An Amendment to BERTology: How Much Does Bert Models Reply on Numbers?

Zixiang Pei
New York University
zp2123@nyu.edu

Zhehan Shi
New York University
zs1113@nyu.edu

Hang Yang
New York University
hy2423@nyu.edu

Abstract

For this project we proposed a brand new method to study Bert models' ability to utilize numeracy in several tasks, namely, classification and numeric-related question answering. RoBerta is a variant of the Bert model that was developed by Facebook AI. We compare roBerta model's performance on the original dataset and on a customized dataset where all numbers are masked with a special <NUM> token. From the results we obtained, we observed that the model performs better when finetuned on the masked dataset and tested on the original datasts. We conclude that masking numbers help Bert models understand and utilize numeracy.

1 Introduction

Transformers have an outsize impact in the field of Natural Language Processing since 2018, (Devlin et al., 2018). While the family of BERT models has achieved state-of-the-art performance on a number of tasks, it is less clear why, which hinders further improvements of the architecture. For example, BERT models struggle with representation of numbers. While there has been past experiments that measure Bert models' abilities on probing tasks, few has targeted on reasoning-related tasks. Our team proposes a novel experiment on how (and whether) Bert models understand numeracy and use it on reasoning tasks. For example, when asking a Bert model to answer "who won the basketball game?", does the model rely on comparing numbers (Team A scored x points, Team B threw y points, $x > y$ so A is the answer) or does the model rely on some other semantic cues (Team A is keeping their lead in the last session.) to answer the question.

2 Motivation

2.1 Sociocultural Role by Numbers

We counter numbers everywhere, from scientific journals to financial documents (Thawani et al., 2021). Numbers could also serve as a source of superstition or sarcasm (Dubey et al., 2019). The ability to understand and work with numbers is critical for many complex reasoning tasks (Wallace et al., 2019). This experiment could also serve as a systematic approach to detect whether BERT models learn and benefit from numerical reasoning when performing downstream tasks.

2.2 Semantic Roles Played by Numbers

Both (Wallace et al., 2019) and (Rogers et al., 2020) discuss that Bert models struggle with representations of numbers. They conducted experiments on probing tasks such as decoding (given a number's embedding, regress its true value) or addition (given the embedding of two numbers, predict their sum). The experiments conclude that Bert struggles on floats and Bert cannot extrapolate. While those tasks tend to evaluate numeracy as is, they fail to uncover enough of numbers' semantic roles when put into a full sentence and carry out regular tasks such as sentence-level classification.

3 Methods

For this study our method is to compare model results on the original dataset and on a processed dataset where all numbers are masked using a special token <NUM> for two tasks, classification and questions answering. For classification we try to predict whether an email or a text message is spam or not. For question answering, we try to predict the exact answers to questions given a context. By comparing the performance of our model on the original and processed datasets, we can assess the

impact of masking numbers on the model's performance and draw conclusions about the effectiveness of this preprocessing technique.

3.1 Dataset Summary

3.1.1 SMS spam dataset

The SMS Spam Collection is a public set of SMS labeled messages that have been collected for mobile phone spam research. The dataset was originally donated in 2012. There are 5574 instances. 425 of them are spam.

3.1.2 Enron email spam dataset

This Enron (Klimt and Yang, 2004) dataset is a variant of of original Enron email spam. In this collection, there are 33716 e-mails total, 17171 of them are spam and 16545 non-spam.

3.1.3 DROP dataset

DROP (Dua et al., 2019) is an acronym for Discrete Reasoning Over the content of Paragraphs. DROP is a crowdsourced, adversarially-created dataset. There are originally 86945 instances. We only selected answers of both type `span` and length one, leaving us with 24655 instances.

3.2 Dataset Manipulation

3.2.1 Masking

The goal of masking our dataset is to help us understand how our model is performing under different conditions.

For the classification task, we replaced all of the numbers with the same token while keeping the rest the same, for example, "100.23" should be converted to "<num> . <num>". "M32XL" should be converted to "M <num> XL", and so on. In addition, we also mask some numbers in English words, for example, "two hundred and fifteen" is converted to "<num> <num> and <num>" token.

For the question answering, we first replicated the same approach for the classification task on the question answering task. We masked all the elements in the three columns, `context`, `question` and `answer`. But we kept the `answer_start`, which is an element of `answer`.

