← Go to **ACL ARR 2024 June** homepage (/group?id=aclweb.org/ACL/ARR/2024/June)

Model Balancing Helps Low-data Training and Fine-tuning

(/pdf?id=tiSCZsHs9V)

Zihang Liu (/profile?id=~Zihang_Liu4), Yuanzhe Hu (/profile?id=~Yuanzhe_Hu1), Tianyu Pang (/profile?id=~Tianyu_Pang2), Yefan Zhou (/profile?id=~Yefan_Zhou1), Pu Ren (/profile?id=~Pu_Ren1), Yaoqing Yang (/profile?id=~Yaoqing_Yang1)

14 Jun 2024 (modified: 22 Aug 2024) ACL ARR 2024 June Submission June, Senior Area Chairs, Area Chairs, Reviewers, Authors, Commitment Readers Revisions (/revisions?id=tiSCZsHs9V)

Reviewing Volunteers For Emergency Reviewing: The volunteers listed above are only willing to serve as regular reviewers.

Abstract:
Recent advances in foundation models have highlighted the need to align pre-trained models with specialized domains using small, curated datasets. This trend has made low-data training and fine-tuning especially crucial in natural language processing (NLP) and scientific machine learning (SciML) fields. To address the limitations inherent in low-data training, we draw inspiration from Heavy-Tailed Self-Regularization (HT-SR) theory, analyzing the shape of empirical spectral densities (ESDs), and reveal a trend of imbalanced training quality across different layers of the model. We adapt a recently proposed layer-wise learning rate scheduler, TempBalance, to effectively balance layers' training qualities, and improve low-data training and fine-tuning in NLP and SciML tasks. Notably, TempBalance yields increasing performance gains as the amount of tuning data decreases. Comparative analyses further reveal the effectiveness of TempBalance and its adaptability to act as an “add-on” method to further improve model performance.

Paper Type: Long
Research Area: Efficient/Low-Resource Methods for NLP
Research Area Keywords: data-efficient training
Contribution Types: Model analysis & interpretability, Approaches to low-resource settings
Languages Studied: English
Reassignment Request Action Editor: This is not a resubmission
Reassignment Request Reviewers: This is not a resubmission
Software: zip (/attachment?id=tiSCZsHs9V&name=software)
Author Submission Checklist: I confirm that the paper is anonymous and that all links to data/code repositories in the paper are anonymous., I confirm that the paper has proper length (Short papers: 4 content pages maximum, Long papers: 8 content pages maximum, Ethical considerations and Limitations do not count toward this limit), I confirm that the paper is properly formatted (Templates for *ACL conferences can be found here: https://github.com/acl-org/acl-style-files (https://github.com/acl-org/acl-style-files).)
A1 Limitations Section: This paper has a limitations section.
A2 Potential Risks: Yes
A2 Elaboration: Appendix A
A3 Abstract And Introduction Summarize Claims: Yes
A3 Elaboration: Section 1
B Use Or Create Scientific Artifacts: Yes
B1 Cite Creators Of Artifacts: Yes
B1 Elaboration: Section 4
B2 Discuss The License For Artifacts: Yes
B2 Elaboration: Section 4
B3 Artifact Use Consistent With Intended Use: Yes
B3 Elaboration: Section 4
B4 Data Contains Personally Identifying Info Or Offensive Content: No
B4 Elaboration: We use publicly available datasets, that does not contain any personally identifying info or offensive content
B5 Documentation Of Artifacts: Yes
B5 Elaboration: Section 4
B6 Statistics For Data: Yes
B6 Elaboration: Appendix B
C Computational Experiments: Yes
C1 Model Size And Budget: Yes
C1 Elaboration: Appendix E
C2 Experimental Setup And Hyperparameters: Yes
C2 Elaboration: Appendix C
C3 Descriptive Statistics: Yes
C3 Elaboration: Section 4
C4 Parameters For Packages: Yes
C4 Elaboration: Section 4
D Human Subjects Including Annotators: No
D1 Instructions Given To Participants: N/A
D2 Recruitment And Payment: N/A
D3 Data Consent: N/A
D4 Ethics Review Board Approval: N/A
D5 Characteristics Of Annotators: N/A
E Ai Assistants In Research Or Writing: No
E1 Information About Use Of Ai Assistants: N/A
Association For Computational Linguistics - Blind Submission License Agreement: On behalf of all authors, I agree
Reviewing No Volunteers Reason: All authors are new to the ACL community.
TLDR: We leverage HT-SR theory to design a layer-wise fine-tuning scheme for LLMs and SciML models.
Preprint: no
Preprint Status: We are considering releasing a non-anonymous preprint in the next two months (i.e., during the reviewing process).
Preferred Venue: EMNLP 2024
Consent To Share Data: yes

Consent To Share Submission Details: On behalf of all authors, we agree to the terms above to share our submission details.
Submission Number: 1834

Discussion (?id=tiSCZsHs9V#discussion)

Filter by reply type... | Filter by author... | Search keywords... | Sort: Newest First

Everyone | Submission1834... | Submission1834 Area... | Submission1834 Authors | Submission1834... | Program Chairs | Submission1834... | Submission1834... | Submission1834... | Submission1834...

