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

## [CSED490X] Recent Trends in ML: A Large-Scale Perspective

Jungtaek Kim

jtkim@postech.ac.kr

POSTECH
Pohang 37673, Republic of Korea
https://jungtaek.github.io

March 30, 2022

# Table of Contents

Introduction

Tasks & Datasets

A Machine Learning Model

A Learning Scheme

Experimental Results

# Introduction

# Bidirectional Encoder Representations from Transformers (BERT)

- BERT has been introduced in the work by Devlin et al. [2018].
- Unlike recent language representation models, it is designed to pre-train deep bidirectional representations.
- It is trained by unlabeled text by jointly conditioning on both left and right contexts.
- As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to solve down-stream tasks.

[Devlin et al., 2018] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

# Today's Lecture

- We will cover the tasks and datasets solved in this paper first.
- Before introducing the BERT model, we will visit a bidirectional RNN model.
- Eventually, we will study the BERT model and the details of learning schemes, used in this paper.
- Finally, we will investigate the experimental results.

# Tasks & Datasets

# Tasks & Datasets

- BERT is tested on four experimental circumstances:

1. General Language Understanding Evaluation (GLUE) benchmark;
2. Stanford Question Answering Dataset v1.1 (SQuAD v1.1);
3. Stanford Question Answering Dataset v2.0 (SQuAD v2.0);
4. Situations With Adversarial Generations (SWAG) dataset.

# General Language Understanding Evaluation (GLUE)

- The GLUE benchmark [Wang et al., 2019] is a collection of diverse natural language understanding tasks.
- **MNLI**
  - The Multi-Genre Natural Language Inference Corpus is a crowdsourced collection of sentence pairs with textual entailment annotations.
  - Given a premise sentence and a hypothesis sentence, the task is to predict whether the premise entails the hypothesis (entailment), contradicts the hypothesis (contradiction), or neither (neutral).
- **QQP**
  - The Quora Question Pairs dataset is a collection of question pairs from the community question-answering website Quora.
  - The task is to determine whether a pair of questions are semantically equivalent.

# General Language Understanding Evaluation (GLUE)

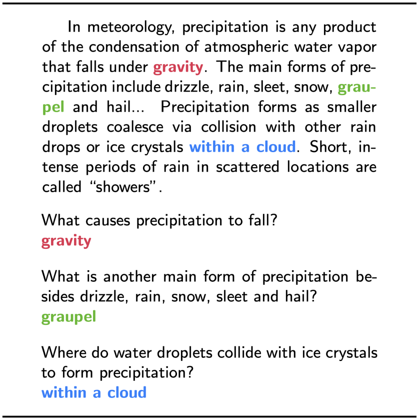
- The GLUE benchmark [Wang et al., 2019] is a collection of diverse natural language understanding tasks.
- **QNLI**
  - The Stanford Question Answering Dataset is a question-answering dataset consisting of question-paragraph pairs, where one of the sentences in the paragraph (drawn from Wikipedia) contains the answer to the corresponding question (written by an annotator).
  - The authors convert the task into sentence pair classification by forming a pair between each question and each sentence in the corresponding context, and filtering out pairs with low lexical overlap between the question and the context sentence.
  - The task is to determine whether the context sentence contains the answer to the question.

# General Language Understanding Evaluation (GLUE)

- The GLUE benchmark [Wang et al., 2019] is a collection of diverse natural language understanding tasks.
- **SST-2**
  - The Stanford Sentiment Treebank consists of sentences from movie reviews and human annotations of their sentiment.
  - The task is to predict the sentiment of a given sentence.

# Stanford Question Answering Dataset (SQuAD) v1.1 & v2.0

- SQuAD v1.1 [Rajpurkar et al., 2016] is a collection of 100k crowdsourced question and answer pairs.
- Given a question and a passage from Wikipedia containing the answer, the SQuAD v1.1 task is to predict the answer text span in the passage.
- The SQuAD v2.0 [Rajpurkar et al., 2018] task extends the SQuAD v1.1 problem definition by allowing for the possibility that no short answer exists in the provided paragraph, making the problem more realistic.



