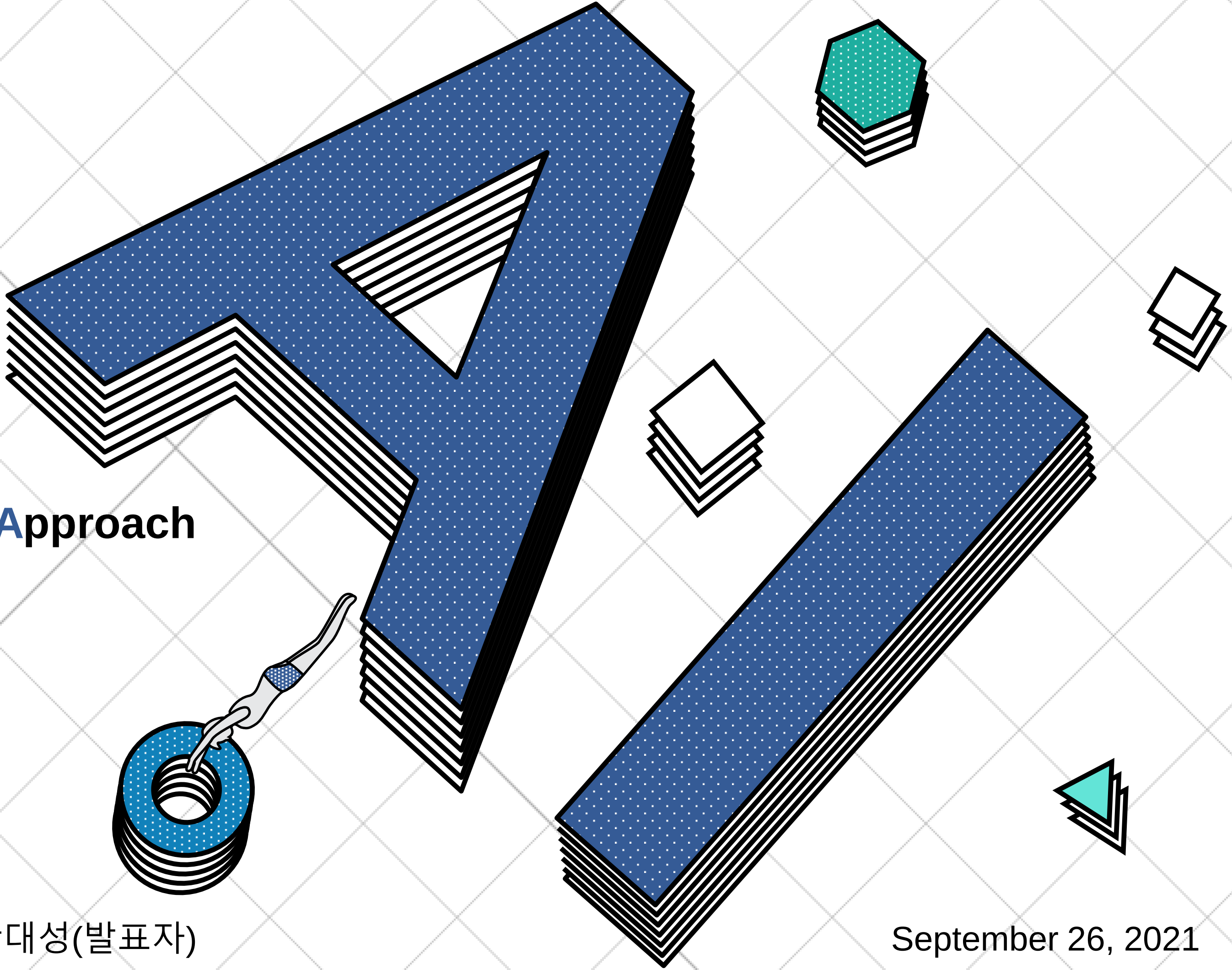
RoBERTa

A Robustly Optimized BERT Pretraining Approach



JipHyeonJeon Study

집현전 중급반 15조 : 김영재(발표자), 김재희(발표자), 왕대성(발표자)