19 / 19 replies shown

Add: **Author-Editor Confidential Comment**

Meta Review of Submission1834 by Area Chair 8Fma

Meta Review by Area Chair 8Fma 03 Aug 2024 at 18:27 (modified: 22 Aug 2024 at 17:14)
Senior Area Chairs, Area Chairs, Authors, Reviewers Submitted, Program Chairs, Commitment Readers Revisions (/revisions?id=aOcV0eOdML)

Metareview:
Cleverly adapting the learning rate to improve the performance of fine tuning in low-data applications is an important and timely research, and these authors have performed extensive experiments to the strength of their proposed techniques. The applied portion of this paper focuses mainly on natural language understanding benchmark tasks, but the techniques presented are likely to work in many other circumstances.

Summary Of Reasons To Publish:
The techniques presented here are tested and examined in many different ways, and in my view the presentation is exemplary. The experiments show results that are pushed into extreme subsampling and there's an impressive range and rigor here. There are many potential avenues to further explore other layer weight schedules.

Summary Of Suggested Revisions:
The reviewers asked for additional experimental support and the authors made substantial additional contributions both by evaluating on an additional benchmark as well as providing statistical tests of significance. This is quite a bit of additional effort and resulted in a much stronger final paper.

Overall Assessment: 5 = The paper is largely complete and there are no clear points of revision
Best Paper Ae: Maybe
Best Paper Ae Justification:
This paper is pretty impressive, although I'm not that familiar with the related work cited and so I don't know how significant the results are. However, the paper is really very well written and explains the motivation and implementation with considerable clarity and provides support for the findings with an impressive range of evaluations.

Ethical Concerns:
There are no concerns with this submission

Needs Ethics Review: No
Author Identity Guess: 1 = I do not have even an educated guess about author identity.

Add: **Author-Editor Confidential Comment**

A gentle reminder that the Author/Reviewer discussion period will end soon.

Official Comment
by Authors (Tianyu Pang (/profile?id=~Tianyu_Pang2), Yaoqing Yang (/profile?id=~Yaoqing_Yang1), Pu Ren (/profile?id=~Pu_Ren1), Yuanzhe Hu (/profile?id=~Yuanzhe_Hu1), +2 more (/group/info?id=aclweb.org/ACL/ARR/2024/June/Submission1834/Authors))
29 Jul 2024 at 20:10 (modified: 22 Aug 2024 at 17:14) Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Commitment Readers
Revisions (/revisions?id=sjmRHDc8JX)

Comment:
Dear Reviewers,

Thank you for your constructive feedback. We have provided additional experimental results in our response. Since it is close to the end of the discussion period, please let us know if we should include anything further in the revised draft. We are more than happy to clarify if anything is unclear.

Best regards,

Authors

Add: **Author-Editor Confidential Comment**

Official Review of Submission1834 by Reviewer Z3xH

Official Review by Reviewer Z3xH 22 Jul 2024 at 16:39 (modified: 22 Aug 2024 at 17:14)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer Z3xH, Commitment Readers Revisions (/revisions?id=LR1maUzDyZ)

Paper Summary:
The paper investigates an important problem of model balancing for low-data training and fine-tuning. The paper first notes an important observation inspired by the Heavy-Tailed Self-Regularization (HT-SR) theory, revealing a trend of imbalanced training quality across different layers of the model through the lens of empirical spectral densities (ESDs). Then, correspondingly, the paper proposes to adapt an existing layer-wise learning rate scheduler, TempBalance, to balance layers' training qualities and improve low-data training and fine-tuning in NLP and SciML tasks. The paper conducts a number of evaluations with RoBERTa models and GLUE tasks, showing the proposed approach improves task performance in ultra-low data budgets.

Summary Of Strengths:
The paper is nicely written and well-contextualized. The illustration is clear and the analysis appears rigorous.

The observations are insightful and relevant and the development of the method is natural.

Evaluation results are carefully analyzed, reporting results for a number of Subsampling Ratio and providing standard deviations. This is very favorable.

Summary Of Weaknesses:
It seems to me the major limitation of this paper is its evaluations.

The paper conducted all evaluations with BERT-style models and GLUE tasks. What's the reason for this choice?

This is a tricky setup. GLUE tasks are diverse and tend to have different preferences for data and training. On the other hand, the performance of some tasks saturates easily and the performance of others might be hard to improve. Other than conducting large-scale pre-training or continued pre-training, it is often difficult to achieve visible improvements in GLUE performance. And it may or may not correlate with the performance of other tasks such as SWAG or SQuAD.

For example, SST-2 is a relatively easy task where its performance reaches 90% with as few as 0.5% training data. When trained on even fewer data, the performance is known to be quite volatile and it could be attributed to many largely random factors. I don't think performance gain achieved in this range is strong evidence.

I like the observations and the method development of this paper. When I see the experiment results, it is not very surprising to me the improvements in GLUE tasks are not very significant– which is difficult for many methods.

Since this paper is posed for low-data regimes and scientific machine learning (SciML), why not try some more domain-specific tasks? For example, tasks from this paper [https://arxiv.org/abs/2004.10964 (https://arxiv.org/abs/2004.10964)].

SST-2 is a relatively easy task where its performance reaches 90% with as few as 0.5% training data. When trained on even fewer data, the performance is known to be quite volatile and it could be attributed to many largely random factors. I don't think performance gain achieved in this range is strong evidence.

Comments Suggestions And Typos:
Why choose this evaluation setup? Why invest all efforts into one evaluation setup (RoBERTa/GLUE) rather than experimenting with more diverse scenarios?

If experiment results can be improved, the manuscript could be made much more impactful.

Confidence: 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.
Soundness: 3.5
Overall Assessment: 3 = Good: This paper makes a reasonable contribution, and might be of interest for some (broad or narrow) sub-communities, possibly with minor revisions.
Best Paper: No
Limitations And Societal Impact:
Discussions are adequate.

