RoBERTa: A Robustly Optimized BERT Pretraining Approach

Idea: Improve BERT performance

Bidirectional **E**ncoder **R**epresentations from **T**ransformers
Popular self-training model
Goal: Increase performance with optimized pretraining

BERT – Setup & Architecture

Takes two concatenated sequences of segments (more than one natural sentence) $x_1, \ldots, x_N$ and $y_1, \ldots, y_M$

$M + N < T$ (maximum sequence length)

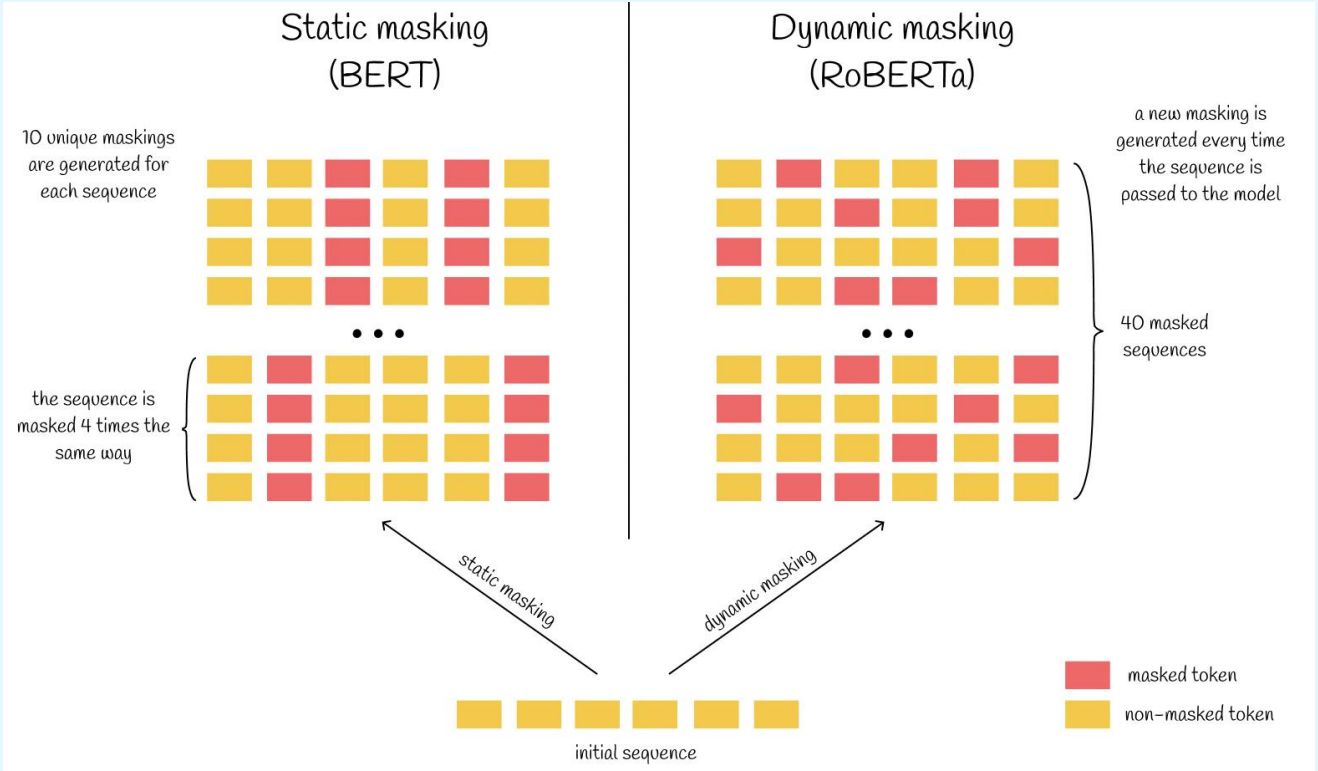
Transformer architecture with layers L, self-attention heads A and hidden dimension H