In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
**gravity**

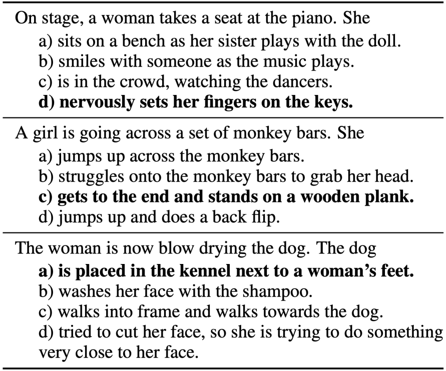
What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
**graupel**

Where do water droplets collide with ice crystals to form precipitation?
**within a cloud**

**Figure 1:** Question-answer pairs for a sample passage in the SQuAD dataset. Each of the answers is a segment of text from the passage.

# Situations With Adversarial Generations (SWAG)

- SWAG [Zellers et al., 2018] contains 113k sentence-pair completion examples that evaluate grounded common-sense inference.
- Given a sentence, the task is to choose the most plausible continuation among four choices.



On stage, a woman takes a seat at the piano. She
a) sits on a bench as her sister plays with the doll.
b) smiles with someone as the music plays.
c) is in the crowd, watching the dancers.
**d) nervously sets her fingers on the keys.**

A girl is going across a set of monkey bars. She
a) jumps up across the monkey bars.
b) struggles onto the monkey bars to grab her head.
**c) gets to the end and stands on a wooden plank.**
d) jumps up and does a back flip.

The woman is now blow drying the dog. The dog
**a) is placed in the kennel next to a woman's feet.**
b) washes her face with the shampoo.
c) walks into frame and walks towards the dog.
d) tried to cut her face, so she is trying to do something very close to her face.

# A Machine Learning Model

# Main Ideas

- **Bidirectional representation**: A left-to-right architecture, at which every token can only attend to previous tokens, is limited, because some tasks require incorporating context from both directions.
- **Pre-training, then fine-tuning**: It reduces the need for many heavily-engineered task-specific architectures.

# Bidirectional Recurrent Neural Networks

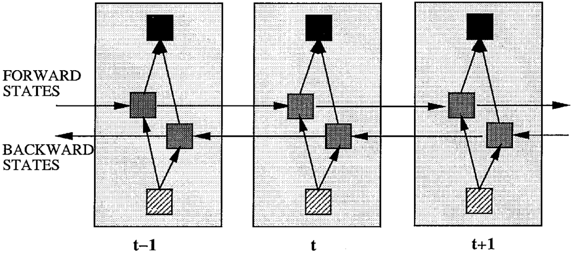


Figure 1: Illustration of bidirectional recurrent neural networks.

- It has a recurrent connection, similar to the vanilla recurrent neural network.
- Compared to the vanilla recurrent neural network, a sequence of input instances is processed by considering forward and backward directions.

Figure 1 is taken from [Schuster and Paliwal, 1997].

[Schuster and Paliwal, 1997] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11): 2673–2681, 1997.

# Transformer

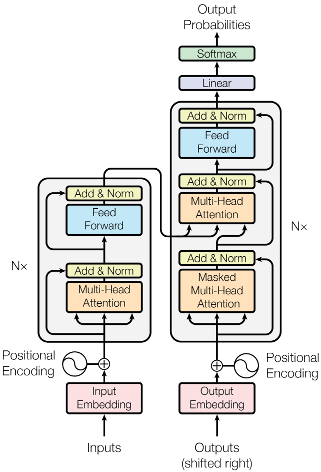
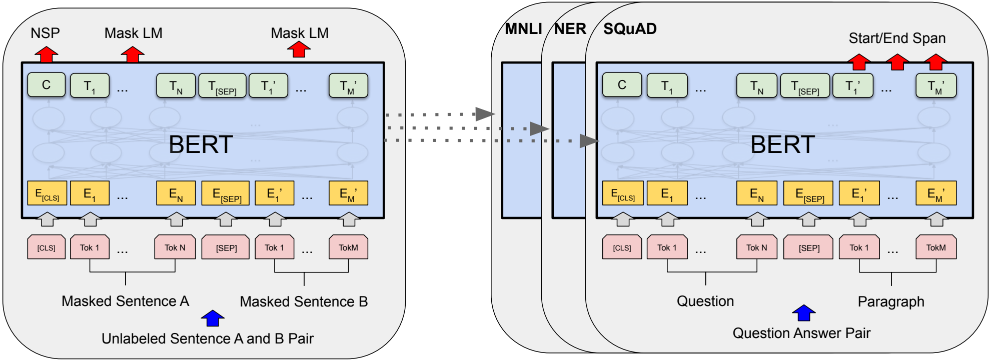