September 26, 2021

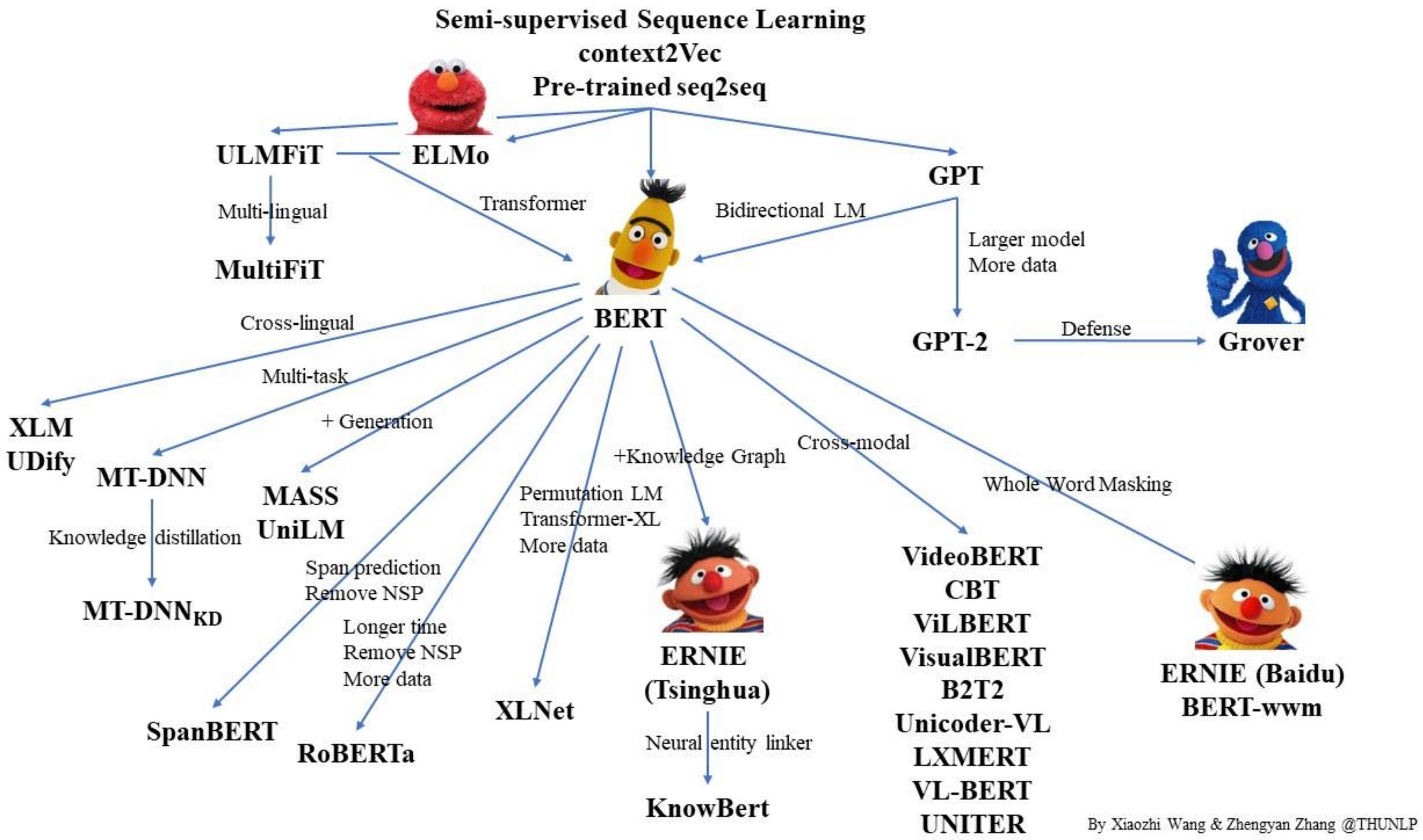
CONTENTS

/01

Background

Background

- Self-training methods have brought significant performance gains.



https://blog.hsutimes.com/2019/10/20/PLMpapers/

Background

- But it can be **challenging to determine which aspects of the methods contribute the most**.
  - Training is computationally expensive
  - Limiting the amount of tuning that can be done, and is often done with private training data of varying sizes,
  - Limiting our ability to measure the effects of the modeling advances.

