View Meta-Reviews

Paper ID
67

Paper Title
Soft Threshold Weight Reparameterization for Learnable Sparsity

META-REVIEWER #3

META-REVIEW QUESTIONS

1. Please provide a meta-review for this paper that explains to both the program chairs and the authors the key positive and negative aspects of this submission. Because authors cannot see reviewer discussions, please also summarize any relevant points that can help improve the paper. Please be sure to make clear what your assessment of the pros/cons of this paper are, especially if your assessment is at odds with the overall reviewer scores. Please do not explicitly mention your recommendation in the meta-review (or you may have to edit it later).
All the authors agree that the paper brings significant contributions, including a convincing experimental evaluation of their method that vastly outperforms the baselines.

In their camera-ready, the authors should address the detailed comments, particularly those mentioned by Reviewers 1 and 2.

8. I agree to keep the paper and supplementary materials (including code submissions and Latex source), and reviews confidential, and delete any submitted code at the end of the review cycle to comply with the confidentiality requirements.
Agreement accepted

9. I acknowledge that my meta-review accords with the ICML code of conduct (see https://icml.cc/public/CodeOfConduct).
Agreement accepted

View Author Feedback

Paper ID
67

Paper Title
Soft Threshold Weight Reparameterization for Learnable Sparsity

AUTHOR FEEDBACK QUESTIONS

1. Author Response to Reviewers Please use this space to respond to any questions raised by reviewers, or to clarify any misconceptions. Please do not include any links to external material, nor include "late-breaking" results that are not responsive to reviewer concerns. We request that you understand that this year is especially difficult for many people, and to be considerate in your response.
We thank the reviewers for the positive reviews and constructive feedback. Code and models will be open-sourced.

R1:
1. L2 vs L1 loss on weights: For L1 regularization, it is not clear how to set the regularization parameter for each layer, as sparsity for each layer can be quite different. Our method allows us to learn the threshold parameter for each layer thus allowing automatic tuning of sparsity per layer while ensuring minor drop in accuracy (by minimizing the standard loss with L2 regularization). Also, weight-decay (L2 regularization) is standard in large-scale CNNs to make training stable (eg., ResNet - He et al., CVPR 2016).
2. Sub-gradient: We use the common sub-gradient for ReLU which is 1 if input is > 0 and 0 otherwise. Sign function is piecewise constant.
3. Relevance of Figure 1: Yes, one can understand the dynamics from Eq 3-4. However, in practice, the weight-decay on “s” contributes more to the threshold’s learning initially and has an inflection when the gradient due to the weights for “s” starts becoming more significant. Figure 1 shows the interaction (fight) between these two factors during the course of training while maximizing the accuracy resulting in stable end-to-end training.
4. Comparison to RigL: We agree with the issues in comparisons with RigL+ERK and SNFS which are end-to-end sparse training methods. However, they are among the very few methods with non-uniform layer-wise sparsity budgets which is a focus of our paper. Table 4 shows the use of heuristic ERK budget with an expensive dense gradient method like DNW having similar accuracy as uniform budget and STR but with >2x inference FLOPs which is very expensive.
STR uses sparse gradients as shown in Eq 4 that can reduce the total training FLOPs as weight-decay is an in-place scaling operation.

R2:
1. Missing references: Thank you, we will incorporate them.
2. Real-world utility of unstructured sparsity: We agree that the current commodity hardware (GPUs) are not optimized for sparse inference and research is ongoing to design faster sparse kernel operations for them (eg., Elsen et al., 2019). But there are millions of single-core processors forming the IoT ecosystem where sparsity in models directly translates to inference time speed-up. Furthermore, even for more powerful processors like EdgeTPU, the available RAM can be quite limited (e.g. 8MB for EdgeTPU), which makes sparse models more attractive.
3. Structured sparsity: We present generalization of STR to structured sparsity (low-rank) on RNNs as they are the operators of interest for typical time-series data analysis in IoT domain. But, our method can be extended to channel pruning in CNNs as well which uses group sparsity. Channel pruning techniques (eg., Li et al., ICLR 2017) typically have an importance scalar learnt for each of the filters. For any layer, STR can be applied to the vector with these importance factors. The dynamics and design choices are very similar to the low-rank experiments presented. We will add a proof of concept experiment to show the adaptability of STR for channel pruning.
4. STE methods: Thank you for the pointer. We will discuss the AutoPrune paper in the related works. We compared with DNW (Worstman et al. NeurIPS 2019), a SOTA STE based method that outperforms AutoPrune for ResNet50 on ImageNet-1K.