Figure 2: Transformer.

Figure 2 is taken from [Vaswani et al., 2017].

# BERT



Pre-training

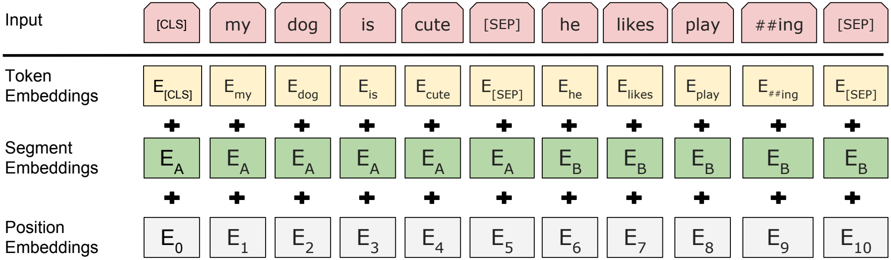
Fine-Tuning

# BERT

- The BERT architecture is a multi-layer bidirectional Transformer encoder based on the original implementation described in Vaswani et al. [2017].
- In this work, denote the number of layers (i.e., Transformer blocks) as $L$, the hidden size as $H$, and the number of self-attention heads as $A$.
- **BASE** model: $L = 12$, $H = 768$, $A = 12$, Total Parameters $= 110$M.
- **LARGE** model: $L = 24$, $H = 1024$, $A = 16$, Total Parameters $= 340$M.

[Vaswani et al., 2017] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS), volume 30, pages 5998–6008, Long Beach, California, USA, 2017.

# Input Representations



| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 | $E_0$ | $E_1$ | $E_2$ | $E_3$ | $E_4$ | $E_5$ | $E_6$ | $E_7$ | $E_8$ | $E_9$ | $E_{10}$ |

- Both a single sentence and a pair of sentences, e.g., $\langle$Question, Answer$\rangle$ are represented in one token sequence.
- The WordPiece embeddings with a 30,000 token vocabulary are used.
- Some special tokens, e.g., `[SEP]` and `[CLS]`, exist.

# A Learning Scheme

# Pre-Training BERT

- BERT is pre-trained using two unsupervised tasks.
- **Task #1 Masked Language Model**: Simply mask some percentage of the input tokens at random, and then predict those masked tokens.
- **Task #2 Next Sentence Prediction**: Given $\mathrm{A}$ and $\mathrm{B}$ sentences contiguously, discriminate whether $\mathrm{B}$ is the next sentence of $\mathrm{A}$, or not.

# Masked Language Model

- 15% of all WordPiece tokens are masked in each sequence at random.
- If some token is chosen, the selected token is replaced with the `[MASK]` token or a random word, or unchanged.
- For example,

  80% of the time: Replace the word with the `[MASK]` token

  `my dog is hairy` → `my dog is [MASK]`,

  10% of the time: Replace the word with a random word

  `my dog is hairy` → `my dog is apple`,

  10% of the time: Keep the word unchanged

  `my dog is hairy` → `my dog is hairy`.

# Next Sentence Prediction

- Many important down-stream tasks such as question answering and natural language inference are based on understanding the relationship between two sentences.

- For example,

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

  Label: `IsNext`

  or

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

  Label: `NotNext`

# Pre-Training BERT

- Training of $\mathrm{BERT}_{\mathrm{BASE}}$ was performed on 4 Cloud TPUs in Pod configuration (16 TPU chips total).

