# RoBERTa: A Robustly Optimized BERT Pretraining Approach

## Stated objective

The paper preforms a replication study of the original BERT model. They analyse the effect of hyperparamters and training set size on the performance of the model. The results indicate that training BERT for *longer, with more data, longer sequences, random token masking, and removing the Next Sentence Prediction objective*, yield better results.

## BERT shortcomigs

In the original BERT paper, the masks for tokens are computed as a pre-processing step and then they are fixed for the duration of the training — *static masking.* In RoBERTa, training data is duplicated 10 times, each sequence is masked 10 different ways over 40 epochs — *dynamic masking.*

In RoBERTa, the authors analyze the necessity of the Next Sentence Prediction (NSP) task on the final models performance. For NSP, the input to BERT, is contiguous segments (one or more sentences) from a document under the constraint that the total tokens $< 512$. The authors try different variants of this, namely

1. SENTENCE-PAIR + NSP: Only a **single** sentence is used from a document. Batch size is increased to account for the small token size.

2. FULL-SENTENCES: Full sentences *without NSP* are sampled from one or more documents, such that length of tokens $< 512$.

3. DOC-SENTENCES: Full Sentences *without NSP*, sampled from **one** document. Batch size proportionally increased.

The results show that training on sentences from a single document — DOC-SENTENCES, without the NSP loss has the best overall performance.

## Training Procedure

Full length sequences, with max size $T = 512$. *Adam* optimizer with $\beta_2 = 0.98$ for larger batch sizes.
**Datasets**: BOOKCORPUS, CC-NEWS, OPENWEBTEXT, STORIES, totalling 160GB of raw uncompressed data.

The authors experimented with larger batch sizes, in accordance with past work. The BERT implementation used a $batch\_size = 256$ which had a perplexity of 3.99. Increasing the size to $batch\_size = 2k$, improved the perplexity to 3.68.

RoBERTa uses a Byte-Pair Encoding (BPE), compared to BERT's character-level encoding, without significantly improved results.

Finally, the authors pre-train RoBERTa on a much larger corupus $160GB$ over BERT's $16GB$. The results also indicate that training for longer, $100k \rightarrow 300k \rightarrow 500k$ steps, give better results.

# Language Models are Unsupervised Multitask Learners

## Stated Objective

Language models for translation, question answering, etc are trained in a supervised methods. In this study the authors demonstrate that models that are trained in an unsupervised fashion on millions of documents learn this implicitly. This shows the success of language models in zero-shot task transfer. The authors use a GPT-2 model trained on WebText dataset

Earlier models used a combination of large datasets, higher capacity models and supervised learning for a task. However, these models are prone to shifts in the data distribution. The prevalence of single task training on single domain datasets is a reason for lack of generalisation in these systems. In this paper, the authors search for more general methdos of transfer learning. They show that language models can perform well on down-stream task in zero-shot setting without any parameter or architecture modification.

## Dataset

Earlier work on language models used single domain text. The authors use a large and diverse natural language dataset. For this, they use a web scrape, which is further manully filtered for quality, known as WebText. It is approximately 40GB in size and contains 8 million documents.

## Approach

Language modeling can be frame as a product of conditional probabilities. Earlier task-specific architectures were used. However, the authors see that most language tasks can be framed as a sequence of symbols. For example, *machine translation* can be thought of as `(translate to french, english-text, french-text)`. The input is encoded using Byte Pair Encoding (BPE). The authors use a Transformer architecture, based on the OpenAI GPT model. Next they introduce GPT-2, which has many more parameters.

## Experiments

The authors test the GPT-2 model on Reading Comprehenstion, Summmarization, Translation, Children's book test, Winograd Schema and Question Answering tasks. Further, the authors analyse the duplication across the downstream datasets and WebText by counting the 8-grams using a bloom filter. This is done so that the performance of the model can be judged correctly. They found a 1.6% overlap.

## Analysis

In reading comprehension task, the GPT-2 models performance in zero-shot setting is comparable to the performance of other models trained under supervision. In summarization, its peformance still lacks according to qualitative metrics. While in many other tasks, the model peforms no better than random guessing. GPT-2 zero-shots to 7 of 8 datasets. This suggests that the high capacity of model is trained to maximize the likelihood.

## Electra: Pre-training text encoders as discriminators rather than generators

### Stated Objective

Masked Language Modeling (MLM) such as BERT requiqe large amounts of compute to be effective. In this paper, the authors propose a more sample-efficient pre-training task called *replaced token detection*. Instead of masking the input, their approach corrupts the input by replacing some tokens with plausible alternatives sampled from a generator network.

### Introduction

MLM models are can thought of as learning denoising autoencoders. However, they only learn from 15% of the tokens per example. This requires a lot of compute. Instead the authors propse *replaced token detection*. The tokens are replaced with output of another language model and the model is trained as a discriminator rather than a generator. The generator is usually trained jointly with the discriminator. Similary to GAN, however the generator is trained with maximum likelihood.

### Experiments

The authors evaluate the model on GLUE benchmarks. The model hyperparamters and architecture are similar to BERT. The weights between the generator and discriminator network are shared, which performs better when not shared. The authors also experiment with different model sizes. The Electra model requires significantly less compute, while providing comparable results. There is an order of magnitude speedup compared to BERT-base model.