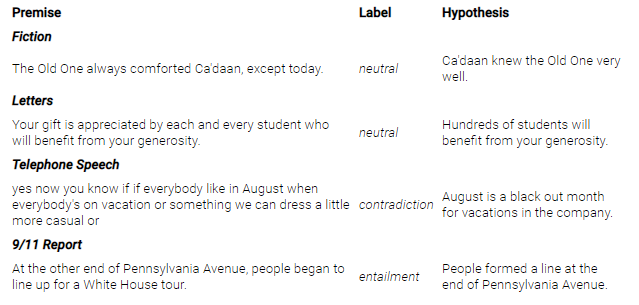
PRESENTATION

* A **robust** model is the one that offers a testing error closer to the training error.
* We will see in **which tasks** ROBERTA had a state-of-the-art performance so we can have an idea of what the model can be useful for
  + Multi Genre Natural Language Inference: premise – hypothesis – label
  + Question-answering Natural Language Inference: determine if the paragraph contains the answer to the question
  + Recognizing Textual Entailment: recognize if the meaning of one text is entailed from the other
  + Semantic Textual Similarity: classify pairs of sentences by their similarity, from 0 to 5
  + ReAding Comprehension dataset from Examinations: reading questions from English exams. It’s for testing human reading skills
* DOC-SENTENCES performs better than FULL-SENTENCES, but the first creates different batch sizes, so FULL-SENTENCES is used for better comparison with other work.
* PPL = perplexity, the lower the better

BERT (Bidirectional Encoder Representations from Transformers) is a self-supervised method release by Google in 2018. It relies on unannotated text obtained from the web, as opposed to a language corpus that’s been labeled specifically for a given task.

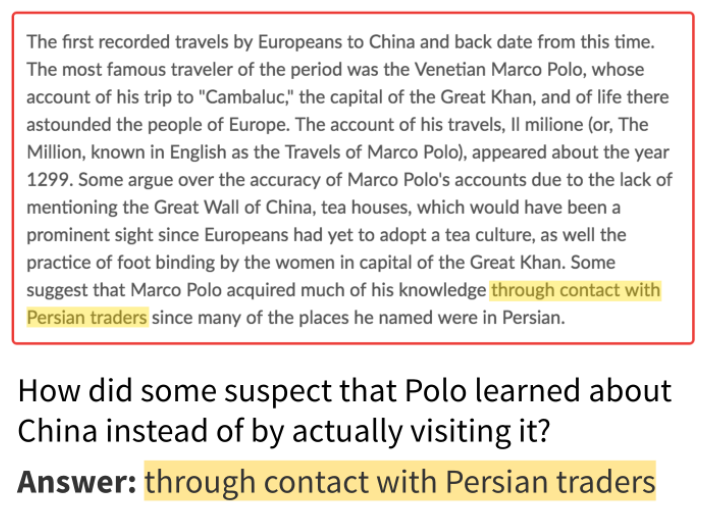
Roberta:

* follows the same strategy as BERT’s language masking, where the system learns to **predict intentionally hidden sections of text** within unlabeled language examples.
* Modifies key hyperparameters in BERT as:
  + Removing BERT’s next-sentence pretraining objective
  + Training with much larger mini-batches and learning rates
* Is trained on a larger dataset than BERT and for a larger amount of time. The dataset is conformed by existing unannotated NLP datasets as well as CC-News, a novel set drawn from public news articles.
* Roberta delivered state-of-the-art performance on the
  + MNLI (Multi-Genre Natural Language Inference corpus is a crowd-sourced collection of 433k sentence pairs annotated with textual entailment information). This data is from different sources: Face-to-face, Telephone, 9/11, Travel, Letters, Oxford University Press, Slate, Verbatim, Government and Fiction



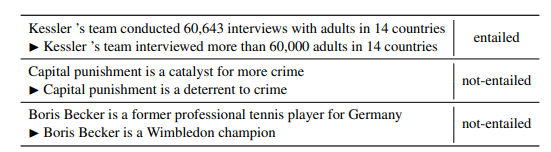


* + QNLI (Question-answering NLI), which is a dataset with question-paragraphs pairs, The task consists in determining if the paragraph contains the answer to the question, with the text being from Wikipedia and the questions being written by an annotator.



