# A Robustly Optimized BERT Pre-training Approach

**Name**
조병웅
NLP
2024/04/30

TRAIN AND TEST

# Contents



TRAIN AND TEST

# Introduction

## BERT was significantly undertrained(underfit)!

### Solution -> Change the Training Procedures

- Training the model longer with more data
- Training the model with bigger batches
- Removing the next sentence prediction objective(NSP)
- Training on longer sequneces
- Dynamically changing the masking pattern
- Text Econding

TRAIN AND TEST

# Training Procedure Analysis

## Dynamic Masking

**- BERT Implementation performed masking once during data preprocessing (static)**

- Data was duplicated 10 times so that each sequence is masked 10 different ways over 40 epoch
- Each training sequence was seen with the same mask four times during training

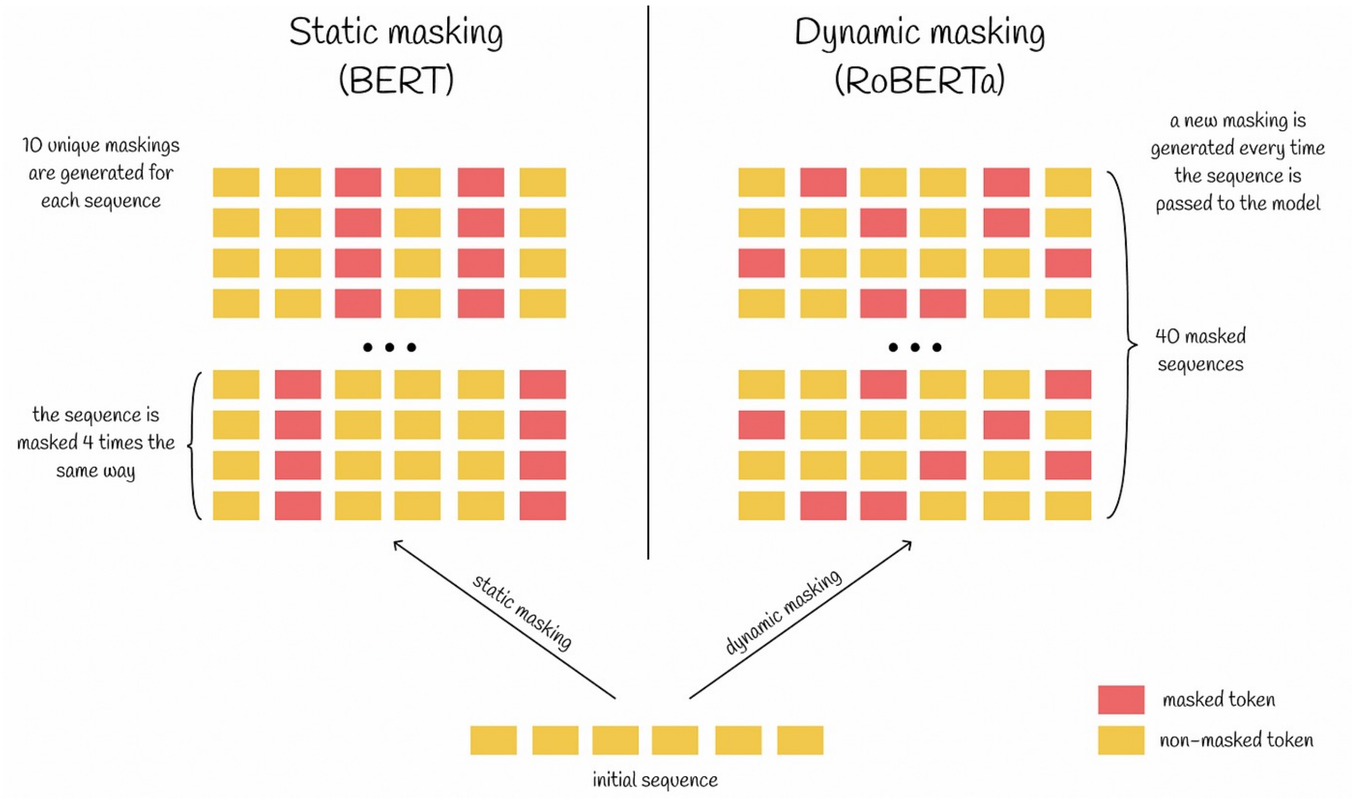
**- Dynamic Masking : generate the masking pattern every time feeding a sequence**

- Crucial when pretraining for more steps or larger datasets
- Comparable or slightly better than static masking

TRAIN AND TEST

# Training Procedure Analysis

## Dynamic Masking




TRAIN AND TEST

# Training Procedure Analysis

**Dynamic Masking**

| 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).