Ethical Concerns:
None

Needs Ethics Review: No
Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.
Datasets: 1 = No usable datasets submitted.
Software: 1 = No usable software released.
Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources
Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Add: **Author-Editor Confidential Comment**

Rebuttal by Authors (1/2)

Official Comment
by Authors (Tianyu Pang (/profile?id=~Tianyu_Pang2), Yaoqing Yang (/profile?id=~Yaoqing_Yang1), Pu Ren (/profile?id=~Pu_Ren1), Yuanzhe Hu (/profile?id=~Yuanzhe_Hu1), +2 more (/group/info?id=aclweb.org/ACL/ARR/2024/June/Submission1834/Authors))
28 Jul 2024 at 20:27 (modified: 22 Aug 2024 at 17:14) Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer Z3xH, Commitment Readers
Revisions (/revisions?id=dcaWOn30Wj)

Comment:
Thanks for your insightful and constructive comments. We address your concerns as follows.

Q1: What's the reason for choosing BERT-style models and GLUE tasks?

1.1 Reason for choosing the RoBERTa/GLUE setup.

We conduct most experiments on the RoBERTa-base model and GLUE tasks because this is one of the most popular choices in recent works on optimization methods in NLP [1,2]. However, this is not the only setting.

1.2 Other settings evaluated in the paper

1. In addition to evaluations with BERT-style models and GLUE tasks, in Appendix D.5 of our paper, we also evaluate our method using LLaMA-7B models with LoRA adapters on the ScienceQA dataset, with results shown in Table 18. We show that our method, LoRA+TB, yields better performance than using LoRA alone when training LLMs with limited data.
2. Furthermore, in Appendix D.3, we explore our method on the more complex SuperGLUE tasks. In Table 16 of our paper, we evaluate our method on six SuperGLUE tasks under different subsampling ratios, and we show that our method effectively increases test performance compared to the baseline method in most low-data regimes.
3. Lastly, In Figure 4, Table 15 and Table 19 of our paper, we consider Neural PDE solving tasks in SciML and train FNO and Unet models on 1D CFD, 2D CFD and DarcyFlow datasets in low-data regimes. We find that our method offers better performance under different subsampling ratios.

All results above are obtained by running 3 random seeds and the hyperparamters we used are provided in Appendix C. These results demonstrate the potential of our method for more diverse scenarios.

[1] Hu et al. "Lora: Low-rank adaptation of large language models", 2022

[2] Bahri et al. "Sharpness-aware minimization improves language model generalization", 2022

Add: **Author-Editor Confidential Comment**

Rebuttal by Authors (2/2)

Official Comment

by Authors (Tianyu Pang (/profile?id=~Tianyu_Pang2), Yaoqing Yang (/profile?id=~Yaoqing_Yang1), Pu Ren (/profile?id=~Pu_Ren1), Yuanzhe Hu (/profile?id=~Yuanzhe_Hu1), +2 more (/group/info?id=aclweb.org/ACL/ARR/2024/June/Submission1834/Authors))

28 Jul 2024 at 20:28 (modified: 22 Aug 2024 at 17:14)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer Z3xH, Commitment Readers Revisions (/revisions?id=sYEWDkTDtK)

Comment:

Q2: Why not experimenting with more diverse scenarios?

2.1 Evaluation on question-answering fine-tuning task (SQuAD)

To better demonstrate the effectiveness of our method (TB) in diverse scenarios, we present the results of applying our method to fine-tune RoBERTa-base models on the SQuAD dataset. In the table below, we show that our method outperforms the baseline full fine-tuning (FT) under different subsampling ratios.

The experimental setup is as follows: For TB and the baseline, we train the RoBERTa-base model for 10 epochs with a batch size of 24 using the AdamW optimizer, with a warmup rate of 0.06 and linear learning rate decay. We follow the detailed hyperparameter settings from [3]. The mean and standard deviation of test accuracy across 3 random seeds on the test set are reported.

SQuAD 1.1 Ratio	1%	5%	10%	20%	50%
FT	45.84±2.26	79.49±0.22	86.88±0.12	88.56±0.14	90.97±0.15
TB	**48.91±1.27**	**81.18±0.07**	**88.08±0.05**	**89.49±0.20**	**91.16±0.03**

2.2 Evaluations on domain-specific NLP fine-tuning tasks

We thank the reviewer for the suggestion to explore more domain-specific tasks. We adopt the training settings from [4] and test our method on five low-resource datasets: SciCite, ChemPort, Hyperpartisan News, RCT (500 samples), and SciERC. We find that our method (TB) consistently yields better performance on these low-resource, domain-specific tasks.

The experimental setup for these tasks is as follows: For TB and the baseline, we train the RoBERTa-base models for 10 epochs with a batch size of 16 and an initial learning rate of 3e-5. We use the AdamW optimizer and apply linear learning rate decay with a 0.06 warmup ratio. The mean and standard deviation of test accuracy across 3 random seeds on the test set are reported.

Dataset	SciCite	ChemPort	Hyperpartisan News	RCT(500 Sample)	SciERC
FT	79.86±1.76	82.63±0.22	88.72±3.63	79.70±0.39	86.11±1.24
TB	**81.06±1.35**	**83.72±0.25**	**93.85±1.26**	**80.12±0.37**	**87.61±0.76**

2.3 Evaluations on domain-specific SciML tasks

To explore diverse scenarios in SciML, we conduct experiments on low-data fine-tuning using the 2DCFD dataset with DPOT-Tiny and DPOT-Small models. In solving PDEs within SciML, we utilize foundational models pre-trained on various fluid dynamics datasets, which are then fine-tuned on a specific physical domain's PDE dataset (2DCFD). In the table below, we show that TB offers better improvements compared to the baseline FT under different subsampling ratios. We also report the error reduced as a percentage.