3.3 Dataset Conversion

We used `metrics.load('drop')` from HuggingFace on DROP dataset; therefore, we converted DROP into SQuAD (Stanford Question Answering Dataset) (Rajpurkar et al., 2016) dataset format by introducing `answer_start` element into the `answer` column. There were three types in the original DROP dataset, `date`, `number` and `span`. We only kept the answers that are of `span` type and of length one and renamed `spans` into `text`.

After the masking process, we have also updated `answer_start` if their position is changed because some of the answers are numbers. Once they are replaced with <num> token, they might appear earlier in the sentence. For example:

```
{'text': ['103-yard opening kickoff return'], 'answer_start': [69]}
{'text': [' <num> -yard opening kickoff return'], 'answer_start': [70]}
```

3.4 roBerta

RoBERTa (Liu et al., 2019) stands for A Robustly Optimized Bert Pretraining Approach. It was first proposed is 2019 and is now one of the most widely used Bert models. The major improvements from Bert are that roBerta uses dynamic masking rather than static masking for it's masked language modeling task, and it removes the next sentence prediction task. It is also trained on more data and it uses a larger batch size. We are hoping to get more trust-worthy results by utilizing roBerta for our downstream tasks.

3.5 Reproducibility

We seeded our program in all the classification tasks and ran the program multiple times on a small dataset to ensure the results are reproducbible.

3.6 Hypothesis

Since roBerta is such a powerful model, we hypothesize that training on unmasked data would lead to better performance of the model on the validation sets. However, we also hypothesize that such a model would not extrapolate when we test on masked data and vice versa since there is the risk of overfitting.

4 Results

4.1 Abbreviation Explanation

In the *Type* column, the original dataset is referred as *num* and the masked dataset *mask*. For table 3 and table 4, it is trained on one dataset and tested on another to probe for generalizability.

4.2 Classification

The following tables are the results for classification using roBerta by training and testing on both numbered dataset and masked dataset.

Type	Accuracy	F1 Score	Precision	Recall
num on num	0.9901	0.9605	**1.0000**	0.9241
num on mask	0.9300	0.6320	**1.0000**	0.4621
mask on mask	**0.9964**	**0.9860**	**1.0000**	**0.9724**
mask on num	0.9946	0.9790	0.9929	0.9655

Table 1 SMS dataset
Train/Val/Test = 0.6/0.2/0.2

Type	Accuracy	F1 Score	Precision	Recall
num on num	0.9835	0.9894	0.9938	0.9849
num on mask	**0.9902**	**0.9903**	0.9931	**0.9876**
mask on mask	0.9810	0.9877	**0.9950**	0.9805
mask on num	0.9831	0.9891	**0.9950**	0.9832

Table 2 Enron dataset
Train/Val/Test = 0.6/0.2/0.2

Type	Accuracy	F1 Score	Precision	Recall
num on num	0.5112	0.2051	0.5969	0.1238
num on mask	0.5015	0.0443	**0.9330**	0.0227
mask on mask	**0.5503**	**0.3006**	0.7227	**0.1897**
mask on num	0.5426	0.2410	0.7779	0.1426

Table 3 trained on SMS tested on Enron dataset
Train/Val/Test = 0.75 Enron/0.25 Enron/1.0 SMS

Type	Accuracy	F1 Score	Precision	Recall
num on num	0.6332	0.3960	0.2541	0.8969
num on mask	0.6648	0.4177	0.2722	0.8969
mask on mask	0.6307	0.3933	0.2522	0.8929
mask on num	**0.6732**	**0.4332**	**0.2822**	**0.9317**

Table 4 trained on Enron & tested on SMS dataset
Train/Val/Test = 0.75 Enron/0.25 Enron/1.0 SMS

4.3 Numerical-Related Reasoning

For the numerical-related reasoning tasks we follow the same approach we did for the classification tasks. In addition to the masking algorithm we introduced, we also preprocess the data to remove all the pure number probing tasks (for those questions the answer requires knowing the actual values of the numbers, which are masked in our case). We then finetuned the pretrained roBerta model for both the masked and unmasked data for 40 epochs and below we report the results as follows:

Type	Exact Matches	F1 Score
num on num	13.37	0.2086
num on mask	14.94	0.1991
mask on mask	**15.28**	**0.2318**
mask on num	14.45	0.1944

Table 5 Question answering on DROP dataset
Train/Val/Test = 0.6/0.2/0.2

5 Discussion

Overall, we discovered unexpected results and our hypotheses were refused. For most of the experiments we ran, roBerta models finetuned on masked datasets perform better on original datasets. Below we go into detail and discuss some possible reasons of this finding.