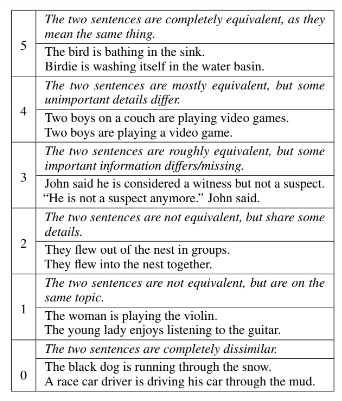


* RTE (Recognizing Textual Entailment) dataset contains pairs of text fragments, with the task requiring go recognize whether the meaning of one text is entailed from the other. Data is from news and Wikipedia text.



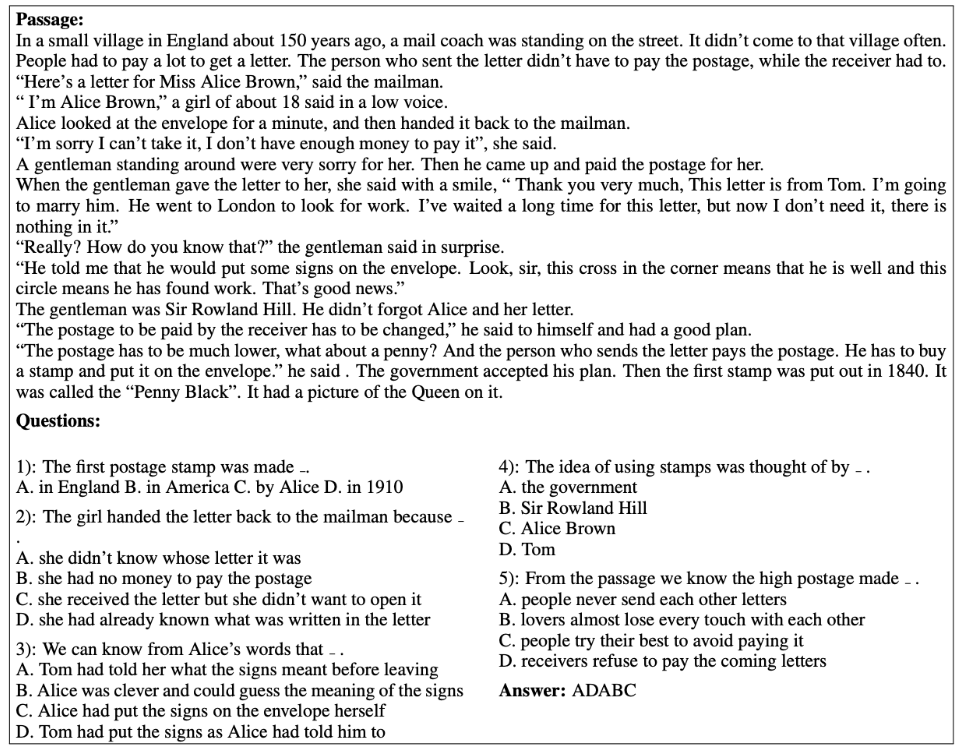


* STS Benchmark (Semantic Textual Similarity), which comprises a selection of datasets formed by image captions, news headlines and user forums.





* RACE (ReAding Comprehension dataset from Examinations) consists of 27,933 passages and 97,867 questions from English exams, targeting Chinese students aged 12-18. Each question has only one correct answer. RACE differs from most machine reading comprehension datasets – instead of generating questions and answers by heuristics or crowd-sourcing, questions are specifically designed for testing human reading skills.