The experimental settings for SciML tasks are as follows: For TB and FT, we train the models for 500 epochs with a batch size of 160 for the Tiny model and 64 for the Small model, and a dropout rate of 1e-6. We test initial learning rates among {0.001, 0.0005, 0.00025, 0.0001, 0.00005}. We use the Adam optimizer, and decay the learning rate by $\gamma = 0.5$ every 50 epochs. The mean and standard deviation of nRMSE (where lower is better) across 3 random seeds on the test set are reported.

DPOT-Tiny Ratio	5%	10%	25%	50%	100%
FT	1.863e-02±1.067e-05	1.747e-02±1.502e-05	1.543e-02±4.008e-05	1.309e-02±2.356e-05	1.096e-02±3.875e-05
TB	**1.856e-02±3.646e-05**	**1.730e-02±1.173e-05**	**1.517e-02±2.807e-05**	**1.283e-02±2.494e-05**	**1.078e-02±4.527e-05**

DPOT-Small Ratio	5%	10%	25%	50%	100%
FT	1.546e-02±3.346e-05	1.426e-02±1.157e-05	1.226e-02±2.094e-05	1.025e-02±2.063e-05	8.400e-03±1.030e-05
TB	**1.539e-02±1.328e-05**	**1.415e-02±1.890e-05**	**1.203e-02±1.313e-05**	**1.005e-02±8.860e-06**	**8.193e-03±1.509e-05**

[3] Liu, Yinhan et al. "Roberta: A robustly optimized bert pretraining approach", 2019

[4] Gururangan, Suchin, et al. "Don't stop pretraining: Adapt language models to domains and tasks", 2020

Add: **Author-Editor Confidential Comment**

Replying to Rebuttal by Authors (2/2)

Thanks for the responses

Official Comment by Reviewer Z3xH 28 Jul 2024 at 22:01 (modified: 22 Aug 2024 at 17:14)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer Z3xH, Commitment Readers Revisions (/revisions?id=lejkffvsey)

Comment:

Thanks to the authors for the responses. I have no further questions. The manuscript is good to me.

Add: **Author-Editor Confidential Comment**

Replying to Thanks for the responses

Thank you for increasing the score!

Official Comment

by Authors (Tianyu Pang (/profile?id=~Tianyu_Pang2), Yaoqing Yang (/profile?id=~Yaoqing_Yang1), Pu Ren (/profile?id=~Pu_Ren1), Yuanzhe Hu (/profile?id=~Yuanzhe_Hu1), +2 more (/group/info?id=aclweb.org/ACL/ARR/2024/June/Submission1834/Authors))

28 Jul 2024 at 22:12 (modified: 22 Aug 2024 at 17:14)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer Z3xH, Commitment Readers Revisions (/revisions?id=SpWhhBDJ2b)

Comment:

Thank you for your positive feedback and for increasing the score. We will include the additional experiments and text in the revised draft.

Add: **Author-Editor Confidential Comment**

Official Review of Submission1834 by Reviewer h1jN

Official Review by Reviewer h1jN 20 Jul 2024 at 03:40 (modified: 22 Aug 2024 at 17:14)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer h1jN, Commitment Readers Revisions (/revisions?id=KgqGO8bRpg)

Paper Summary:

The paper proposes a method of adjusting layer-wise learning rate of model that helps model finetuning in a low-data training setting. The method is based on the Heavy-tailed self-regularization (HT-SR) theory, which states that well-trained neural network models have a heavy-tail phenomenon in the histogram of eigenvalue distribution. Using this theory, the paper adapts TempBalance, a layer-wise learning rate scheduler, to set the learning rate based on the coefficient of a power law regression of the eigenvalues of the parameter matrices, where a smaller power coefficient means a better distribution and a lower learning rate, and a larger one means a higher learning rate. The method increases the test accuracy of finetuned models on scientific datasets.

Summary Of Strengths:

1. The paper provides an efficient way of tuning models by applying different learning rates on different layers.
2. The method performs comparably better when the data is smaller, showing the efficiency of the method in low-data setting.
3. The paper is neatly written, with clear figures and formulae, and easy to read.

Summary Of Weaknesses:

1. The method is based on the HT-SR theory where the key concept is the "well trained" model. This concept is clearly different from the "well trained" when talking about a specific dataset, since a model can be "well trained" on a general dataset, but not trained on a scientific dataset. Therefore, this HT phenomenon has limited relation to the knowledge a model has acquired. It is dubious that this would be a reasonable indicator for adjusting learning rate to better utilize the data.
2. Although the method does improve model performance, it seems that the improvement is somewhat marginal from the results. This could be due to the selection of tasks. Even if the subsampling ratio is very small, the results from figure 2 only drop no more than 0.4%. This cannot prove that the method makes a fundamental difference.
3. In ablation study, only two different learning rate scheduling functions are compared. There are possibly other curves that perform better.