5.1 Classification

Judging by f1-score, *mask on num* outperform *num on num* in 75% of the time, three out of four tables. If we use accuracy, *mask on num* still outperform *num on num* in 75% of the time, three out of four tables. From these results, we can clearly see that the model performs better when finetuned on the *mask* dataset and tested on *num*. This indicates that assigning a consistent meaning to all numbers as a whole helps the model realize numeracy rather than having the model interpret different numbers differently. The model is thought to be forced to focus on other words when numbers are masked.

5.2 Numeric-Related Reasoning

Similar situation occurs for question ansewring task, where *mask on num* outperform *num on num* in both metrics, exact matches and f1 score. This led us to reach the relatively safe conclusion that when the roBerta model is trained on a dataset where numbers are masked, it is able to generalize better on a dataset where numbers are not masked.

One drawback of this experiment is that due to time constraint, and as question answering is a much more complicated task than sentence classification, we did not run the models to convergence. As we reported in the results section, what we did was to train both models (on masked and unmasked data) for the same amount of epochs (40) and compare how well each of them converged. If more time is allowed, we will run both models to convergence and we might draw some additional conclusions.

6 Conclusion

From the results of our study, we can see that the performance of the RoBERTa model is better when finetuned on processed dataset and test on the original dataset.

This phenomenon is generally consistent across both classification and question answering, regardless of the datasets. Numbers, whether in digits or words, often serve as a distractions rather than additional helpful information. This insight is valuable. Our findings suggest that preprocessing the training data by masking numbers can improve the performance of roBerta. In real-world scenarios where numbers are present, it is reasonable to assume that models trained on a dataset where numbers are masked will perform slightly better, providing practitioners with an added advantage in their model deployment.

Overall, our study provides valuable insights into how the roBerta model functions and the potential benefits of using preprocessing techniques to improve its performance.

7 Future Work

The experiments mentioned in the motivation section proposed some future methods to test on how well the model can extrapolate with numbers and we believe we can add those to in our potential future experiments. In the future, we will train our model on only a narrow range of numbers (i.e. from 1 to 200), and then test our model on out-of-the-range data (i.e. from 200 to infinity). From this test we can learn whether Bert models can generalize away from the training data for downstream tasks.

Another process we could do better is our treatment on punctuation. For keeping the structure of our dataset as much as possible we keep the punctuation when we tokenize and mask numbers. For instance, as we mentioned in our masking method, we convert "100.23" to "<num> . <num>" while in reality, it might be better to discard the "." punctuation. If more time allowed, we will look into how much does punctuation help/distract Bert models.

Finally, we have changed a little bit on the way some words are tokenized. For example, when "M32XL" was converted to "M <num> XL", what used to be one token now becomes 3 tokens. In the future, we can also find ways to measure the effect of such operations if we were to carry out the experiments again.

8 Author Contribution Statement

Zixiang Pei: Early research and proposal. Assisting Zhehan for finetuning roBerta for the classification task. Main reports write up.

Zhehan Shi: Main coding for the two experimental running and reporting the results.

Hang Yang: Main coding for the mask algorithm. Assisting Zhehan for finetuning roBerta for the numeric reasoning task.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

Abhijeet Dubey, Lakshya Kumar, Arpan Somani, Aditya Joshi, and Pushpak Bhattacharyya. 2019. "when numbers matter!": Detecting sarcasm in numerical portions of text. In *Proceedings of the tenth workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 72–80.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *European conference on machine learning*, pages 217–226. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Avijit Thawani, Jay Pujara, Pedro A Szekely, and Filip Ilievski. 2021. Representing numbers in nlp: a survey and a vision. *arXiv preprint arXiv:2103.13136*.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do nlp models know numbers? probing numeracy in embeddings. *arXiv preprint arXiv:1909.07940*.