- We present a replication study of BERT, which includes a **careful evaluation of the effects of hyperparameter tuning and training set size**.
  - Batch Size
  - Epoch
  - Learning Rate
  - Dataset
  - Objective function

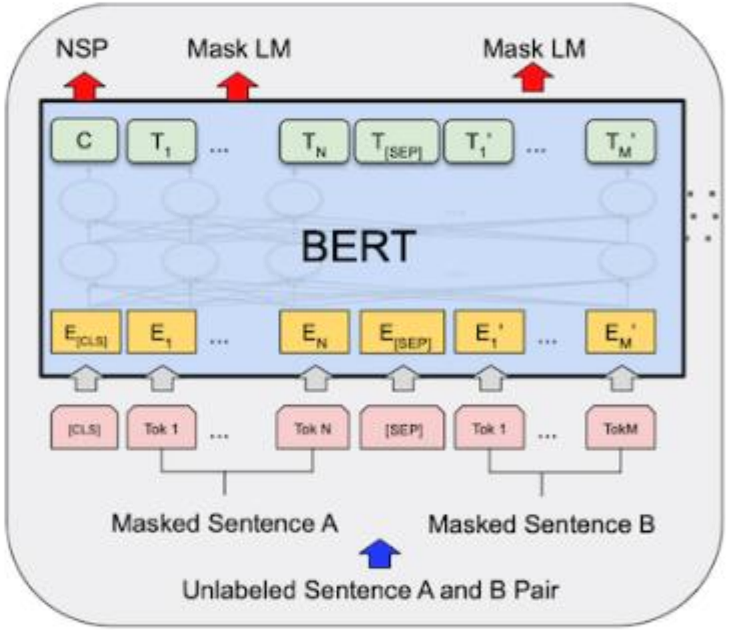
Background

- **BERT was significantly undertrained**

- Proposal
  - Training the model longer, with bigger batches, over more data
  - Removing the next sentence prediction objective
  - Training on longer sequences
  - Dynamically changing the masking pattern applied to the training data
  - Text Encoding

Background of BERT

- BERT takes as input a **concatenation of two segments**.
  - $[CLS], x_1, \dots, x_N, [SEP], y_1, \dots y_M, [SEP]$
- Training Objectives
  - Masked Language Model (MLM)
  - Next Sentence Prediction (NSP)
- Optimization
  - Adam
  - dropout of 0.1 on all layers and attention weights
  - GELU activation function
  - 1,000,000 updates
  - 256 minibatch
  - 512 tokens.
- Data (16 GB)
  - BOOKCORPUS
  - English Wikipedia



Background of BERT

- Masked Language Model (MLM)

  - 12% : Original → [MASK]
  - 1.5 % : Original → Alternatives

Make	everyday	cooking	fun
▽		▽	
Masking		Random	
▽		▽	
[MASK]	everyday	swimming	fun

Background of BERT

- Next Sentence Prediction (NSP)
  - NSP : **binary classification loss for predicting whether two segments follow each other**
  - Positive examples : consecutive sentences from the text corpus.
  - Negative examples : pairing segments from different documents.
  - The **NSP objective was designed to improve performance on downstream tasks**, such as Natural Language Inference, **which require reasoning about the relationships between pairs of sentences**.

IsNext	The man entered a university. $[SEP]$ He studied mathematics.
NotNext	I went to the park. $[SEP]$ Japan is a country

Background of BERT

A B

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

https://paul-hyun.github.io/bert-01/

/02

RoBERTa

What is Different

- Training the model longer, with bigger batches, over more data
- Removing the next sentence prediction objective
- Training on longer sequences
- Dynamically changing the masking pattern applied to the training data\
- Text Encoding

Training the model longer, with bigger batches, over more data

- we compare perplexity and end-task performance of $\text{BERT}_{\text{BASE}}$ as **we increase the batch size, controlling for the number of passes through the training data**.
- We observe **that training with large batches improves perplexity for the masked language modeling objective**, **as well as end-task accuracy**.
- Large batches **are also easier to parallelize via distri**. **buted data parallel training**, and in later experiments **we train with batches of 8K sequences**

Training the model longer, with bigger batches, over more data

bsz	steps	lr	ppl	MNLI-m	SST-2
256	1M	1e-4	3.99	84.7	92.7
2K	125K	7e-4	**3.68**	**85.2**	**92.9**
8K	31K	1e-3	3.77	84.6	92.8

Table 3: Perplexity on held-out training data (*ppl*) and development set accuracy for base models trained over BOOKCORPUS and WIKIPEDIA with varying batch sizes (*bsz*). We tune the learning rate (*lr*) for each setting. Models make the same number of passes over the data (epochs) and have the same computational cost.

Training the model longer, with bigger batches, over more data

- Pre-Train Data
  - BOOKCORPUS plus English WIKIPEDIA (16GB)
    - original data used to train BERT.
  - CC-NEWS (76GB)
    - collected from the English portion of the Common Crawl News dataset
    - The data contains 63 million English news articles crawled between September 2016 and February 2019.
  - OPENWEBTEXT (38GB)
    - open-source recreation of the WebText corpus described in GPT2. (2019)
    - The text is web content extracted from URLs shared on Reddit with at least three upvotes.
  - STORIES (31GB)
    - a dataset introduced in Trinh and Le (2018) containing a subset of Common Crawl data filtered to match the story-like style of Winograd schemas.

