RoBERTa: A Robustly Optimized BERT Pretraining Approach

Paper By

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du et al., 2019

Presented By

Yash Aggarwal and Harsh Gunwant

UC RIVERSIDE

1 Line Paper Summary

BERT is undertrained. We will fine-tune the original BERT and change input and pre-training settings to achieve state-of-the-art results

~ Authors of the paper

UC RIVERSIDE

Index

Abstract

- Language model pretraining has led to significant performance gains.
- Several challenges when comparing these approaches
  - Computationally Expensive
  - Private Datasets
  - Hyperparameter Choices
- BERT was significantly undertrained
- Authors believe to match or exceed the performance after BERT (including GPT and GPT2)

Introduction

- Self-Training Models like BERT, BART, GPT are all the rage right now.
- Large Model == Interpretability Issue
- Language models go through 2 steps of training
  - Pre-training
  - Fine-Tuning
- Authors believed BERT was undertrained and propose a new training method in the paper
- **R**obustly **o**ptimized **BERT** **a**pproach or **RoBERTa**

Introduction (2)

Modifications Suggested for RoBERTa

- Training the model longer, with bigger batches, and more data
- Removing the NSP objective
- Training on longer Sequences
- Using dynamic masking as a pre-training objective instead of static masking

Contributions of the paper

- BERT Design Choices
- Novel Dataset
- Dynamic Masking as pre-training
- Pre-trained and fine-tuned model

Background

- Setup
- Architecture
- Training Objectives
- Dataset

Experimental Setup

BERT Implementation

- BERT was implemented in FAIRSEQ with original architecture and hyperparameters
- Changes were made to
  - Peak Learning Rate
  - Number of warm up steps
  - Adam Epsilon Term
  - β2 Term was changed from 0.999 to 0.98
- Training is done with full length sequences unlike BERT that randomly injects short sequences.
- Training is done on 8x32GB NVIDIA V100 GPUs

Experimental Setup (2)

Dataset Changes

- Pre-training of large language models depend heavily on dataset.
- Authors tried to create a new dataset that is combination of the following 5 datasets
  - BOOKCORPUS + English Wikipedia - 16GB
  - CC - News - 76GB
  - Open WebText - 38GB
  - STORIES - 31GB

  Total ~160 GB

Experimental Setup (3)

Evaluation

- After pre-training RoBERTa fine-tuning is same as that of BERT
- The model is evaluated on following downstream tasks similar to BERT

GLUE

- Collection of 9 datasets for single-sentence or sentence-pair classification

SQuAD

- Answer question by extraction relevant span from context

RACE

- 28,000 passages with 100,000 questions.
- Context is longer.

Training Procedure Analysis

Base Configuration

- RoBERTa is trained by keeping the same configuration as $\text{BERT}_{\text{Base}}$
  - Number of Layers = 12
  - Embedding Dimensions = 768
  - Attention Heads = 12
  - Number of parameters = 110 Million

Training Procedure Analysis (2)

Static Vs 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

Training Procedure Analysis (3)

Model Input Format and NSP

- Next Sentence Prediction Task in BERT and recent developments
  - SEGMENT - PAIR + NSP
  - SENTENCE - PAIR + NSP
  - FULL - SENTENCES
  - DOC - SENTENCES

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

Training Procedure Analysis (4)

Training with larger batches

- Past works have shown that NMTs trained on large mini-batches have better performance.
- Same is the case with BERT
- BERT was trained for 1M steps with batch size of 256
- Authors try to find a sweet spot for different steps and batch sizes

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

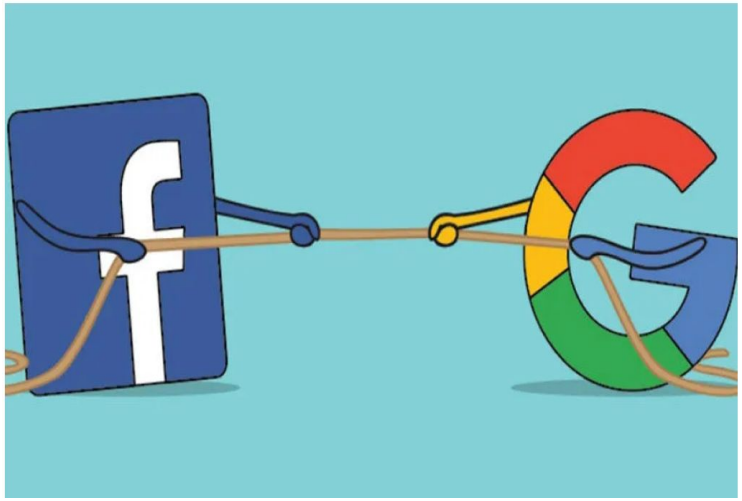
Training Procedure Analysis (5)

Text Encoding

- Character Level Byte Pair Encoding
- BERT has a vocabulary of 30k
- RoBERTa has a vocab of 50k words
- Even though evidence suggests this is slightly worse, authors believe a standard encoding scheme is better than slight performance boosts.

There is room for improvement in BERT !!

RoBERTa



- Trained with dynamic masking
- Full Sentences without NSP Loss
- Large Mini- Batches
- More Training Data (16Gb vs 160Gb)
- Larger byte level BPE
- Training the model Longer

COMPARISON BETWEEN BERT AND RoBERTa

	BERT	RoBERTa
Size (millions)	**Base**: 110 **Large**: 340	**Base**: 110 **Large**: 340
Training Time	**Base**: 8 x V100 x 12 days* **Large:** 64 TPU Chips x 4 days (or 280 x V100 x 1 days*)	**Large:** 1024 x V100 x 1 day; 4-5 times more than BERT.
Performance	Outperforms state-of-the-art in Oct 2018	2-20% improvement over BERT
Data	16 GB BERT data (Books Corpus + Wikipedia). 3.3 Billion words.	160 GB (16 GB BERT data + 144 GB additional)
Method	BERT (Bidirectional Transformer with MLM and NSP)	BERT without NSP**

Development set results for RoBERTa

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

GLUE

- GLUE (General Language Understanding Evaluation) is a benchmark for evaluating the performance of natural language processing models on a diverse range of tasks.

- The benchmark provides a suite of tasks that cover a broad range of linguistic phenomena and include tasks such as sentiment analysis, question answering, and text classification.

- Two fine tuning settings are considered for GLUE - Single task,dev and ensembles, test

GLUE RESULTS

	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**

SQuAD RESULTS:

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**†

UC RIVERSIDE

RACE RESULTS:

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**

CONCLUSION:

- When pretraining BERT models, a number of design decisions were evaluated, and it was determined that BERT's performance can be significantly enhanced by **training the model longer, with larger batches over more data; removing the next sentence prediction objective; training on longer sequences; and dynamically adjusting the masking.**
- RoBERTa achieves state-of-the-art results on GLUE, RACE, and SQuAD without the need for multi-task fine-tuning for GLUE or additional data for SQuAD.
- These results demonstrate the significance of these formerly neglected design decisions and show that BERT's pretraining objective remains competitive in comparison to recently presented alternatives.

THANK YOU !!

UC RIVERSIDE