View Reviews

Paper ID
8418

Paper Title
On the Softmax Bottleneck of Recurrent Language Models

Track Name
AAAI2021

Reviewer #2

Questions

1. {Summary} Please summarize the main claims/contributions of the paper in your own words. (Do not provide any review in this box)
This paper raises a question of whether language models can achieve the better performance when they have the higher rank of log-probability distribution. Recent studies hypothesize that neural language models have limited performance due to the limitation of softmax, which is called softmax bottleneck. Softmax bottleneck is that matrices composed of log-softmax output have lower ranks than the matrices composed of true log-probability distributions. In contrast to these studies, this paper experimentally shows that the ranks and the performances are not necessarily correlated through experiments with various rank models and evaluation of word embeddings.
Since the paper reveals that the performance of language models does not correlate to the rank of models, the paper also investigates the cause of high performance of Linear Monotonic Softmax with Piecewise Linear Increasing Functions" (LMS-PLIF) and mixture of softmax (MoS), which are proposed to increase the rank.

2. {Novelty} How novel is the paper?
Paper makes non-trivial advances over past work

3. {Soundness} Is the paper technically sound?
The paper has minor technical flaws that are easily fixable

4. {Impact} How important is the paper likely to be, considering both methodological contributions and impact on application areas?
The paper will impact a moderate number of researchers

5. {Clarity} Is the paper well-organized and clearly written?
Good: paper is well organized but language can be improved

6. {Evaluation} Are claims well supported by experimental results?
Moderate: Experimental results are weak: important baselines are missing, or improvements are not significant

7. {Resources} How impactful will this work be via sharing datasets, code and/or other resources? (It may help to consult the paper's reproducibility checklist.)
Not applicable: no shared resources

8. (Reproducibility) Would the experiments in the paper be easy to reproduce? (It may help to consult the paper's reproducibility checklist.)
Good: e.g., code/data available, but some details of experimental settings are missing/unclear

9. {Reasons to Accept} Please describe the paper's key strengths.
1. This paper raises an important and interesting problem about neural language models.
Since high-rank recurrent language model is an attractive research direction, negative results about this direction have a large impact on researchers. In addition, this paper provides some insightful results to support the claim. For example, this paper shows that LMS-PLIF with randomly frozen parameters can achieve as good performance as

LMS-PLIF with learnable parameters. These results might inspire researchers to propose a new method or theory for neural language modeling.

2. The claim is supported by a thorough experiments that evaluate the correlation between ranks and performance from various points of view.
To investigate the correlation, the paper evaluates four recent methods and the proposed methods, which can control the rank of the models.

10. {Reasons to Reject} Please describe the paper's key weaknesses.
This paper only shows the problem of the existing hypothesis and does not present a new algorithm for improvements of language modeling.
Even so, I think it is important to reconsider the cause of improvements in the existing methods because most of the deep learning techniques are still black-box and based on the hypothesis that is not evaluated well.

This paper only uses two datasets for the language model. Thus, the experimental evidence supports the claim only on these datasets. To obtain more convincing results, I suggest the paper additionally uses other large datasets e.g. One Billion Words Benchmark, or WikiText-103 since the large language datasets are expected to require high capacity of models.

Results in Figure 2 indeed support the claim that correlation coefficient between ranks and performance is not high. However, Figure 2 also shows that high-rank models do not have low performance at least; the perplexities of models that have the rank of about 10000 are always smaller than 65.5 on WT2. I think the results indicate that ranks of language models can contribute to good performance even though other factors also affect the performance. Therefore, the claim that the high rank is neither necessary nor sufficient for the better performance might be overstatement.
If I have misunderstood the results, please point out any errors to me.

11. {Detailed Comments} Please provide other detailed comments and constructive feedback.
I suggest that you should clarify the hypothesis in softmax bottleneck. I think softmax bottleneck is based on the hypothesis that the true probability distribution is spread over the high rank spaces. So, if the high rank models are not required for good performance, the true probability distribution in natural language might be spread over the low lank space.

I think Figure 7 might be misleading visualization since we cannot see the small difference due to the normalization and linear-scale graph. I would like to see singular values in log-scale since singular values can be spread over the large range and quickly drop at the point corresponding to the rank in a semi-log graph.
Since sizes and vocabularies of datasets are larger than 10000, I think that it is not surprising that the lowest singular value can be lower than 1/10000*largest singular values. So, I would like to know why equation (6) is more suitable metric for rank than Press et al. (2007).

--Comments after author response--
I have read author feedback. I think the evaluation of effective ranks of eq. (6) is still not sufficient. Since it is difficult to evaluate the exact ranks of log P of the real datasets, I suggest you use some toy problems to obtain more convincing results that support your claims.

12. {QUESTIONS FOR THE AUTHORS} Please provide questions for authors to address during the author feedback period. (Please number them)
Q1 If I have misunderstood and made wrong comments in the above forms, please point out them.
Q2 In experiments of Figure 2, did you observe that high-rank models have low performance?