Training the model longer, with bigger batches, over more data

Model	data	bsz	steps	SQuAD (v1.1/2.0)	MNLI-m	SST-2
RoBERTa						
with Books + Wiki	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	**94.6/89.4**	**90.2**	**96.4**
$\text{BERT}_{\text{LARGE}}$						
with Books + Wiki	13GB	256	1M	90.9/81.8	86.6	93.7
$\text{XLNet}_{\text{LARGE}}$						
with Books + Wiki	13GB	256	1M	94.0/87.8	88.4	94.4
+ additional data	126GB	2K	500K	94.5/88.8	89.8	95.6

Table 4: Development set results for RoBERTa as we pretrain over more data (16GB → 160GB of text) and pretrain for longer (100K → 300K → 500K steps). Each row accumulates improvements from the rows above. RoBERTa matches the architecture and training objective of $\text{BERT}_{\text{LARGE}}$. Results for $\text{BERT}_{\text{LARGE}}$ and $\text{XLNet}_{\text{LARGE}}$ are from Devlin et al. (2019) and Yang et al. (2019), respectively. Complete results on all GLUE tasks can be found in the Appendix.

Training the model longer, with bigger batches, over more data

Model	data	bsz	steps	SQuAD (v1.1/2.0)	MNLI-m	SST-2
RoBERTa						
with BOOKS + WIKI	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	**94.6/89.4**	**90.2**	**96.4**
$\text{BERT}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	90.9/81.8	86.6	93.7
$\text{XLNet}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	94.0/87.8	88.4	94.4
+ additional data	126GB	2K	500K	94.5/88.8	89.8	95.6

Table 4: Development set results for RoBERTa as we pretrain over more data (16GB → 160GB of text) and pretrain for longer (100K → 300K → 500K steps). Each row accumulates improvements from the rows above. RoBERTa matches the architecture and training objective of $\text{BERT}_{\text{LARGE}}$. Results for $\text{BERT}_{\text{LARGE}}$ and $\text{XLNet}_{\text{LARGE}}$ are from Devlin et al. (2019) and Yang et al. (2019), respectively. Complete results on all GLUE tasks can be found in the Appendix.

Removing the next sentence prediction objective

- The NSP loss was hypothesized to be an important factor in training the original BERT model. Devlin et al. observe that removing NSP hurts performance, with **significant performance degradation on QNLI, MNLI, and SQuAD 1.1**.

Tasks	MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
	Dev Set				
$\text{BERT}_{\text{BASE}}$	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.

Removing the next sentence prediction objective

- The NSP loss was hypothesized to be an important factor in training the original BERT model. Devlin et al. observe that removing NSP hurts performance, with **significant performance degradation on QNLI, MNLI, and SQuAD 1.1**.
- However, some recent work(XLNet) has questioned the necessity of the NSP loss.

Removing the next sentence prediction objective

- To better understand this discrepancy, we compare several alternative training formats:
  - SEGMENT-PAIR+NSP
  - SENTENCE-PAIR+NSP
  - FULL-SENTENCES
  - DOC-SENTENCES

Removing the next sentence prediction objective

- SEGMENT-PAIR+NSP
    - the **original input format** used in BERT, with the NSP loss.
    - Each input has a pair of segments, which can **each contain multiple natural sentences**, but the total combined length must be less than 512 tokens

Removing the next sentence prediction objective

- SEGMENT-PAIR+NSP
  - the **original input format** used in BERT, with the NSP loss.
  - Each input has a pair of segments, which can **each contain multiple natural sentences**, but the total combined length must be less than 512 tokens

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task specific architecture modifications.

Language model pretraining has led to significant performance gains but careful comparison between different approaches is challenging. Training is computationally expensive, often done on private datasets of different sizes, and, as we will show, hyperparameter choices have significant impact on the final results. We present a replication study of BERT pretraining that carefully measures the impact of many key hyperparameters and training data size. We find that BERT was significantly undertrained, and can match or exceed the performance of every model published after it. Our best model achieves state-of-the-art results on GLUE, RACE and SQuAD.

Removing the next sentence prediction objective

- SEGMENT-PAIR+NSP
  - the **original input format** used in BERT, with the NSP loss.
  - Each input has a pair of segments, which can **each contain multiple natural sentences**, but the total combined length must be less than 512 tokens

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task specific architecture modifications.