- Training of $\mathrm{BERT}_{\mathrm{LARGE}}$ was performed on 16 Cloud TPUs (64 TPU chips total).

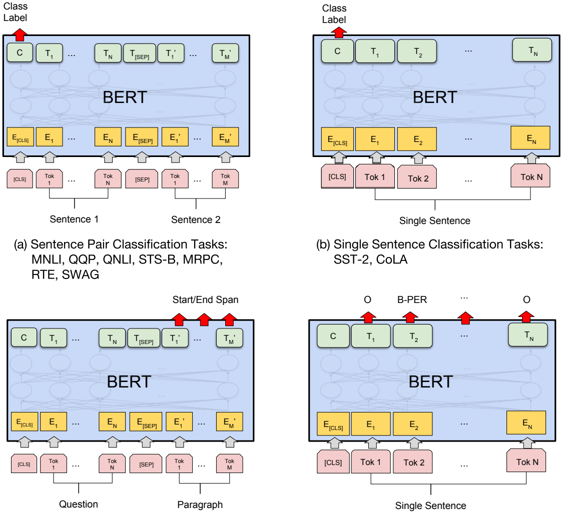
- Each pre-training took 4 days to complete.

- To speed up pre-training, the model is pre-trained with sequence length of 128 for 90% of the steps. Then, it is trained with sequence of 512 for the rest 10% of the steps to learn the positional embeddings.

# Fine-Tuning BERT

- Fine-tuning is straightforward since the self-attention mechanism allows BERT to model many downstream tasks.
- BERT uses the self-attention mechanism by encoding a concatenated text pair to effectively include bidirectional cross attention between two sentences.
- For each task, we simply plug in the task-specific inputs and outputs into BERT and fine-tune all the parameters end-to-end.
- All of the results in the paper can be replicated in at most 1 hour on a single Cloud TPU, or a few hours on a GPU, starting from the exact same pre-trained model.

# Fine-Tuning BERT



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

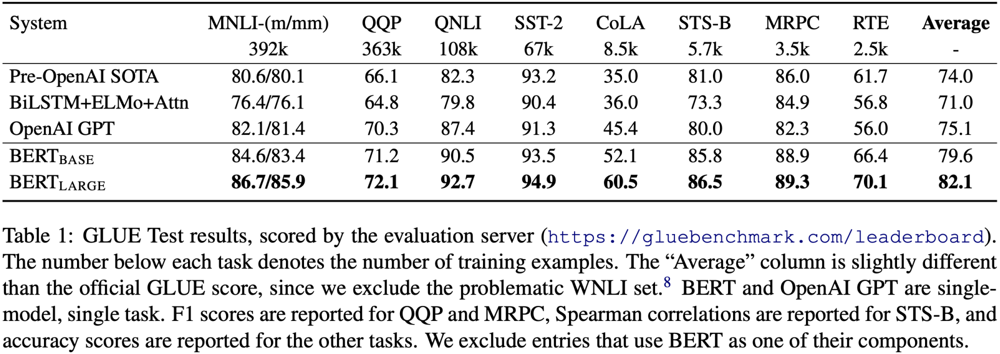
(b) Single Sentence Classification Tasks:
SST-2, CoLA

(c) Question Answering Tasks:
SQuAD v1.1

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

# Experimental Results

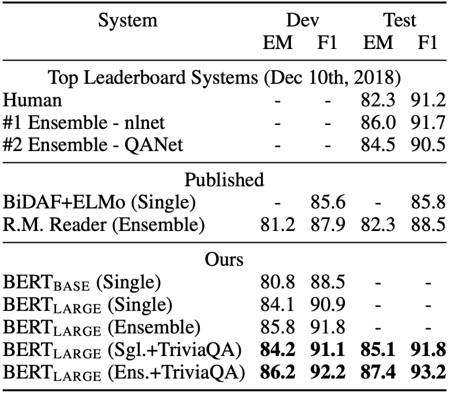
# Experimental Results



| 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 | 79.8 | 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** |

