Replication Progress

Should You Mask 15% in Masked Language Modeling?

Team 7

Katharina Strauss Soegaard
Aaron Lingchi Chao
Timo Stephan Hermann Stoll
Max Huppertz

Problem

- Masking rate - 15%
- Masking strategy

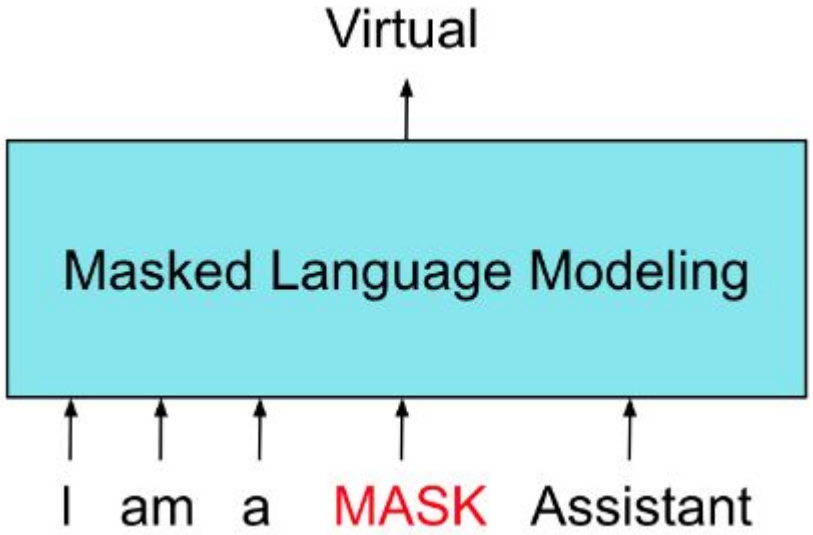


Fig. https://miro.medium.com/v2/resize:fit:880/1*0vMq-b5y1-XoA4TGJ3Akew.png

Problem

“15% masking is used ubiquitously by BERT’s successors”

BUT

“15% is not universally optimal“

➡ Which impact do masking rate and strategy have?

https://aclanthology.org/2023.eacl-main.217.pdf

Problem

Masking rate and strategy effect:

- Performance on downstream tasks (GLUE)
- Sample Efficiency
- Training Time

➡ Masking rate and strategy choice is important!

Problem

In the paper they:

- performed pre-training with 15%, 40% and 80% masking rate
- analyzed and compared downstream task performance (GLUE, SqUAD)

Findings:

- larger models benefited from higher masking rates
- even 80% masking rate performs well

Approach

Fine-tuning on GLUE and on SqUAD v1.1 given the instructions in the paper

Pretrained models are available for different masking rates

Writing a training script for the usage of Huggingface checkpoints supplied by the authors

Execution of the script on 6/9 GLUE tasks and evaluation using the given metrics

Dataset

Replication:

- SQuAD v1.1, reading comprehension tasks
- GLUE, tasks test a variety of NLP areas, including acceptability, sentiment, similarity and inference tasks

Replication results (comparison to paper)

- Using pretrained models and fine-tuning them on GLUE and SQuAD
- We ran into issues with the given scripts, therefore we rewrote them
- Bug fixing and rewriting the scripts took a lot of time but resulted in a working script for GLUE tasks
- The script for SQuAD requires manual transformation of given fairseq checkpoints where bugs still occur

Replication results (comparison to paper)