Language model pretraining has led to significant performance gains but careful comparison between different approaches is challenging. Training is computationally expensive, often done on private datasets of different sizes, and, as we will show, hyperparameter choices have significant impact on the final results. We present a replication study of BERT pretraining that carefully measures the impact of many key hyperparameters and training data size. We find that BERT was significantly undertrained, and can match or exceed the performance of every model published after it. Our best model achieves state-of-the-art results on GLUE, RACE and SQuAD.

Removing the next sentence prediction objective

- SENTENCE-PAIR+NSP
  - Each input contains **a pair of natural sentences**, either sampled from a contiguous portion of one document or from separate documents.
  - Since these inputs are significantly shorter than 512 tokens, we **increase the batch size so that the total number of tokens remains similar to SEGMENT-PAIR+NSP**.

Removing the next sentence prediction objective

- SENTENCE-PAIR+NSP
  - Each input contains **a pair of natural sentences**, either sampled from a contiguous portion of one document or from separate documents.
  - Since these inputs are significantly shorter than 512 tokens, we **increase the batch size so that the total number of tokens remains similar to SEGMENT-PAIR+NSP**.

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task specific architecture modifications.

Language model pretraining has led to significant performance gains but careful comparison between different approaches is challenging. Training is computationally expensive, often done on private datasets of different sizes, and, as we will show, hyperparameter choices have significant impact on the final results. We present a replication study of BERT pretraining that carefully measures the impact of many key hyperparameters and training data size. We find that BERT was significantly undertrained, and can match or exceed the performance of every model published after it. Our best model achieves state-of-the-art results on GLUE, RACE and SQuAD.

Removing the next sentence prediction objective

- SENTENCE-PAIR+NSP
  - Each input contains **a pair of natural sentences**, either sampled from a contiguous portion of one document or from separate documents.
  - Since these inputs are significantly shorter than 512 tokens, we **increase the batch size so that the total number of tokens remains similar to SEGMENT-PAIR+NSP**.

> We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task specific architecture modifications.

> Language model pretraining has led to significant performance gains but careful comparison between different approaches is challenging. Training is computationally expensive, often done on private datasets of different sizes, and, as we will show, hyperparameter choices have significant impact on the final results. We present a replication study of BERT pretraining that carefully measures the impact of many key hyperparameters and training data size. We find that BERT was significantly undertrained, and can match or exceed the performance of every model published after it. Our best model achieves state-of-the-art results on GLUE, RACE and SQuAD.

Removing the next sentence prediction objective

- FULL-SENTENCES
  - Each input is packed with **full sentences sampled contiguously from one or more documents**, such that the total length is at most 512 tokens.
  - When we **reach the end of one document**, we begin sampling sentences from the next document and add an extra separator token between documents.
  - We **remove the NSP loss**.

Removing the next sentence prediction objective

- FULL-SENTENCES
  - Each input is packed with **full sentences sampled contiguously from one or more documents**, such that the total length is at most 512 tokens.
  - When we **reach the end of one document**, we begin sampling sentences from the next document and add an extra separator token between documents.
  - We **remove the NSP loss**.

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task specific architecture modifications.

Language model pretraining has led to significant performance gains but careful comparison between different approaches is challenging. Training is computationally expensive, often done on private datasets of different sizes, and, as we will show, hyperparameter choices have significant impact on the final results. We present a replication study of BERT pretraining that carefully measures the impact of many key hyperparameters and training data size. We find that BERT was significantly undertrained, and can match or exceed the performance of every model published after it. Our best model achieves state-of-the-art results on GLUE, RACE and SQuAD.

Removing the next sentence prediction objective

- DOC-SENTENCES
  - Inputs are constructed similarly to FULL-SENTENCES, except that **they may not cross document boundaries**.
  - Inputs **sampled near the end of a document may be shorter than 512 tokens**, so **we dynamically increase the batch size in these cases to achieve a similar number of total tokens** as FULLSENTENCES.
  - We **remove the NSP loss**.

Removing the next sentence prediction objective

- DOC-SENTENCES
  - Inputs are constructed similarly to FULL-SENTENCES, except that **they may not cross document boundaries**.
  - Inputs **sampled near the end of a document may be shorter than 512 tokens**, so **we dynamically increase the batch size in these cases to achieve a similar number of total tokens** as FULLSENTENCES.
  - We **remove the NSP loss**.