14. (OVERALL SCORE)
6 - Above threshold of acceptance

19. I acknowledge that I have read the author's rebuttal and made whatever changes to my review where necessary.
Agreement accepted

Reviewer #3

Questions

1. {Summary} Please summarize the main claims/contributions of the paper in your own words. (Do not provide any review in this box)
Consider language modeling, let P be a |C|x|V| matrix, representing the conditional probability P(v|c) of the next word v given the context c. Yang et al. proved that the rank of log P is no more than (d + 1) where d is the word embedding dimension; they also hypothesized that the rank of log P is positively correlated to quality of the language model. This paper, however, (experimentally) shows that higher rank of log P is neither necessary nor sufficient for better test perplexity of the language model. This paper also conducts an extensive set of experiments to investigate which factors might be the true reasons that lead to the improved performance of “Mixture of Softmaxes” approaches proposed in Yang et al. 2018.

2. {Novelty} How novel is the paper?
Paper contributes some new ideas

3. {Soundness} Is the paper technically sound?
I have not checked all details, but the paper appears to be technically sound

4. {Impact} How important is the paper likely to be, considering both methodological contributions and impact on application areas?
The paper will impact a moderate number of researchers

5. {Clarity} Is the paper well-organized and clearly written?
Good: paper is well organized but language can be improved

6. {Evaluation} Are claims well supported by experimental results?
Good: Experimental results are sufficient, though more analysis would significantly add support to the claims

7. {Resources} How impactful will this work be via sharing datasets, code and/or other resources? (It may help to consult the paper’s reproducibility checklist.)
Fair: some may find shared resources useful in future work

8. (Reproducibility) Would the experiments in the paper be easy to reproduce? (It may help to consult the paper’s reproducibility checklist.)
Good: e.g., code/data available, but some details of experimental settings are missing/unclear

9. {Reasons to Accept} Please describe the paper’s key strengths.
- The conclusions of this paper are interesting and even surprising to some extent. I think the results from this paper might potentially change people’s understanding of the softmax bottleneck proposed in Yang et al.
- The paper is well-organized.
- The experiments were carefully done.

10. {Reasons to Reject} Please describe the paper’s key weaknesses.

- This paper experimentally shows that the rank of log P is only weakly correlated to test perplexity, but it does not give us a deeper understanding of why this is the case.
- Other conclusions made in this paper are also based on a set of experiments. We still lack the understanding of what might be the key factors to improve a language model.

11. {Detailed Comments} Please provide other detailed comments and constructive feedback.

- Can you provide more high-level intuitions on why the Generalized SigSoftmax should be defined in this way, and why “when used to replace the softmax function in the baseline language model, 92 is capable of producing log P matrices with diverse ranks”?

- The experiment section is a bit overwhelmed by the experimental details. Might be helpful to defer some of the details to the appendix and provide more intuitions/discussions etc.

- Figure 6, I think any conclusions made based on 6 contexts is not reliable. Also it would be better to provide a little bit more context on how the 6 contexts were picked.

12. {QUESTIONS FOR THE AUTHORS} Please provide questions for authors to address during the author feedback period. (Please number them)
1. Line 280 “...except for the MoS model on the WT2 dataset, which was trained only for 500 epochs”. Can you explain why use 500 instead of 1000?

14. (OVERALL SCORE)
6 - Above threshold of acceptance

19. I acknowledge that I have read the author's rebuttal and made whatever changes to my review where necessary.
Agreement accepted

Reviewer #4

Questions

1. {Summary} Please summarize the main claims/contributions of the paper in your own words. (Do not provide any review in this box)
This work investigates the recent solutions proposed to deal with what is called the "softmax bottleneck" in language modeling. It is the idea that a language model is basically trying to obtain a matrix P of conditional probabilities of output words given input context - and this matrix is a product of the matrices containing the vectors representing the words (output weights) and contexts (hidden representations), plus a few operations (exponential, normalization ~ softmax). However, the dimension of the word/context representation is a bound on the rank of the resulting matrix P, hindering the representational capacity of a language model.
While several recent articles propose solutions to allow the rank of this matrix to be high, and demonstrate that their solutions can provide better perplexity results, this work attempts to demonstrate empirically that this rank is only weakly correlated with performance.
Their experiment follow two directions:
- Demonstrating that high rank does not necessarily imply better performance on perplexity - which is done by building several models with different ranks and observing their behavior. The authors also look at the correlation with performance on word similarity tasks.
- Demonstrating that several factors linked to performance in language models may have opposing effect on the rank - notably, a higher capacity of the model allows for higher rank, but regularization (with dropout), which increase performances, gives lower rank. The authors suggest that these factors may explain the recent good results obtained by solutions trying to avoid this softmax bottleneck.

2. {Novelty} How novel is the paper?
Paper makes non-trivial advances over past work

3. {Soundness} Is the paper technically sound?
I have not checked all details, but the paper appears to be technically sound

4. {Impact} How important is the paper likely to be, considering both methodological contributions and impact on application areas?