BERT – Pretraining & Optimization

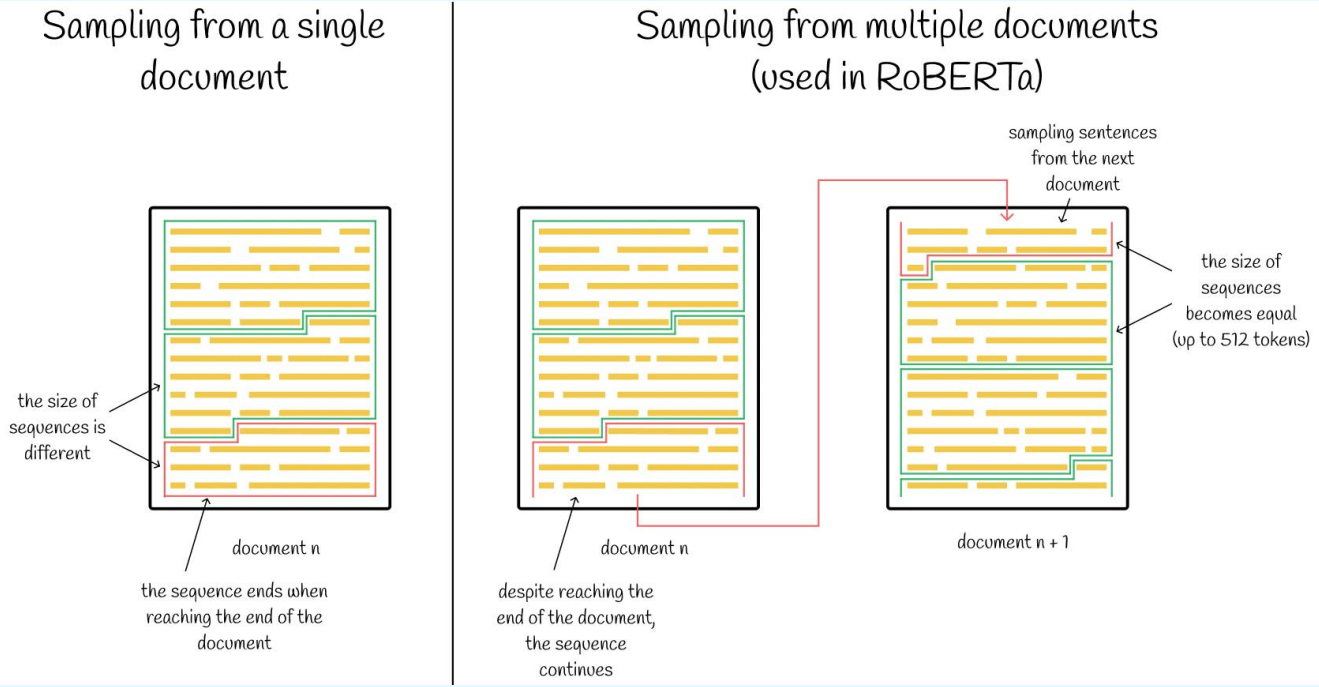
Two objectives:
- Masked Language Model (MLM)
- Next Sentence Prediciton (NSP)

