View Author Feedback

Paper ID
8418

Paper Title
On the Softmax Bottleneck of Recurrent Language Models

Track Name
AAAI2021

AUTHOR FEEDBACK QUESTIONS

1. Author Response
We genuinely thank all the reviewers for their invaluable comments and suggestions.

To Reviewer #1:

Our choice of datasets is the standard used in the literature for softmax bottleneck. However, we will make an effort to include more results on additional datasets if time and our computing resources permit.

We wish to clarify our claim that high rank is neither sufficient nor necessary for good performance. By this statement, we mean that high rank models may not give good performance and that low rank models may not give poor performance, thus justifying our claim of “insufficient” and “unnecessary”.

The fundamental hypothesis justifying the limitation imposed by the softmax bottleneck is that a language model needs to have a high-rank logP matrix in order to model the rich context dependency in natural languages. This work may appear to have raised doubts on this hypothesis. We however like to note that the evidence presented in this paper also raises the possibility that the hypothesis is indeed correct but the estimation of rank is not sufficiently accurate (see our discussion on effective ranks in the paper and in later this document). Further study is necessary to fully address this.

Regarding your complaint of Figure 7, we can also include a log-scale plot in the revision. Note that even when plotted in the log-scale, it can be seen that the singular values may not drop suddenly at a number near the Press rank. Our purpose of plotting the singular values in linear scale and introducing effective ranks in equation (6) is to suggest that the singular values drop too rapidly in all compared models and that the Press rank might be an overestimate of the true rank. In particular, for example, using 0.001-effective rank, Table 6 shows that if “rank” is defined as the number of largest singular values that account for 99.9% of the total “power” (sum of squares) of all singular values, the rank of all models are no higher than 100. This observation should give a negative answer to your question regarding the spread of singular values: they do not spread over a large range. This observation seems to also suggest that all models have effectively low ranks (much less than the word-embedding dimension), for which the low-rank limit imposed by the “softmax bottleneck” on the logP matrix does not present itself as a real bottleneck. That is, the models still have room to improve their ranks before they hit the low-rank limit.

Regarding your question on Figure 2, it turns out within the GSS family, high-rank models appear to all have good performances, but such performances may also be achieved by low-rank models. If we go beyond the GSS family (e.g., using ReLU activation instead), a high-rank model may also have low performance, as shown in (Kanai et. al 2018).

To Reviewer #3:

We agree that further study is needed to arrive at a deeper understanding of the observations in this paper. We hope that our observations serve as a pointer to further research along that direction.

Our definition of GSS is motivated by extending the sigsoftmax to a rich family, which also includes the original softmax as a special case. Our inspiration is that since two members (sigsoftmax and softmax) of this family give distinct ranks, other members of this family may allow us to produce diverse ranks.

Following your suggestion, we will make an effort to reorganize the experimental results and make spaces to include more intuitions and discussions.

We agree that any set of 6 contexts will not provide a reliable conclusion. In the revision, we will make an effort to run a large scale experiment using many random sets of contexts. We will also include more details on our experimental procedure in selecting the random and biased sets of contexts.

Regarding your question concerning our training of MoS for only 500 epochs on WT2, we note that this training task is the most time consuming. Luckily, the training has converged before 500 epochs.

To Reviewer #4:

We will follow your suggestion and make our best effort revising the paper to improve its clarity and ease of reading. Progresses in reasoning and intermediate summaries will be given on each set of experiments.

Regarding your comments on word dimensions, we agree that it may not be fair to compare two models with different word dimensions. In fact we have made an effort in using the same word dimension when comparing models, except for the MoS and MoC models. For these two models, we adopted the hyper-parameter settings reported in Yang et. al (2018), since they appear to have been tuned to give the best performances. This is addressed in lines 390-405 (Column 2) and 436-469 (Column 1). Also when comparing MoS and MoC, the same word embedding dimension is used.

Regarding your comments on the word similarity tasks, we would like to note that we think such downstream tasks may serve as another metric, in addition to perplexity, to evaluate the compared models. We will make an effort however to make the presentation more concise.

Like you, we also think the model capacity and regularization aspects play a significant role in our experimental observations. We will follow your suggestion to expand that part if space permits.

We agree that the paper will greatly benefit from some more theoretical justifications. But before we see a clear handle to an in-depth theoretical study, we hope the presented work in this paper is sufficiently interesting to inspire the efforts of the research community to further this study.

To Reviewer #5:

We agree that the results of the paper are mostly established empirically. However the generalization of sigsoftmax to the GSS family is, in our opinion, an elegant mathematical construction. On the other hand, whether a research is of an empirical nature does not mean it has low impact, as is witnessed by the many breakthroughs in the deep learning revolution.

Agreeably there are many other metrics for other NLP tasks such as text generation or machine translation. However, for word-level language models, perplexity has been adopted as the primary metric in the research literature. Nonetheless in this work, we go beyond evaluation solely based on perplexity by also using some down-stream word similarity tasks.

Regarding your comments on GPT2 and BERT, in fact we note that the notion of “softmax bottleneck”, if existing at all, may not have been well defined. It would be nonetheless interesting to explore along that direction.

Regarding your comments on t-test, we note that running each experiment 10 times is near the limit of our affordable computation. Although a larger number of repetitions is more preferred, we believe that the results and conclusions presented in this work are adequately convincing. The computational complexities of the compared models are beyond the scope of this research.