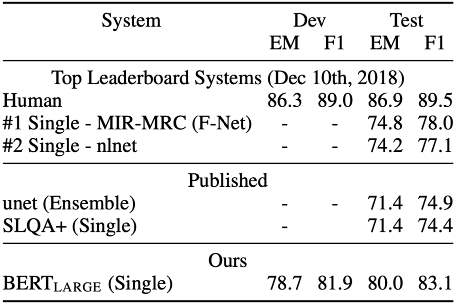
Table 1: GLUE Test results, scored by the evaluation server (https://gluebenchmark.com/leaderboard). The number below each task denotes the number of training examples. The "Average" column is slightly different than the official GLUE score, since we exclude the problematic WNLI set.[8] BERT and OpenAI GPT are single-model, single task. F1 scores are reported for QQP and MRPC, Spearman correlations are reported for STS-B, and accuracy scores are reported for the other tasks. We exclude entries that use BERT as one of their components.

# Experimental Results



| 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 | | | | |
| BERT$_{\text{BASE}}$ (Single) | 80.8 | 88.5 | - | - |
| BERT$_{\text{LARGE}}$ (Single) | 84.1 | 90.9 | - | - |
| BERT$_{\text{LARGE}}$ (Ensemble) | 85.8 | 91.8 | - | - |
| BERT$_{\text{LARGE}}$ (Sgl.+TriviaQA) | **84.2** | **91.1** | **85.1** | **91.8** |
| BERT$_{\text{LARGE}}$ (Ens.+TriviaQA) | **86.2** | **92.2** | **87.4** | **93.2** |

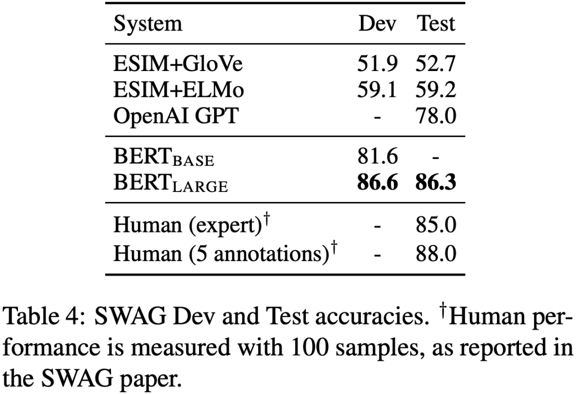
Table 2: SQuAD 1.1 results. The BERT ensemble is 7x systems which use different pre-training checkpoints and fine-tuning seeds.



| 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 | | | | |
| BERT$_{\text{LARGE}}$ (Single) | 78.7 | 81.9 | 80.0 | 83.1 |

Table 3: SQuAD 2.0 results. We exclude entries that use BERT as one of their components.

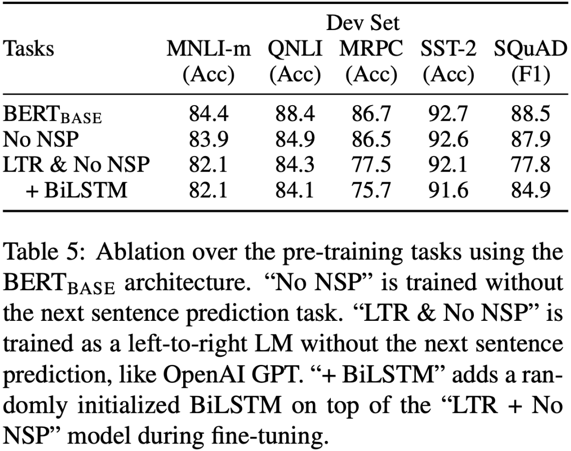
# Experimental Results



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

Table 4: SWAG Dev and Test accuracies. †Human performance is measured with 100 samples, as reported in the SWAG paper.

# Experimental Results



| | Dev Set | | | | |
|---|---|---|---|---|---|
| Tasks | MNLI-m<br>(Acc) | QNLI<br>(Acc) | MRPC<br>(Acc) | SST-2<br>(Acc) | SQuAD<br>(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.

# Any Questions?

# References I

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

P. Rajpurkar, R. Jia, and P. Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 784–789, 2018. Short Paper.

M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008, Long Beach, California, USA, 2017.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.