5. Ablation of g(): The practical constraints for g() are in Appendix A.1 and we will discuss the potential functions and their effects on dynamics in the next revision.

R3:
1. Setting of **λ**: Weight-decay parameter indeed controls the budget. All the hyper-parameters settings are in Appendix A.5. From our experience, one can intuitively figure them out or do a binary search after the initial couple of experiments. We also present workarounds to alleviate the problem of searching for the weight-decay parameter in lines 408-415 (right).

R5:
1. Dependence on hyperparameters: We acknowledge the dependence on **λ** and “s_init” as a drawback and provide some workarounds in section 5. We look forward to further exploring their interaction and necessity in the future. Please see R3’s rebuttal for more information.
2. Thresholding behavior: There is no thresholding schedule, but it is rather the learnt behavior of the threshold. The initial values are ~0 due to the nature of the sigmoid function and this range (eg., 20 epochs) is controlled by both s_init and **λ**. Please refer to point 3) in R1’s rebuttal for further explanation.

---

3. I certify that this author response conforms to the ICML Code of Conduct (https://www.icml.cc/public/CodeOfConduct)
Agreement accepted

View Reviews

Paper ID
67

Paper Title
Soft Threshold Weight Reparameterization for Learnable Sparsity

Reviewer #1

Questions

1. Please summarize the main claim(s) of this paper in two or three sentences.
The paper introduces an adaptive pruning method. Unlike some other pruning methods that fix a sparsity budget uniformly for all layers, this method optimizes the sparsity pattern directly. This is achieved by introducing a "re-parameterization" of the weights that includes a shrinkage operator, as well as adding weight decay to the standard objective. Extensive experiments are presented that demonstrate the merits of the proposed approach.

2. Merits of the Paper. What would be the main benefits to the machine learning community if this paper were presented at the conference? Please list at least one.
The proposed method is a novel and effective way to prune networks. The proposed method is also quite interesting and has some connections to sparse inference. There are also some interesting observations made in Section 4.1.2 that contradict previously observed results -- which can spark some potentially enlightening debate in the community.

3. Please provide an overall evaluation for this submission.
Very good paper, I would like to see it accepted.

4. Score Justification Beyond what you've written above as "merits", what were the major considerations that led you to your overall score for this paper?
Although I am not intimately familiar with this sub-field (pruning/sparse training), the experimental results presented in the paper seem convincing and competitive/exceeding SOTA. The method is inspired by previous work in sparse inference, particularly using the shrinkage function to sparsify weights. I particularly like sections 4.1.2 and 5 that explicitly state interesting findings as well as detail the limitations of the approach.

5. Detailed Comments for Authors Please comment on the following, as relevant: - The significance and novelty of the paper's contributions. - The paper's potential impact on the field of machine learning. - The degree to which the paper substantiates its main claims. - Constructive criticism and feedback that could help improve the work or its presentation. - The degree to which the results in the paper are reproducible. - Missing references, presentation suggestions, and typos or grammar improvements.
-Although interesting and well presented, the approach is a little bit contrived. The authors introduce weight decay (L2 loss on weights) and use the shrinkage function with adaptive threshold to minimize the weight decay loss + classification loss. In effect the authors are using an L1 sub-gradient with adaptive threshold to minimize an L2 loss. Why not use L1 loss on the weights instead of L2?
-Line 183: please provide a reference for "popular sub-gradient".
-Line 211: "Figure 1 shows that the threshold's dynamics are guided by..." -- how can you tell what the threshold's dynamics are guided by from that figure? I can more easily tell what they are guided by from Equations 3-4.
-Finally comparison to RigL isn't exactly fair since it does sparse training and not just pruning. The authors use dense gradients in this work whereas RigL uses sparse gradients AFAIK.