Adam for optimization
Trained with BookCorpus + English Wikipedia

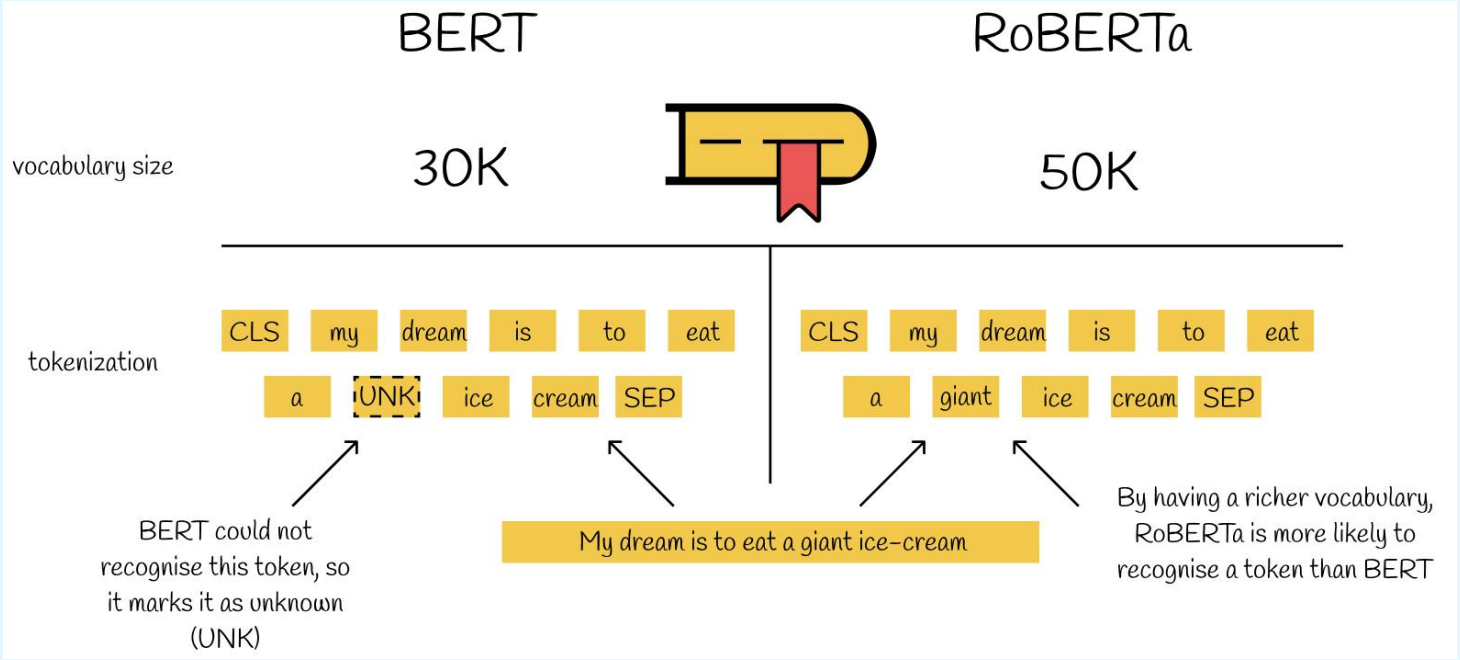
RoBERTa improvements – Masking



RoBERTa improvements – Training data



RoBERTa improvements – Encoding



RoBERTa improvements – Overview

- Dynamic masking
- Next sentence prediction
- Byte-pair encoding
- Larger batch size

Evaluation

Benchmarks

- **GLUE**
- SQuAD
- RACE

General Language Understanding Evaluation

Contents	independent sentences and sentence pairs
Task	Answering questions, assigning correct context to ambiguous words …
Challenge	Number of tasks

Evaluation

Benchmarks

- GLUE
- **SQuAD**
- RACE

Stanford Question Answering Dataset

Contents	paragraph of context and a question
Task	Answer the question
Challenge	It contains unanswerable questions

Evaluation

Benchmarks

- GLUE
- SQuAD
- **RACE**

ReAding Comprehension from Examinations

Contents	large dataset with 28,000 passages and nearly 100,000 questions
Task	select one correct answer from four options
Challenge	significantly longer context

Evaluation

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

Evaluation

Model Input Format and Next Sentence Prediction

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

Evaluation

Training with large batches

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

Evaluation

Text Encoding

Differences to BERT

	BERT	RoBERTa
Masking	static masking	dynamic masking
Data training format	SEGMENT-PAIR	FULL-SENTENCES
Batch size (Pre training)	256	2048
vocab size	30K	50K
Data amount (Pre training)	3.3 billion word corpus	30.3 billion word corpus
Epochs (Pre training)	40	20
Steps (Pre training)	1,000,000	500,000

BERT – RoBERTa comparison

	GLUE	SQuAD (v1.1/2.0)	RACE
BERT	84.05	90.9/81.8	72.0
RoBERTa	88.5	94.6/89.4	83.2

Sources

Liu, Zhuang, et al. "A robustly optimized BERT pre-training approach with post-training." China National Conference on Chinese Computational Linguistics. Cham: Springer International Publishing, 2021.

Vyacheslav Efimov. "Large Language Models: RoBERTa — A Robustly Optimized BERT Approach" Toward Data Science, 25 September 2023, https://towardsdatascience.com/roberta-1ef07226c8d8. Accessed 9 January 2024.

SQuAD

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

	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

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			$\mathbf{87.0}^{\dagger}$	$\mathbf{89.9}^{\dagger}$

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**