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task specific architecture modifications.

Language model pretraining has led to significant performance gains but careful comparison between different approaches is challenging. Training is computationally expensive, often done on private datasets of different sizes, and, as we will show, hyperparameter choices have significant impact on the final results. We present a replication study of BERT pretraining that carefully measures the impact of many key hyperparameters and training data size. We find that BERT was significantly undertrained, and can match or exceed the performance of every model published after it. Our best model achieves state-of-the-art results on GLUE, RACE and SQuAD.

Removing the next sentence prediction objective

Model	SQuAD 1.1/2.0	MNLI-m	SST-2	RACE
Our reimplementation (with NSP loss):				
SEGMENT-PAIR	90.4/78.7	84.0	92.9	64.2
SENTENCE-PAIR	88.7/76.2	82.9	92.1	63.0
Our reimplementation (without NSP loss):				
FULL-SENTENCES	90.4/79.1	84.7	92.5	64.8
DOC-SENTENCES	90.6/79.7	84.7	92.7	65.6
$\text{BERT}_{\text{BASE}}$	88.5/76.3	84.3	92.8	64.3
$\text{XLNet}_{\text{BASE}}$ (K = 7)	–/81.3	85.8	92.7	66.1
$\text{XLNet}_{\text{BASE}}$ (K = 6)	–/81.0	85.6	93.4	66.7

Table 2: Development set results for base models pretrained over BOOKCORPUS and WIKIPEDIA. All models are trained for 1M steps with a batch size of 256 sequences. We report F1 for SQuAD and accuracy for MNLI-m, SST-2 and RACE. Reported results are medians over five random initializations (seeds). Results for $\text{BERT}_{\text{BASE}}$ and $\text{XLNet}_{\text{BASE}}$ are from Yang et al. (2019).

Removing the next sentence prediction objective

- This setting **outperforms the originally published BERT$_{BASE}$ results** and that **removing the NSP loss matches or slightly improves downstream task performance**, in contrast to Devlin et al..
- Finally we find **that restricting sequences to come from a single document performs slightly better than packing sequences from multiple documents**.
- However, because the **DOC-SENTENCES format results in variable batch sizes**, we use **FULL-SENTENCES** in the remainder of our experiments for easier comparison with related work.

Training on longer sequences

- We pretrain with **sequences of at most T = 512 tokens**. Unlike Devlin et al., we **do not randomly inject short sequences**, and **we do not train with a reduced sequence length for the first 90% of updates**.
    - BERT : To speed up pretraining in our experiments, we **pre-train the model with sequence length of 128 for 90% of the steps**. Then, we **train the rest 10% of the steps of sequence of 512** to learn the positional embeddings.
- We train **only with full-length sequences**.

Dynamically changing the masking pattern applied to the training data

- The original BERT implementation **performed masking once during data preprocessing, resulting in a single static mask**.
- **To avoid using the same mask for each training instance in every epoch**, training data was **duplicated 10 times** so that **each sequence is masked in 10 different ways over the 40 epochs of training**.
- Thus, **each training sequence was seen with the same mask four times during training**.

Dynamically changing the masking pattern applied to the training data

- We compare this strategy with **dynamic masking** where **we generate the masking pattern every time we feed a sequence to the model**.
- This becomes **crucial when pretraining for more steps or with larger datasets**.

Masking	SQuAD 2.0	MNLI-m	SST-2
reference	76.3	84.3	92.8
Our reimplementation:			
static	78.3	84.3	92.5
dynamic	78.7	84.0	92.9

Table 1: Comparison between static and dynamic masking for $\text{BERT}_{\text{BASE}}$. We report F1 for SQuAD and accuracy for MNLI-m and SST-2. Reported results are medians over 5 random initializations (seeds). Reference results are from Yang et al. (2019).

Text Encoding

- BPE achieving **slightly worse end-task performance** on some tasks.
- Nevertheless, we believe **the advantages of a universal encoding scheme outweighs the minor degradation in performance** and use this encoding in the remainder of our experiments.
  - Using bytes makes **it possible to learn a subword vocabulary of a modest size (50K units) that can still encode any input text without introducing any “unknown” tokens**

	Unit	Vocabulary Size
BERT	Unicode Character	30K
GPT2	Byte	50K

/03

Result

Evaluation Dataset

- GLUE
- SQuAD
- RACE

Evaluation Dataset

- GLUE
  - The General Language Understanding Evaluation (GLUE) benchmark is a collection of **9 datasets for evaluating natural language understanding systems**.
  - Tasks are framed as either **single-sentence classification** or **sentence-pair classification tasks**.
    - Single-sentence : CoLA, SST-2
    - Sentence-pair : MRPC, QQP, STS-B, MNLI, RTE, QNLI, WNLI