TRAIN AND TEST

# Training Procedure Analysis

## Model Input Format and NSP

**NSP에 대한 의문 제기 -> 검증을 위한 학습 포맷 비교 (기존의 변형)**

**# SEGMENT – PAIR + NSP / 기존 BERT 입력 포맷, 총길이 512 토큰 이하, 연속된 문장**

**->** 아침에 일어났다. 이상하게 배가 아팠다. 오늘은 왠지 술을 한잔 하고 싶다. 비가 오고 일도 잘 안된다.

**# SENTENCE – PAIR + NSP / 문장 두개만 이용, 문장 단위, 512보다 훨씬 짧음 -> 배치사이즈를 증가시킴**

**->** 아침에 일어났다. 이상하게 배가 아팠다.

**# FULL - SENTENCES (NO NSP) / 임의의 segment를 한 개 이상의 문단에서 가져옴, 512까지 잇기**

**->** 왠지 모르게 다 내려놓고 놀고 싶은 날이다. 우와 엄마가 선물을 줬다

**# DOC – SENTENCES (NO NSP) / 한 개의 문단에서만 데이터를 가져옴 / 512보다 짧을 가능성 있음 -> 배치사이즈를 증가시킴**

**->** 병원은 너무 멀었다. 차로 한참을 가다가 나는 잠이 들고 말았다.

TRAIN AND TEST

# Training Procedure Analysis

## Model Input Format and NSP

Individual Sentence = low performance
➔ Not able to learn long-range dependencies

**But**, No NSP is better than NSP predictions

| 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 |
| $BERT_{BASE}$ | 88.5/76.3 | 84.3 | 92.8 | 64.3 |
| $XLNet_{BASE}$ (K = 7) | –/81.3 | 85.8 | 92.7 | 66.1 |
| $XLNet_{BASE}$ (K = 6) | –/81.0 | 85.6 | 93.4 | 66.7 |

TRAIN AND TEST

# Training Procedure Analysis

## Training with large Batches

**Previous Research ->**
BERT is also amenable to large batch training

| 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.