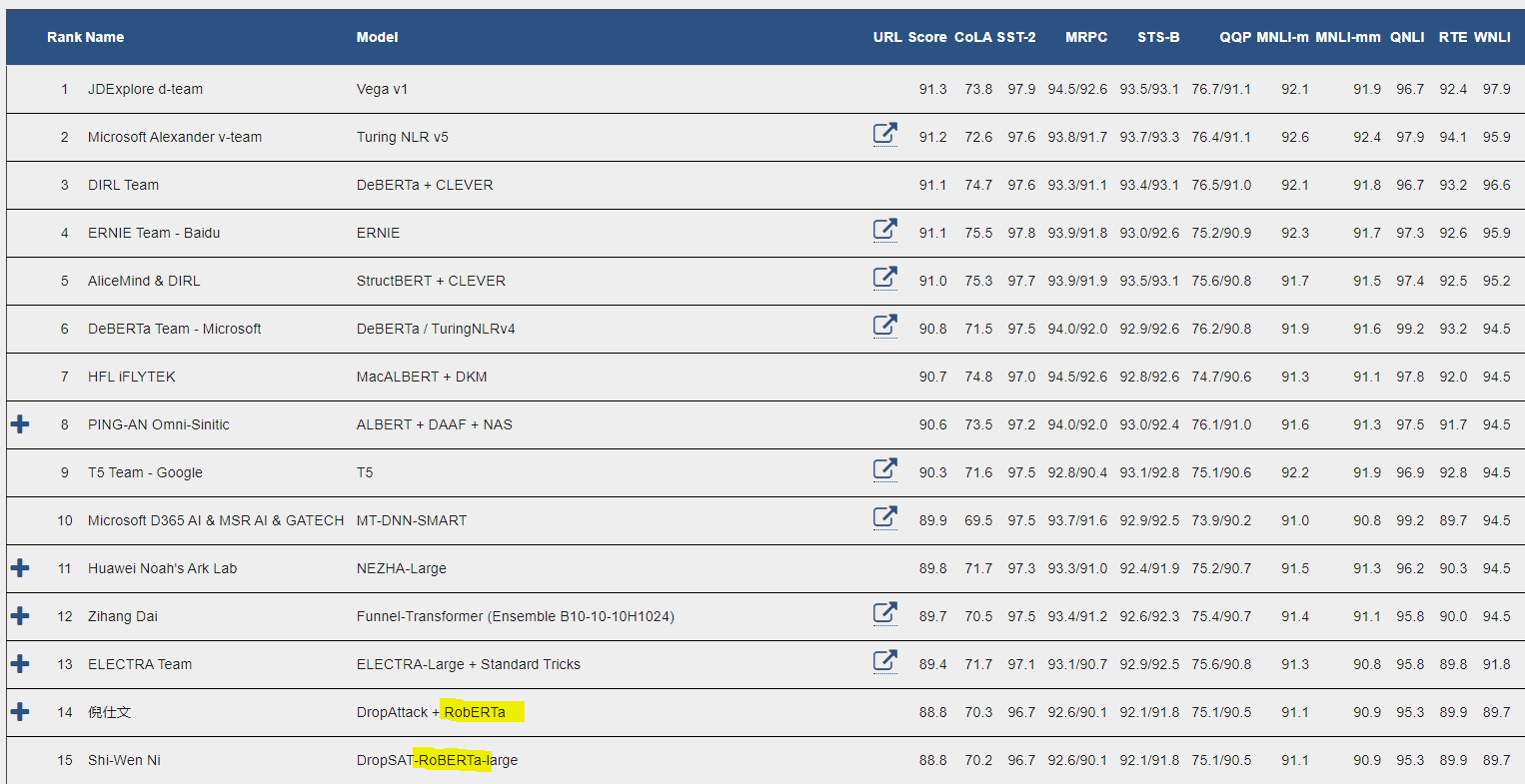




At the time RoBERTa was released in 2019, it reached the top position of the GLUE leaderboard.

Glue (General Language Understanding Evaluation) is a collection of nine natural language understanding tasks that aim to evaluate natural language understanding models.

At the actual moment Roberta is on the 19th position, even though some models that use it combined with other systems occupy higher positions.



https://datascience.stackexchange.com/questions/97310/what-is-the-difference-between-bert-and-roberta

<https://www.youtube.com/watch?v=-MCYbmU9kfg>

BERT

Bert is an **encoder**. Training of bert is done in two phases:

1. **Pretrain** BERT to understand language

What is language? What is context? Bert learns language by training in two tasks simultaneously:

* 1. **Masked Language Model**: mask one word of the sentence
  2. **Next Sentence Prediction**: does sentence B follow sentence A

Both happen at the same time with tokenized sentences

Output **Word vectors** all have the same size and they are all generated simultaneously.