Evaluation Dataset

- GLUE

	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	WNLI	Avg
Single-task single models on dev										
$\text{BERT}_{\text{LARGE}}$	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-	-
$\text{XLNet}_{\text{LARGE}}$	89.8/-	93.9	91.8	83.8	95.6	89.2	63.6	91.8	-	-
RoBERTa	**90.2/90.2**	**94.7**	**92.2**	**86.6**	**96.4**	**90.9**	**68.0**	**92.4**	**91.3**	-
Ensembles on test (from leaderboard as of July 25, 2019)										
ALICE	88.2/87.9	95.7	**90.7**	83.5	95.2	92.6	**68.6**	91.1	80.8	86.3
MT-DNN	87.9/87.4	96.0	89.9	86.3	96.5	92.7	68.4	91.1	89.0	87.6
XLNet	90.2/89.8	98.6	90.3	86.3	**96.8**	**93.0**	67.8	91.6	**90.4**	88.4
RoBERTa	**90.8/90.2**	**98.9**	90.2	**88.2**	96.7	92.3	67.8	**92.2**	89.0	**88.5**

Table 5: Results on GLUE. All results are based on a 24-layer architecture. $\text{BERT}_{\text{LARGE}}$ and $\text{XLNet}_{\text{LARGE}}$ results are from Devlin et al. (2019) and Yang et al. (2019), respectively. RoBERTa results on the development set are a median over five runs. RoBERTa results on the test set are ensembles of *single-task* models. For RTE, STS and MRPC we finetune starting from the MNLI model instead of the baseline pretrained model. Averages are obtained from the GLUE leaderboard.

Evaluation Dataset

- SQuAD
  - The Stanford Question Answering Dataset (SQuAD) **provides a paragraph of context and a question**.
  - The task is **to answer the question by extracting the relevant span from the context**.
  - SQuAD V1.1
    - the context always contains an answer
  - SQuAD V2.0
    - some questions are not answered in the provided context.

Evaluation Dataset

- SQuAD

Model	SQuAD 1.1 EM	SQuAD 1.1 F1	SQuAD 2.0 EM	SQuAD 2.0 F1
Single models on dev, w/o data augmentation				
$\text{BERT}_{\text{LARGE}}$	84.1	90.9	79.0	81.8
$\text{XLNet}_{\text{LARGE}}$	**89.0**	94.5	86.1	88.8
RoBERTa	88.9	**94.6**	**86.5**	**89.4**
Single models on test (as of July 25, 2019)				
$\text{XLNet}_{\text{LARGE}}$			86.3†	89.1†
RoBERTa			86.8	89.8
XLNet + SG-Net Verifier			**87.0**†	**89.9**†

Table 6: Results on SQuAD. † indicates results that depend on additional external training data. RoBERTa uses only the provided SQuAD data in both dev and test settings. $\text{BERT}_{\text{LARGE}}$ and $\text{XLNet}_{\text{LARGE}}$ results are from Devlin et al. (2019) and Yang et al. (2019), respectively.

Evaluation Dataset

- RACE
  - The ReAding Comprehension from Examinations (RACE) task is a large-scale **reading comprehension dataset** with more than **28,000 passages and nearly 100,000 questions**.
  - The dataset is collected from English examinations in China, which are designed for middle and high school students.
  - In RACE, each passage is associated with multiple questions.
  - For every question, the task **is to select one correct answer from four options**.
  - RACE has **significantly longer context than other popular reading comprehension datasets** and **the proportion of questions that requires reasoning is very large**.

Evaluation Dataset

- RACE

Model	Accuracy	Middle	High
Single models on test (as of July 25, 2019)			
$\text{BERT}_{\text{LARGE}}$	72.0	76.6	70.1
$\text{XLNet}_{\text{LARGE}}$	81.7	85.4	80.2
RoBERTa	**83.2**	**86.5**	**81.3**