The paper will impact a moderate number of researchers

5. {Clarity} Is the paper well-organized and clearly written?
Good: paper is well organized but language can be improved

6. {Evaluation} Are claims well supported by experimental results?
Good: Experimental results are sufficient, though more analysis would significantly add support to the claims

7. {Resources} How impactful will this work be via sharing datasets, code and/or other resources? (It may help to consult the paper's reproducibility checklist.)
Not applicable: no shared resources

8. (Reproducibility) Would the experiments in the paper be easy to reproduce? (It may help to consult the paper's reproducibility checklist.)
Meets Minimum Standard: e.g., code/data unavailable, but paper is clear enough that an expert could confidently reproduce

9. {Reasons to Accept} Please describe the paper's key strengths.
- A recent observation (the rank of the matrix P characterizing a language model being low because of the dimension of the word vectors being low) and and associated hypothesis (low rank limit representational power) led to solutions being proposed - models with high rank P - which in turn led to better performances.
This paper provides further analysis and attempts to rectify the (maybe rash) conclusion that low rank was indeed an issue and high rank leads to better performance. This kind of work seems to be rare and should be appreciated.

10. {Reasons to Reject} Please describe the paper's key weaknesses.
- As this paper is focused on experiments to demonstrate that previous reasoning may be false, the experimental setup needs to be particularly clear and rigorous, which is not always the case.
- The paper, while well-structured and correctly written, is often confusing.

11. {Detailed Comments} Please provide other detailed comments and constructive feedback.
On clarity:
- The introduction could give a clearer overall idea of the reasoning behind the experiments, as now, the reader is lost in the numerous details of the various models and the successions of experimental setups.
- As the paper progresses, it is difficult to follow the thread, as the experiments follow each other and intermediate summaries of the progress made on the reasoning, or of contribution that have been made, are very rarely done.

On experiments: while for the most part the paper seems technically sounds, some details seem to stick out to me.
- First, the word dimensions. It seems unfair to compare several setups that don't use same dimensions on word representations, as it is done several times.
- Secondly, the word similarity tasks and cherry picked contexts. While I understand the idea of countering examples provided in the original paper, it seems that these experiments take up a lot of space while bringing almost nothing to the overall reasoning.
Personally, I found the study done on the latter part of the paper, on capacity/regularization, to be very interesting and I believe it should be expanded on to re-enforce its impact.

12. {QUESTIONS FOR THE AUTHORS} Please provide questions for authors to address during the author feedback period. (Please number them)
Finally, I believe the paper could use some theoretical justifications. For example, on the way the rank is evaluated: an approximation is made - why is that enough ?

14. (OVERALL SCORE)
6 - Above threshold of acceptance

19. I acknowledge that I have read the author's rebuttal and made whatever changes to my review where necessary.
Agreement accepted

Reviewer #5

Questions

1. {Summary} Please summarize the main claims/contributions of the paper in your own words. (Do not provide any review in this box)
This work conducts a series of empirical studies on the relationship between the perplexity and the rank of the conditional distribution matrix. They claim that the rank has a weak correlation with the performance (perplexity) of a language model.

2. {Novelty} How novel is the paper?
Paper makes non-trivial advances over past work

3. {Soundness} Is the paper technically sound?
The paper has minor technical flaws that are easily fixable

4. {Impact} How important is the paper likely to be, considering both methodological contributions and impact on application areas?
The paper will have a broad and significant impact

5. {Clarity} Is the paper well-organized and clearly written?
Excellent: paper is well organized and clearly written

6. {Evaluation} Are claims well supported by experimental results?
Moderate: Experimental results are weak: important baselines are missing, or improvements are not significant

7. {Resources} How impactful will this work be via sharing datasets, code and/or other resources? (It may help to consult the paper's reproducibility checklist.)
Good: shared resources are likely to significantly impact future work

8. (Reproducibility) Would the experiments in the paper be easy to reproduce? (It may help to consult the paper's reproducibility checklist.)
Excellent: e.g., code/data available and paper comprehensively describes experimental settings

9. {Reasons to Accept} Please describe the paper's key strengths.
They make a different conclusion from the existing methods.

10. {Reasons to Reject} Please describe the paper's key weaknesses.
Their conclusion comes from empirical studies. What's more, they only study the correlation between the rank of the conditional distribution matrix and the perplexity. For text generation and machine translation, there are still many other metrics have to be considered.

11. {Detailed Comments} Please provide other detailed comments and constructive feedback.
I am very interested that whether their conclusion is still true or not on other popular language models such as GPT2, BERT.

12. {QUESTIONS FOR THE AUTHORS} Please provide questions for authors to address during the author feedback period. (Please number them)
Why you choose the number of samples as 10 for the t-test? Are there any considerations to reduce the compuational costs? In pratice, it should be larger or equals 30.

14. (OVERALL SCORE)
6 - Above threshold of acceptance