understanding context, makes it unsuitable for generation tasks (e.g., text generation, machine translation). This fundamental limitation hinders its application in generative AI domains.
  - The lack of autoregressive capabilities prevents BERT from being used in tasks that require sequential prediction.
  - This limitation has had a significant impact, ultimately contributing to the inability of BERT to compete with newer models designed for both understanding and generation, such as GPT series or unified general-purpose models.

Thank you for listening