Apply a softmax activation to the output word vector to create a distribution. This distribution is a one-hot encoding vector. Compare and train the network with softmax layer vs actual word (one hot encoded)

1. **Fine tune** BERT to a specific task
   1. **Fine tuning**: how to use language for specific task?

Question-Answering: modify input and output. Pass in a question and a passage containing the answer. In the output layer we would output the start and end words.

video:

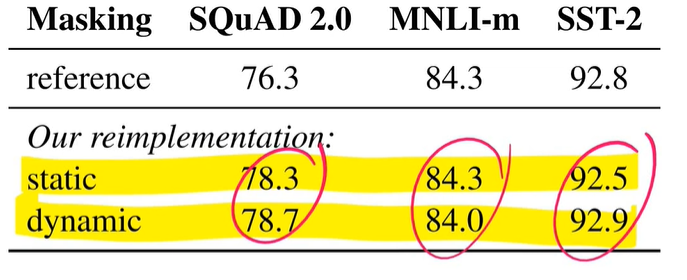
The creators of ROBERTA explored the necessity of Next Sentence Prediction binary classification **loss**

**Static vs Dynamic masking**

**Masking**: mask words from a sequence of input

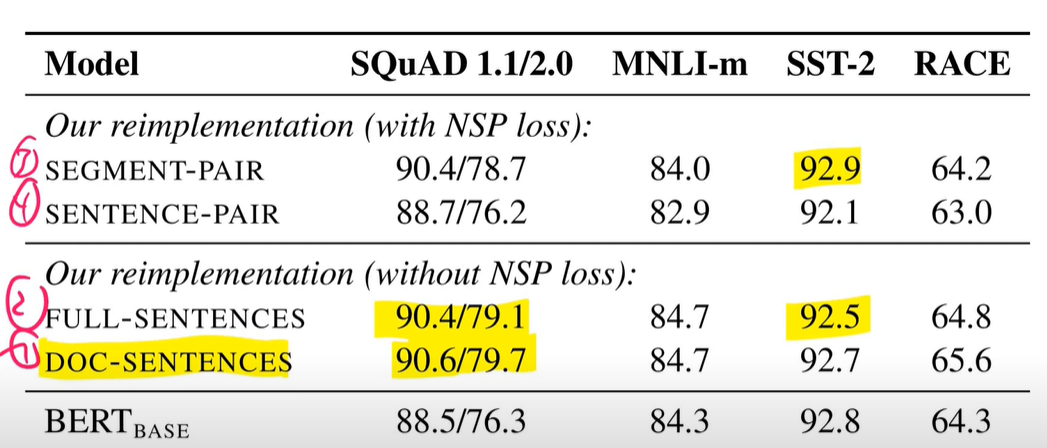
Bert performs masking **once** during data preprocessing = **static mask.** Masks 15% of the words and learns the bidirectional representation of sentences.

Roberta generates the mask on the flight, without precomputing it. It uses a different mask for each training instance in every epoch. dynamic masking so that masked token changes during each epoch which made the model to learn to predict intentionally hidden secrets of text.



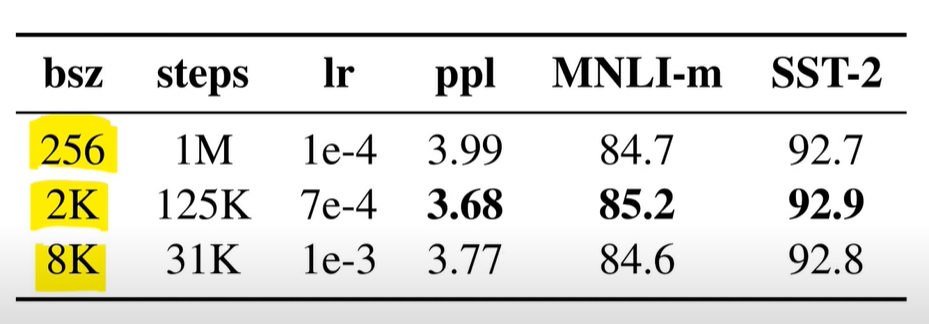
BERT uses Next Sentence Prediction loss on the training of the model. They compared alternative training models:

* Segment-pair+NSP: each input has a pair of segments which can each contain multiple sentences.
* Sentence-pair+NSP: each input is a pair of sentences
* Full-sentences: each input is packed with full sentences sampled contiguously from one or more docs
* Doc-sentences: input are constructed similarly to full-sentences, except that they may not cross document boundaries.