Comments Suggestions And Typos:

See above

Confidence: 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.
Soundness: 3 = Acceptable: This study provides sufficient support for its major claims/arguments. Some minor points may need extra support or details.
Overall Assessment: 3 = Good: This paper makes a reasonable contribution, and might be of interest for some (broad or narrow) sub-communities, possibly with minor revisions.
Best Paper: No
Needs Ethics Review: No
Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.
Datasets: 1 = No usable datasets submitted.
Software: 4 = Useful: I would recommend the new software to other researchers or developers for their ongoing work.
Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources
Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Add: **Author-Editor Confidential Comment**

Thank you for the responses

Official Comment by Reviewer h1jN 29 Jul 2024 at 20:19 (modified: 22 Aug 2024 at 17:14)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer h1jN, Commitment Readers Revisions (/revisions?id=DCSM7d6cxf)

Comment:

I would like to thank the authors for their responses in detail. These answer many of my questions and confusions, and I have raised my score accordingly.

Add: **Author-Editor Confidential Comment**

Thank you for your reply!

Official Comment

by Authors (Tianyu Pang (/profile?id=~Tianyu_Pang2), Yaoqing Yang (/profile?id=~Yaoqing_Yang1), Pu Ren (/profile?id=~Pu_Ren1), Yuanzhe Hu (/profile?id=~Yuanzhe_Hu1), +2 more (/group/info?id=aclweb.org/ACL/ARR/2024/June/Submission1834/Authors))

29 Jul 2024 at 20:29 (modified: 22 Aug 2024 at 17:14)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer h1jN, Commitment Readers Revisions (/revisions?id=JNAM4eatit)

Comment:

Thank you for your positive response and for raising the score. We will include the additional experimental results and texts in our revised draft.

Add: **Author-Editor Confidential Comment**

Rebuttal by Authors (1/3)

Official Comment

by Authors (Tianyu Pang (/profile?id=~Tianyu_Pang2), Yaoqing Yang (/profile?id=~Yaoqing_Yang1), Pu Ren (/profile?id=~Pu_Ren1), Yuanzhe Hu (/profile?id=~Yuanzhe_Hu1), +2 more (/group/info?id=aclweb.org/ACL/ARR/2024/June/Submission1834/Authors))

28 Jul 2024 at 20:42 (modified: 22 Aug 2024 at 17:14) Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer h1jN, Commitment Readers

Revisions (/revisions?id=dTEryqY73V)

Comment:

Thanks for your insightful and constructive comments. We address your concerns as follows.

Q1: This HT phenomenon has limited relation to the knowledge a model has acquired during fine-tuning. It is dubious that this would be a reasonable indicator for adjusting learning rate to better utilize the data.

We clarify that the HT phenomenon we measured is related to the knowledge a model acquires during fine-tuning. We note that the HT phenomena change because of weight matrices updated during fine-tuning (a process of acquiring task-specific knowledge), and we estimate it using the PL_Alpha_Hill metric throughout this process. This dynamic HT estimation accounts for the task-specific knowledge acquired during fine-tuning. We determine the learning rate schedule dynamically based on the updated HT estimation. Therefore, the HT metric used in our current method already incorporates the data.

A new ablation study shows that our current method outperforms a baseline (TB_Fix) that doesn't account for the knowledge a model acquires during fine-tuning. To demonstrate that the HT phenomenon estimation in our current method is closely related to the knowledge a model acquires during fine-tuning, we designed another baseline method called **TB_Fix**. TB_Fix does not incorporate information during fine-tuning and only uses information from the pre-trained model. It adjusts the learning rate based ONLY on the initial ESDs of the pre-trained model throughout fine-tuning. We show that our method, TB, which incorporates fine-tuning and specific dataset information, outperforms TB_Fix, which does not. The experimental setup is training a RoBERTa-base model on the QNLI task with different subsampling ratios. The results below are obtained by running 3 random seeds, and the hyperparameters we used are the same as those in Appendix C.

Subsampling Ratio	0.02%	0.05%	0.1%	0.5%	1%	5%
FT	53.69±0.44	71.31±1.29	73.57±0.90	82.68±0.43	84.09±0.36	87.94±0.08
TB_Fix	57.83±6.14	71.39±2.19	75.50±0.97	83.43±0.31	84.31±0.49	88.00±0.28
TB (ours)	**58.11±6.29**	**72.83±1.65**	**75.78±0.47**	**84.30±0.46**	**84.47±0.55**	**88.24±0.08**

Add: **Author-Editor Confidential Comment**

Rebuttal by Authors (2/3)

Official Comment

by Authors (Tianyu Pang (/profile?id=~Tianyu_Pang2), Yaoqing Yang (/profile?id=~Yaoqing_Yang1), Pu Ren (/profile?id=~Pu_Ren1), Yuanzhe Hu (/profile?id=~Yuanzhe_Hu1), +2 more (/group/info?id=aclweb.org/ACL/ARR/2024/June/Submission1834/Authors))

28 Jul 2024 at 20:44 (modified: 22 Aug 2024 at 17:14)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer h1jN, Commitment Readers Revisions (/revisions?id=BVT274qBMc)

Comment:

Q2: Improvement is somewhat marginal from the results.

2.1 Results in our paper

Our method significantly improves performance in low-data training. In Tables 13 and 15, we show that our method TB can improve the RoBERTa-base model trained on the SST-2 dataset by up to 9.9% and reduce the nRMSE of the FNO trained on the 2D Compressible Navier-Stokes dataset by 14.47%. These are not marginal improvements.

2.2 Statistical testing on the significance of improvement

To further demonstrate that our method brings significant performance improvements, we perform statistical testing on the test results of our algorithm and the baseline method. We define the Null Hypothesis (H0) as "There is no significant difference in performance between our algorithm and the baseline," and the Alternative Hypothesis (H1) as "Our algorithm performs significantly better than the baseline." We run experiments of training RoBERTa-base on SST-2 for 10 random seeds and perform t-tests on the results. We present the results in the table below:

Ratio	0.02%	0.1%	0.5%	1%	5%
P-value	3.85e-9	0.13	0.003	0.003	4.06e-5

As shown in the table above, the t-test yielded p-values below the standard significance threshold of 0.05 across almost all subsampling ratios. This indicates that the observed differences are statistically significant at a 95% confidence level, suggesting that the improvement of our method compared to the baseline is indeed significant.