6. Please rate your expertise on the topic of this submission, picking the closest match.
I have seen talks or skimmed a few papers on this topic, and have not published in this area.

7. Please rate your confidence in your evaluation of this paper, picking the closest match.
I tried to check the important points carefully. It is unlikely, though possible, that I missed something that could affect my ratings.

12. I agree to keep the paper and supplementary materials (including code submissions and Latex source) confidential, and delete any submitted code at the end of the review cycle to comply with the confidentiality requirements.
Agreement accepted

13. I acknowledge that my review accords with the ICML code of conduct (see https://icml.cc/public/CodeOfConduct).
Agreement accepted

Reviewer #2

Questions

1. Please summarize the main claim(s) of this paper in two or three sentences.
The paper proposes a method to apply non-uniform unstructured pruning to DNNs with learnable sparsity targets for each layer. It further shows how to apply the same technique for structured pruning in RNNs. It beats all the recent SOTA methods by a significant margin for ultra high sparsity rates (98-99%).

2. Merits of the Paper. What would be the main benefits to the machine learning community if this paper were presented at the conference? Please list at least one.
It provides an easy to implement, non-heuristic method to learn layerwise sparsity targets, as well as the masks themselves. Contrary to the common intuition that later layers should be more heavily pruned, it shows that keeping these layers denser provides a better trade off between accuracy and FLOPS. It achieves the highest accuracy for ultra-sparse networks.

Moreover the paper is very well written and covers the literature quite well, and is therefore a good survey at the same time.

3. Please provide an overall evaluation for this submission.
Very good paper, I would like to see it accepted.

4. Score Justification Beyond what you've written above as "merits", what were the major considerations that led you to your overall score for this paper?
As the paper claims, it is an extremely hard combinatorial problem to set the right sparsity levels for each layer, which is why uniform pruning is common, and most non-uniform pruning methods are heuristics based (e.g. magnitude based). This paper shows that a very simple method beats all these other methods, with the help of just two hyper-parameters to fine tune.

Comparisons to other SOTA papers are excellent, and the results are significantly better.

5. Detailed Comments for Authors Please comment on the following, as relevant: - The significance and novelty of the paper's contributions. - The paper's potential impact on the field of machine learning. - The degree to which the paper substantiates its main claims. - Constructive criticism and feedback that could help improve the work or its presentation. - The degree to which the results in the paper are reproducible. - Missing references, presentation suggestions, and typos or grammar improvements.
The paper uses a very simple idea - learning a common magnitude threshold for each layer - and provides a straightforward algorithm to train it. The results are very significant, and I'd imagine they would be very easy to reproduce.

While there are other methods that the paper does not cite, which can result in non-uniform sparsity levels (e.g. MorphNet, AutoSlim), these methods are meant for structured pruning. AutoPrune (AAAI 2019) is similar and does unstructured pruning though, so it should be mentioned.

The results show that an ultra sparse network pruned with this method can achieve 10% higher accuracy than the other SOTA methods. One concern though, is that unstructured pruning is not easy to make use of to improve speed, as most frameworks only have dense implementations for the GPU. I think this is not addressed

too well in the paper. The Elsen paper is cited, but that's it. The paper does not give a sense of what the theoretical decrease in FLOPS would correspond to in terms of gains in speed. And if there is no gain in speed, why would one prefer unstructured pruning to structured one? This should be better clarified in the paper.

The structured pruning section feels like a bit of afterthought to me. There is no discussion on CNNs, just a low rank approximation of RNNs. I'd be much more interested in reading how this method could be applied for channel pruning for CNNs.