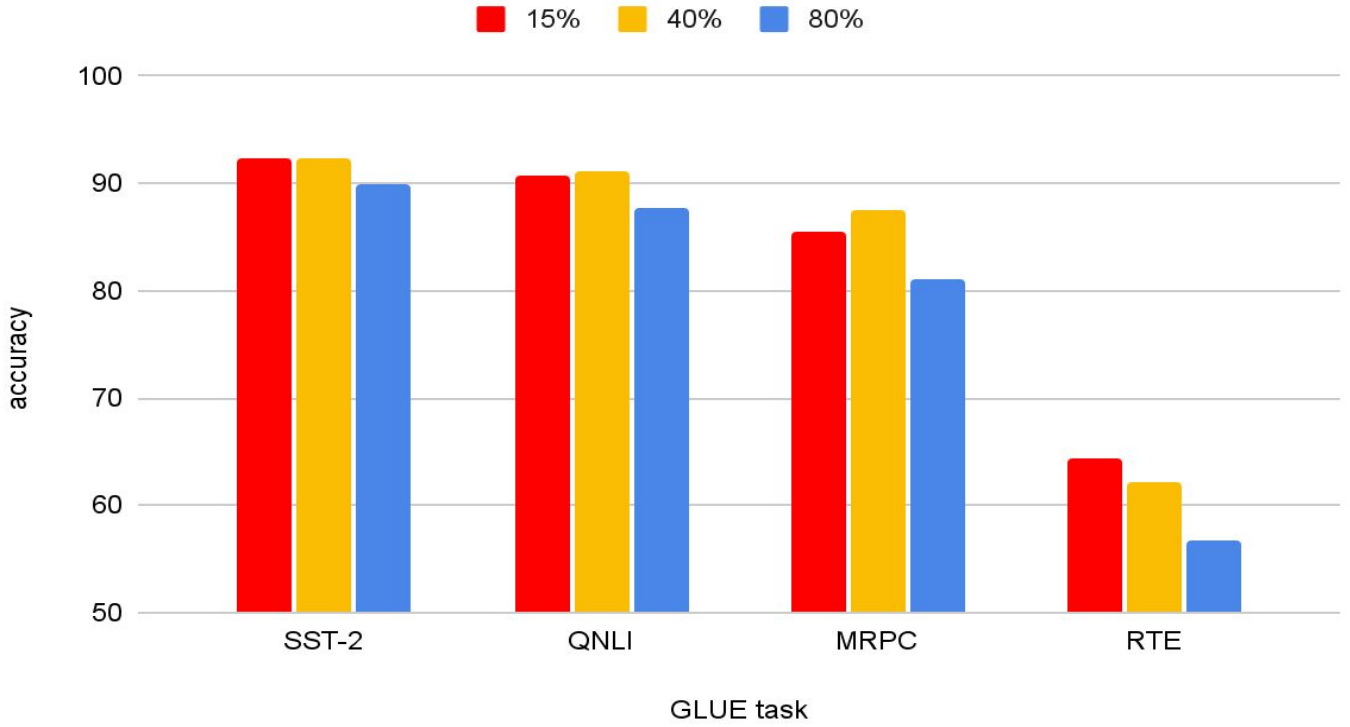
- Experiment setting
  - Large RoBERTa structure with 354M parameters
  - Using a custom script to automate the training process using the Huggingface transformers library
- We managed to replicate 6/10 fine-tuning tasks, and we plan to run the rest during this week.
- Executing the remaining GLUE tasks just requires additional time but not any further code changes

Replication results (comparison to paper)

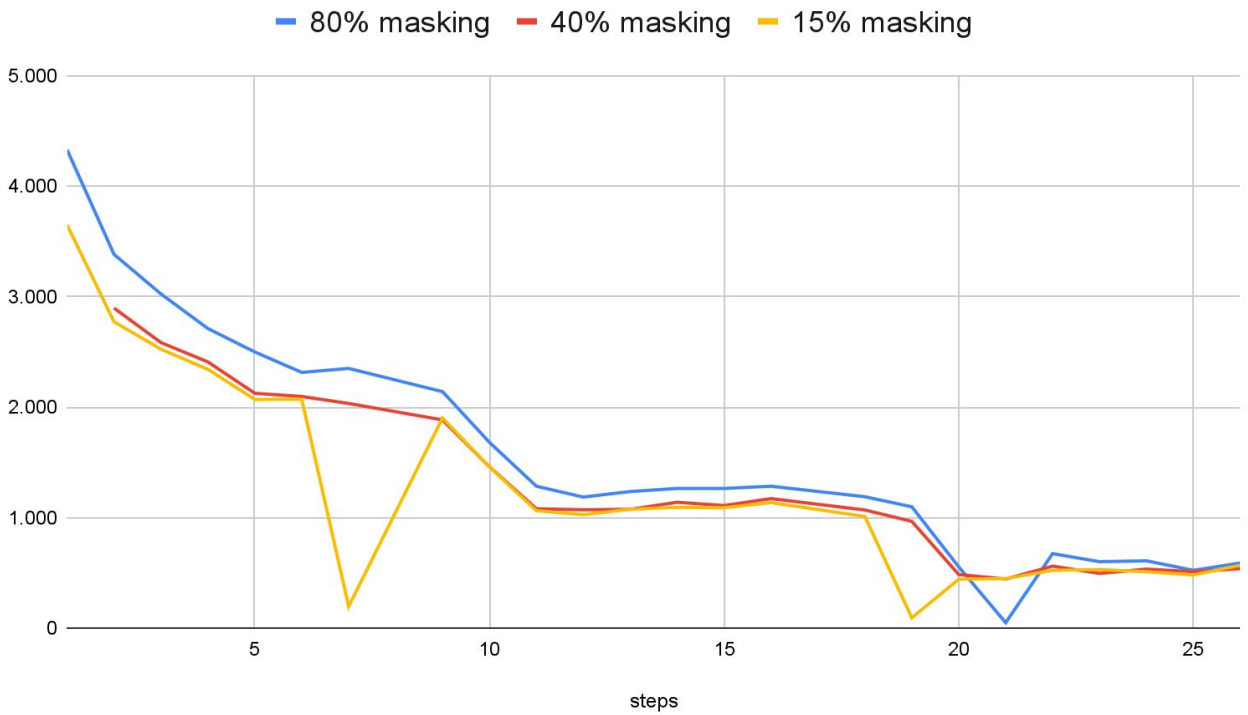
	MNLI-m	MNLI-mm	QNLI	QQP	RTE	SST-2	MRPC	CoLA	STS-B	SQuAD
15%	/84.2	/84.6	**90.8**/ 90.9	/87.8	**64.3**/ 67.3	**92.4**/ 93.3	**85.5**/ 77.0	**57.8**/ 59.2	**86.7**/87.7	/88.0
40%	/84.5	/84.8	**91.2**/ 91.6	/88.1	**62.1**/ 67.0	**92.3**/ 92.8	**87.5**/ 76.9	**59.3**/ 61.0	**86.7**/88.2	/89.8
80%	/80.8	/81.0	**87.8**/ 87.9	/87.1	**56.7**/ 58.6	**89.9**/ 90.5	**81.1**/ 72.1	**37.7**/ 38.7	**85.5**/86.3	/86.2