2.3 More evaluation on question-answering task (SQuAD)

To better demonstrate the effectiveness of our method (TB) in more diverse scenarios, we present the results of applying our method to fine-tune RoBERTa-base models on the SQuAD dataset. In the table below, we show that our method outperforms baseline full fine-tuning (FT) under different subsampling ratios.

Here is the experimental setup: For TB and the baseline, we train the RoBERTa-base model for 10 epochs with a batch size of 24 using the AdamW optimizer with a warmup rate of 0.06 and linear learning rate decay. We follow the detailed hyperparameter settings from [1]. The mean and standard deviation of test accuracy across 3 random seeds on the test set are reported.

SQuAD 1.1 Ratio	1%	5%	10%	20%	50%
FT	45.84±2.26	79.49±0.22	86.88±0.12	88.56±0.14	90.97±0.15
TB	**48.91±1.27**	**81.18±0.07**	**88.08±0.05**	**89.49±0.20**	**91.16±0.03**

2.4 More evaluations on domain-specific NLP tasks

Additionally, we also test our method on more domain-specific tasks in low-data regimes. We adopt the training setting from [2] and evaluate our method on five low-resource datasets: SciCite, ChemPort, Hyperpartisan News, RCT (500 samples), and SciERC. We find that our method (TB) consistently yields better performance on these low-resource domain-specific tasks.

The experimental setup for these tasks is as follows: For TB and the baseline, we train the RoBERTa-base models for 10 epochs with a batch size of 16 and an initial learning rate of 3e-5. We use the AdamW optimizer and linear learning rate decay with a 0.06 warmup ratio. The mean and standard deviation of test accuracy across 3 random seeds on the test set are reported.

Dataset	SciCite	ChemPort	Hyperpartisan News	RCT(500 Sample)	SciERC
FT	79.86±1.76	82.63±0.22	88.72±3.63	79.70±0.39	86.11±1.24
TB	**81.06±1.35**	**83.72±0.25**	**93.85±1.26**	**80.12±0.37**	**87.61±0.76**

[1] Liu, Yinhan et al. "Roberta: A robustly optimized bert pretraining approach", 2019

[2] Gururangan, Suchin, et al. "Don't stop pretraining: Adapt language models to domains and tasks", 2020

Add: **Author-Editor Confidential Comment**

Replying to Rebuttal by Authors (2/3)

Rebuttal by Authors (3/3)

Official Comment

by Authors (Tianyu Pang (/profile?id=~Tianyu_Pang2), Yaoqing Yang (/profile?id=~Yaoqing_Yang1), Pu Ren (/profile?id=~Pu_Ren1), Yuanzhe Hu (/profile?id=~Yuanzhe_Hu1), +2 more (/group/info?id=aclweb.org/ACL/ARR/2024/June/Submission1834/Authors))

28 Jul 2024 at 20:45 (modified: 22 Aug 2024 at 17:14)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer h1jN, Commitment Readers Revisions (/revisions?id=msBcYTTfQk)

Comment:

Q3: Only two different learning rate scheduling functions are compared. There are possibly other curves that perform better.

We evaluated three alternative learning rate assignment functions: Square root (Sqrt), Log2, and Exponent (Exp):

- Sqrt : $f_t(i) = \eta_t \frac{\sqrt{\alpha_t^i}}{\frac{1}{L}\sum_{j=1}^{L}\sqrt{\alpha_t^j}}$,
- Log2: $f_t(i) = \eta_t \frac{log(\alpha_t^i)}{\frac{1}{L}\sum_{j=1}^{L} log(\alpha_t^j)}$,
- Exp: $f_t(i) = \eta_t 10^{(\alpha_t^i - \overline{\alpha_t})s}$

Here, η_t denotes the base global learning rate at epoch t, α_t^i is the PL_Alpha_Hill estimate of the layer i at epoch t, $\overline{\alpha_t}$ is the mean PL_Alpha_Hill across all layers, s is a tunable hyperparameter and L is the total number of model layers.

As shown in the table below, TB with the sigmoid learning rate schedule surpasses all the other TB designs when tested on Roberta-base model on QNLI dataset. All hyperparameters are consistent with our paper. Each experiment was conducted with 3 random seeds.

Ratio	0.05%	0.1%	0.5%	1%
FT	71.31±1.29	73.57±0.90	82.68±0.43	84.09±0.36
Sigmoid (ours)	**72.83±1.65**	**75.78±0.47**	**84.30±0.46**	84.47±0.55
Linear_map	72.76±1.54	73.49±2.92	83.87±0.61	**84.60±0.07**
Sqrt	70.82±2.18	74.20±2.51	82.73±0.34	84.12±0.67
Log2	70.41±3.22	74.71±1.97	82.93±0.63	84.23±0.29
Exp	66.09±2.29	71.88±3.58	82.45±0.32	83.67±0.39

Add: **Author-Editor Confidential Comment**

Replying to Rebuttal by Authors (3/3)

A gentle reminder that the deadline for the Author/Reviewer discussion is approaching.