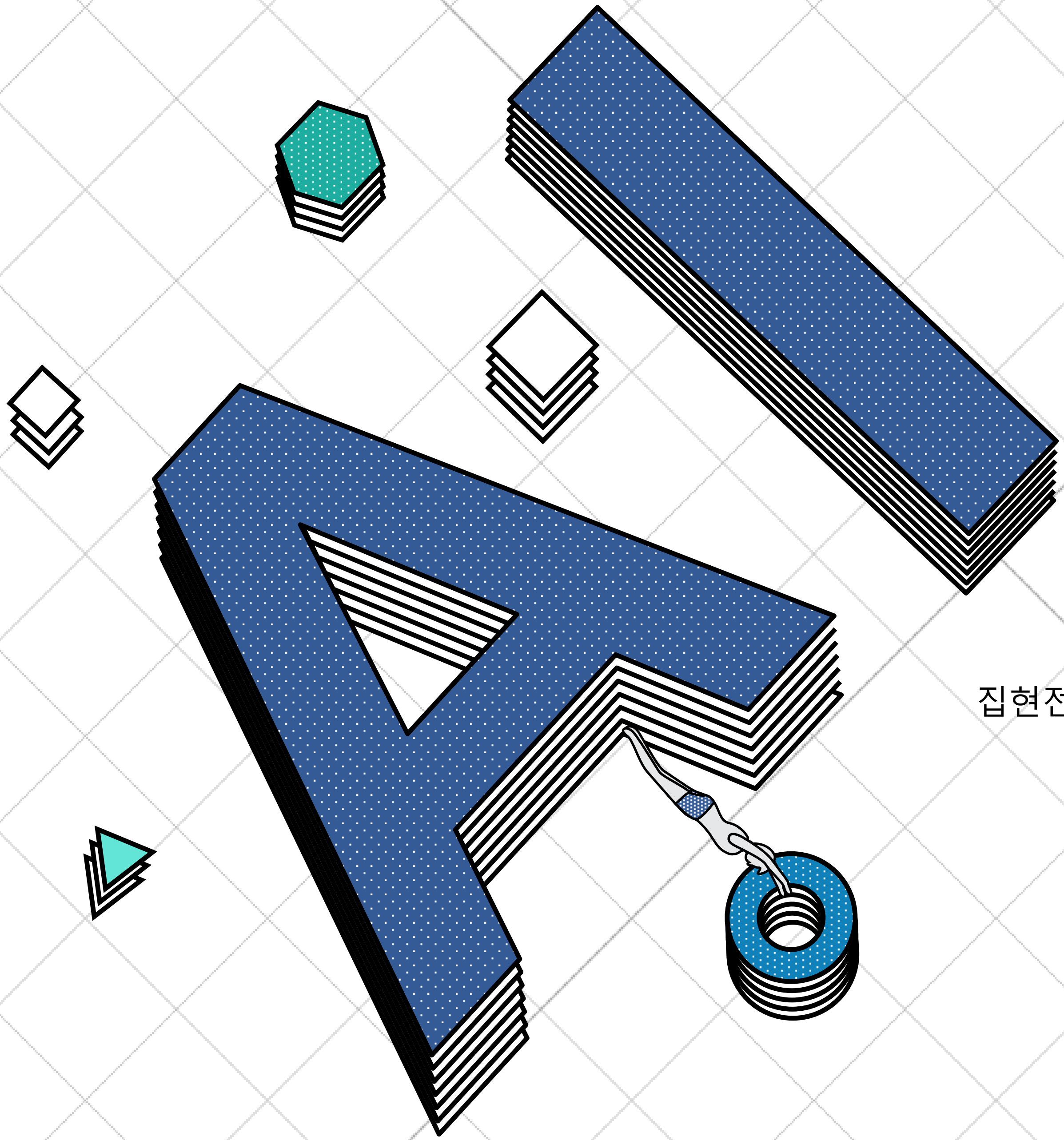
Table 7: Results on the RACE test set. $\text{BERT}_{\text{LARGE}}$ and $\text{XLNet}_{\text{LARGE}}$ results are from Yang et al. (2019).

/04

Conclusion

Conclusion

- We carefully evaluate a number of design decisions when pretraining BERT models.
- We find that **performance can be substantially improved by training the model longer, with bigger batches over more data; removing the next sentence prediction objective; training on longer sequences; and dynamically changing the masking pattern applied to the training data**.
- Our improved pretraining procedure, which we call RoBERTa, **achieves state-of-the-art results** on GLUE, RACE and SQuAD, without multi-task finetuning for GLUE or additional data for SQuAD.
- These results illustrate the **importance of these previously overlooked design decisions** and suggest that **BERT’s pretraining objective remains competitive with recently proposed alternatives**.



Thanks

집현전 중급반 15조 : 김영재(발표자), 김재희(발표자), 왕대성(발표자)

September 26, 2021