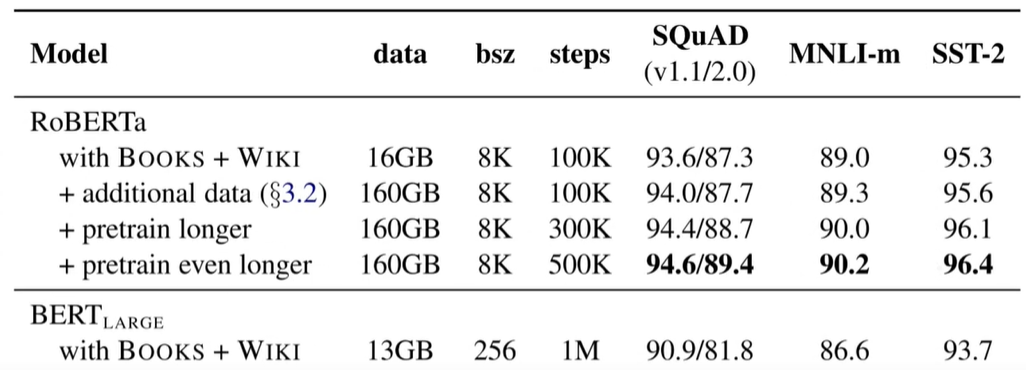


**Removing the NSP loss matches or slightly improves downstream task performance**.

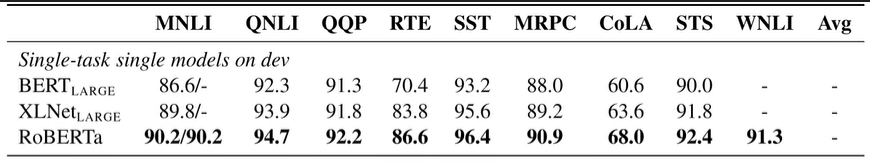
Next thing they studied was **batch size**. Batch size is the number of inputs used to train a single forward and backward pass. Improvement with larger batch sizes. Optimal 2k:



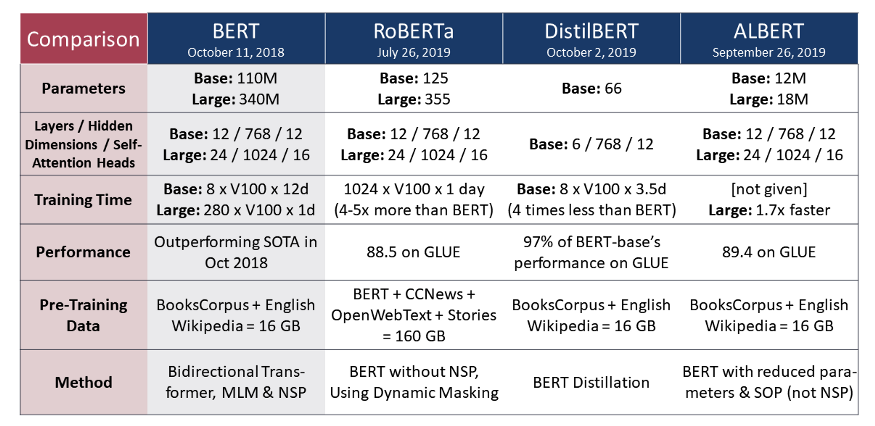
**Roberta**: dynamic masking, FULL-SENTENCES without NSP loss, large mini-batches and a larger byte-level BPE.



GLUE:



Conclusions:



<https://medium.com/analytics-vidhya/6-steps-to-build-roberta-a-robustly-optimised-bert-pretraining-approach-e508ebe78b96>

example python:

https://towardsdatascience.com/transformers-retraining-roberta-base-using-the-roberta-mlm-procedure-7422160d5764