Official Comment

by Authors (Tianyu Pang (/profile?id=~Tianyu_Pang2), Yaoqing Yang (/profile?id=~Yaoqing_Yang1), Pu Ren (/profile?id=~Pu_Ren1), Yuanzhe Hu (/profile?id=~Yuanzhe_Hu1), +2 more (/group/info?id=aclweb.org/ACL/ARR/2024/June/Submission1834/Authors))

29 Jul 2024 at 11:15 (modified: 22 Aug 2024 at 17:14)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer h1jN, Commitment Readers Revisions (/revisions?id=7UPxrBpoM8)

Comment:

Dear Reviewer h1jN:

Thank you for your constructive feedback. As the discussion period is closing soon, we'd like to follow up and see if the Reviewer has the chance to consider our response. The additional experimental results have been provided in our response. Please let us know if we should include anything further in the revised draft. We are more than happy to clarify if anything is unclear.

Best,

Authors

Add: **Author-Editor Confidential Comment**

Official Review of Submission1834 by Reviewer VV6m

Official Review by Reviewer VV6m 19 Jul 2024 at 17:18 (modified: 22 Aug 2024 at 17:14)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer VV6m, Commitment Readers Revisions (/revisions?id=Wb4EW7MsIR)

Paper Summary:
The paper presents an adaptation of the TempBalance algorithm, which dynamically adjusts learning rates based on the empirical spectral density of weight matrices to improve model performance in low-data regimes.

Summary Of Strengths:
- Extensive experiments and clearly defined contributions: The paper offers a thorough experimental evaluation. The contributions are well-defined and demonstrate improvements in specific low-data scenarios.
- Compatibility and versatility: TempBalance is compatible with existing optimization methods like AdaFactor and improves performance when used in conjunction.
- Diagnostic purposes: The algorithm leverages heavy-tailed self-regularization theory to diagnose and improve training quality, providing a way to evaluate the state of training via the empirical spectral density.

Summary Of Weaknesses:
- Marginal gains in performance: While TempBalance generally improves performance in low-data regimes compared to standard fine-tuning, the gains can vary depending on the task, and they are often marginal. Additionally, it is regularly outperformed by AdaFactor.
- Lack of statistical testing: The paper does not include statistical testing to rigorously validate the performance improvements of TempBalance.
- Computational costs: The approach requires frequent calculations of ESD, which can significantly increase the total training time. As the authors pointed out, the computational overhead primarily comes from the SVD process, which could be problematic with larger model sizes.

Comments Suggestions And Typos:

Typos

L374: selct => select

Comments after author discussion

The authors have addressed some of my concerns in their response and I adjusted my score accordingly.

Confidence: 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.
Soundness: 4 = Strong: This study provides sufficient support for all of its claims/arguments. Some extra experiments could be nice, but not essential.
Overall Assessment: 4 = This paper represents solid work, and is of significant interest for the (broad or narrow) sub-communities that might build on it.
Best Paper: No
Needs Ethics Review: No
Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.
Datasets: 1 = No usable datasets submitted.
Software: 1 = No usable software released.
Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources
Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Add: **Author-Editor Confidential Comment**

Rebuttal by Authors (1/2)

Official Comment

by Authors (Tianyu Pang (/profile?id=~Tianyu_Pang2), Yaoqing Yang (/profile?id=~Yaoqing_Yang1), Pu Ren (/profile?id=~Pu_Ren1), Yuanzhe Hu (/profile?id=~Yuanzhe_Hu1), +2 more (/group/info?id=aclweb.org/ACL/ARR/2024/June/Submission1834/Authors))

28 Jul 2024 at 20:31 (modified: 22 Aug 2024 at 17:14) Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer VV6m, Commitment Readers

Revisions (/revisions?id=mhsZUS1AcE)

Comment:
Thanks for your insightful and constructive comments. We address your concerns as follows.

Q1: Marginal gains in performance.

1. Our method significantly improve performance in low-data training. In Tables 13 and 15, we show that our method TB can improve the RoBERTa-base model trained on the SST-2 dataset by up to 9.9% and reduce the nRMSE of the FNO trained on the 2D Compressible Navier-Stokes dataset by 14.47%. These are not marginal improvements.
2. To better demonstrate the effectiveness of our method (TB) in more diverse scenarios, we present the results of applying our method to fine-tune RoBERTa-base models on the SQuAD dataset. In the table below, we show that our method outperforms baseline full fine-tuning (FT) under different subsampling ratios. Here is the experimental setup: For TB and the baseline, we train the RoBERTa-base model for 10 epochs with a batch size of 24 using the AdamW optimizer with a warmup rate of 0.06 and linear learning rate decay. We follow the detailed hyperparameter settings from [1]. The mean and standard deviation of test accuracy across 3 random seeds on the test set are reported.

SQuAD 1.1 Ratio	1%	5%	10%	20%	50%
FT	45.84±2.26	79.49±0.22	86.88±0.12	88.56±0.14	90.97±0.15
TB	**48.91±1.27**	**81.18±0.07**	**88.08±0.05**	**89.49±0.20**	**91.16±0.03**

Additionally, we also test our method on more domain-specific tasks in low-data regimes. We adopt the training setting from [2] and evaluate our method on five low-resource datasets: SciCite, ChemPort, Hyperpartisan News, RCT (500 samples), and SciERC. We find that our method (TB) consistently yields better performance on these low-resource domain-specific tasks.

The experimental setup for these tasks is as follows: For TB and the baseline, we train the RoBERTa-base models for 10 epochs with a batch size of 16 and an initial learning rate of 3e-5. We use the AdamW optimizer and linear learning rate decay with a 0.06 warmup ratio. The mean and standard deviation of test accuracy across 3 random seeds on the test set are reported.