Our results are highlighted in blue, for MNLI-m/MNLI-mm, QQP and SQuAD the results are not available yet.

Comparison of replication results



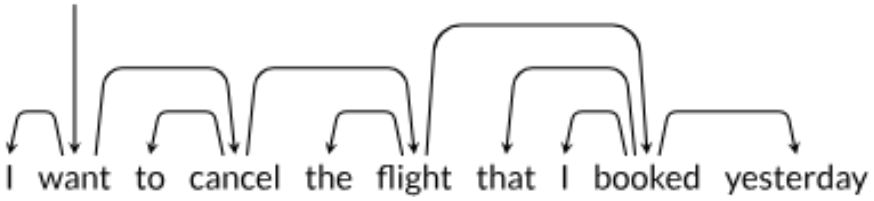
SST-2 - training loss

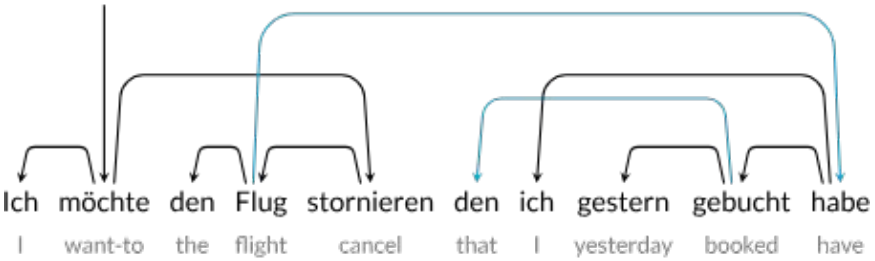


German

Language Specific

- Performance → GottBERT >= Multi-Modal Language Models
- Different Grammar and Longer Words
- Non-Projective Syntax - *syntactic structures are not drawable without crossing 2 arcs*
- Less Data Compared to English





https://explosion.ai/blog/german-model

Plan

Replication Outcome

- Similar Results to Paper

Approach for Future Work

- Use scripts given, and the new one we created
- Hope to train on same size, but flexible to down-size the model
- Doing fixed masking rate first then alternating masking rate if possible

Old Timeline

Week #	Tasks
9/10	Run Original Code and Replicate Results
11	**Upload Replication Presentation Video**
12	Use New German Language Dataset and Try Out Changes in Masking Strategies and Specific Corpuses
13/14	Write Report Using Results of Changes
15	**Final Presentation** (12/5-12/7)
16	**Final Report** (12/14)

Division of Labor

Programming Team

(*Timo and Max*)

1. Replication Code
2. Improvement Code

Technical Writing Team

(*Katharina and Aaron*)

1. Replication Presentation
2. Final Presentation
3. Final Report

New Timeline

Week #	Tasks
12	Figure out the Improvement Implementation (New Dataset) - Dataset Pre-Processing - Training & Fine-tuning - Results
13/14	Get Results from Changed Implementation and Write Report - Analysis using Given Metrics - Discussion of the Results - Theoretical View on Masking Rates
15	**Final Presentation** (12/5-12/7)
16	**Final Report** (12/14)