**Q1. Why are large batch training good?**

**Q2. Why did they choose 8K?**

TRAIN AND TEST

# Training Procedure Analysis

**Text Encoding**

**Byte-Pair Encoding (BPE) is a hybrid between character- and word-level representations.**

➔ No OOV
➔ Original BERT = 30k vocab size / RoBERTa = 50k vocab size

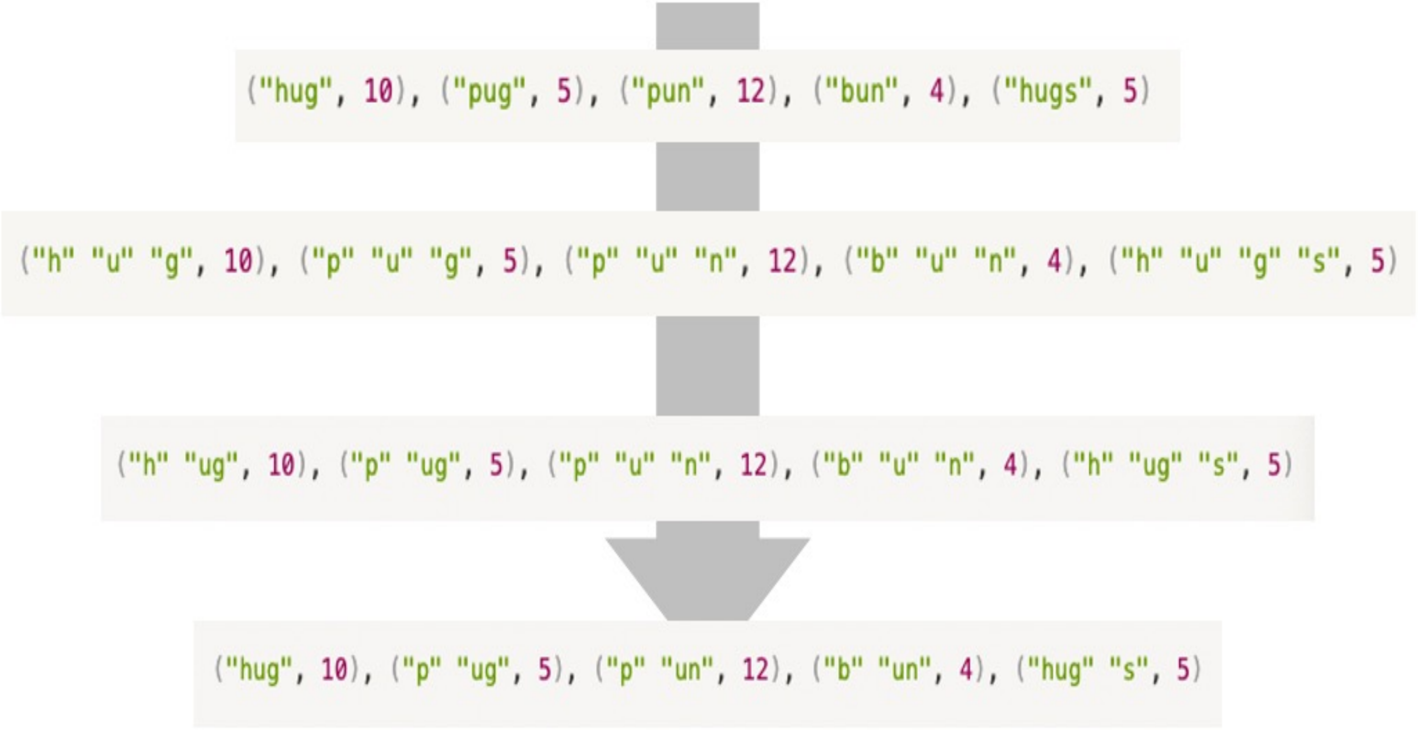
TRAIN AND TEST

# Training Procedure Analysis

**Text Encoding**

**Byte-Pair Encoding (BPE)**
**: Subword tokenizer**

**Step1. Pre-tokenize**
**Step2. Calculate frequency**
**Step3. Merge from big one**
**Step4. return step1 until size**




TRAIN AND TEST

# LoBERTa

➔ In the previous section we learn about **modifications to the BERT pretraining procedure** that improve end-task performance.

**aggregate these improvement ->**

**RoBERTa(Robustly optimized BERT approach)**

- Dynamic Masking
- NO NSP
- Large Batch
- Large BPE

# LoBERTa

## → Performance comparison

| 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 |
| $BERT_{LARGE}$ | | | | | | |
| with BOOKS + WIKI | 13GB | 256 | 1M | 90.9/81.8 | 86.6 | 93.7 |
| $XLNet_{LARGE}$ | | | | | | |
| 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 |

# LoBERTa

## ➔ Performance comparison

| | 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** |

| 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** |

| Model | SQuAD 1.1 | | SQuAD 2.0 | |
|---|---|---|---|---|
| | EM | F1 | EM | 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)* | | | | |
| $\text{XLNet}_{\text{LARGE}}$ | | | $86.3^{\dagger}$ | $89.1^{\dagger}$ |
| RoBERTa | | | 86.8 | 89.8 |
| XLNet + SG-Net Verifier | | | **$87.0^{\dagger}$** | **$89.9^{\dagger}$** |

# Conclusion

- Presents several ways for improve the existing model.
- The longest, the more, and the largest are important
- There is a need to focus on training rather than structure.
- BERT’s pre-training objective is still competitive.

TRAIN AND TEST

TRAIN AND TEST