Dataset	SciCite	ChemPort	Hyperpartisan News	RCT(500 Sample)	SciERC
FT	79.86±1.76	82.63±0.22	88.72±3.63	79.70±0.39	86.11±1.24

TB	**81.06±1.35**	**83.72±0.25**	**93.85±1.26**	**80.12±0.37**	**87.61±0.76**

[1] Liu, Yinhan et al. "Roberta: A robustly optimized bert pretraining approach", 2019

[2] Gururangan, Suchin, et al. "Don't stop pretraining: Adapt language models to domains and tasks", 2020

Q2: Our method TB is regularly outperformed by AdaFactor.

We agree with the reviewer that TB may be outperformed by AdaFactor. However, one advantage of our method is that TB can be used as an add-on with existing optimization methods to achieve further improvements, as noted by the reviewer in the "Strengths" section. Designing new optimizers often involves many baselines, and we believe it is more crucial to demonstrate that new approaches explore angles orthogonal to existing methods, thereby bringing in new improvements.

Add: **Author-Editor Confidential Comment**

Rebuttal by Authors (2/2)

Official Comment

by Authors (Tianyu Pang (/profile?id=~Tianyu_Pang2), Yaoqing Yang (/profile?id=~Yaoqing_Yang1), Pu Ren (/profile?id=~Pu_Ren1), Yuanzhe Hu (/profile?id=~Yuanzhe_Hu1), +2 more (/group/info?id=aclweb.org/ACL/ARR/2024/June/Submission1834/Authors))

28 Jul 2024 at 20:33 (modified: 22 Aug 2024 at 17:14)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer VV6m, Commitment Readers Revisions (/revisions?id=ugSfvtUWhp)

Comment:

Q3: The paper does not include statistical testing to rigorously validate the performance improvements of TempBalance.

We perform statistical testing to verify the effectiveness of our algorithm compared to baseline methods. We define the Null Hypothesis (H0) as "There is no significant difference in performance between our algorithm and the baseline," and the Alternative Hypothesis (H1) as "Our algorithm performs significantly better than the baseline." We run experiments of training RoBERTa-base on SST-2 for 10 random seeds and perform t-tests on the results. We present the results in the table below:

Ratio	**0.02%**	**0.1%**	**0.5%**	**1%**	**5%**
P-value	3.85e-9	0.13	0.003	0.003	4.06e-5

As shown in the table above, the t-test yielded p-values below the standard significance threshold of 0.05 across almost all subsampling ratios. This indicates that the observed differences are statistically significant at a 95% confidence level, suggesting that the improvement of our method compared to the baseline is indeed significant.

Q4: Computational costs.

We provide additional study on the computational costs of our TB method, with results presented in the tables below.

The following two tables depicts the time of one TB calculation on different models. For NLP tasks, we choose RoBERTa-base and RoBERTa-large:

Model	**RoBERTa-base**	**RoBERTa-large**
Time of Using TB once (sec)	3.13	11.15

For SciML tasks, we choose FNO, UNet, DPOT-Tiny and DPOT-Small:

Model	**FNO**	**UNet**	**DPOT-Tiny**	**DPOT-Small**
Time of Using TB once (sec)	0.012	0.59	0.239	1.37

The following table depicts the total time percentage of TB calculation during training under different subsampling ratios.

For NLP task, we measure the computation cost of TB when training RoBERTa-base model on the SST-2 dataset:

Time Percentage of TB During Training	**0.02%**	**0.05%**	**0.1%**	**0.5%**	**1%**	**5%**	**100%**
RoBERTa-base SST-2	24.03%	10.48%	7.98%	6.46%	2.19%	2.14%	0.87%

For SciML task, we measure the computation cost of TB when training FNO and UNet model on the 2DCFD dataset:

Time Percentage of TB During Training	**2.5%**	**10%**	**25%**	**50%**	**100%**
FNO, 2DCFD	0.97%	0.70%	0.31%	0.16%	0.084%
UNet, 2DCFD	19.25%	9.40%	4.62%	2.56%	1.41%

We can see that in both settings, the computational overhead of TB is acceptable and does not constitute a significant portion of the total training time.

Q5: Some typos

We thank the reviewer for pointing out the typo (selct => select), and we will fix it in the revised version of the paper.

Add: **Author-Editor Confidential Comment**

Replying to Rebuttal by Authors (2/2)

Official Comment by Reviewer VV6m

Official Comment by Reviewer VV6m 01 Aug 2024 at 05:38 (modified: 22 Aug 2024 at 17:14)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer VV6m, Commitment Readers Revisions (/revisions?id=ovAoaHzypU)

Comment:

Thank you for providing additional results. I have raised my score accordingly.

Add: **Author-Editor Confidential Comment**

Thank you for increasing the score and the positive feedback!

Official Comment

by Authors (Tianyu Pang (/profile?id=~Tianyu_Pang2), Yaoqing Yang (/profile?id=~Yaoqing_Yang1), Pu Ren (/profile?id=~Pu_Ren1), Yuanzhe Hu (/profile?id=~Yuanzhe_Hu1), +2 more (/group/info?id=aclweb.org/ACL/ARR/2024/June/Submission1834/Authors))

01 Aug 2024 at 05:46 (modified: 22 Aug 2024 at 17:14)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer VV6m, Commitment Readers Revisions (/revisions?id=aRBdoAQtBv)

Comment:

Dear Reviewer VV6m,

We sincerely appreciate your decision to increase the score and your positive feedback. We will include the additional experimental results and texts in our revised draft.

Add: **Author-Editor Confidential Comment**