The paper uses a binary form of straight through estimator (STE) as the subgradient. It would be interesting to see a comparison of different alternatives here. An uncited but relevant paper, "AutoPrune: Automatic Network Pruning by Regularizing Auxiliary Parameters", compares many different STEs for instance.

While there are some missing references and the ablation study could be stronger (also in terms of comparing different g() functions), the paper is still quite strong and has very significant results and is easy to implement. So I think it is worth accepting into the conference.

After feedback:
Thanks for the explanations.

6. Please rate your expertise on the topic of this submission, picking the closest match.
I have published one or more papers in the narrow area of this submission.

7. Please rate your confidence in your evaluation of this paper, picking the closest match.
I am very confident in my evaluation of the paper. I read the paper very carefully and I am very familiar with related work.

8. Datasets If this paper introduces a new dataset, which of the following norms are addressed? (For ICML 2020, lack of adherence is not grounds for rejection and should not affect your score; however, we have encouraged authors to follow these suggestions.)
This paper does not introduce a new dataset (skip the remainder of this question).

12. I agree to keep the paper and supplementary materials (including code submissions and Latex source) confidential, and delete any submitted code at the end of the review cycle to comply with the confidentiality requirements.
Agreement accepted

13. I acknowledge that my review accords with the ICML code of conduct (see https://icml.cc/public/CodeOfConduct).
Agreement accepted

Reviewer #3

Questions

1. Please summarize the main claim(s) of this paper in two or three sentences.
This paper proposes soft threshold reparameterization (STR) to induce non-uniform sparsity based on learning the layer-wise pruning thresholds. For this, weight parameters are directly optimized in the projection.

2. Merits of the Paper. What would be the main benefits to the machine learning community if this paper were presented at the conference? Please list at least one.
This paper shows a good contribution in the sparsity-induced learning. STR can be useful to achieve non-uniform and layer-wise pruning in a deep neural network.

3. Please provide an overall evaluation for this submission.
Very good paper, I would like to see it accepted.

4. Score Justification Beyond what you've written above as "merits", what were the major considerations that led you to your overall score for this paper?
the weight decay parameter seems to be important as it controls the budget. How can it be obtained?

5. Detailed Comments for Authors Please comment on the following, as relevant: - The significance and novelty of the paper's contributions. - The paper's potential impact on the field of machine learning. - The degree to which the paper substantiates its main claims. - Constructive criticism and feedback that could help improve the work or its presentation. - The degree to which the results in the paper are reproducible. - Missing references, presentation suggestions, and typos or grammar improvements.
- The significance and novelty of the paper's contributions: the paper proposes STR to tackle non-uniform and layer-wise sparsity budget allocation in the pruning.
- The paper's potential impact on the field of machine learning: the current deep neural network is complicated, so the proposed method can alleviate the computational complexity.
- The degree to which the paper substantiates its main claims: it demonstrates the effectiveness in the results.
- Constructive criticism and feedback that could help improve the work or its presentation.
- The degree to which the results in the paper are reproducible.

6. Please rate your expertise on the topic of this submission, picking the closest match.
I have seen talks or skimmed a few papers on this topic, and have not published in this area.

7. Please rate your confidence in your evaluation of this paper, picking the closest match.
I tried to check the important points carefully. It is unlikely, though possible, that I missed something that could affect my ratings.

8. Datasets If this paper introduces a new dataset, which of the following norms are addressed? (For ICML 2020, lack of adherence is not grounds for rejection and should not affect your score; however, we have encouraged authors to follow these suggestions.)
This paper does not introduce a new dataset (skip the remainder of this question).

12. I agree to keep the paper and supplementary materials (including code submissions and Latex source) confidential, and delete any submitted code at the end of the review cycle to comply with the confidentiality requirements.
Agreement accepted

13. I acknowledge that my review accords with the ICML code of conduct (see https://icml.cc/public/CodeOfConduct).
Agreement accepted

Reviewer #5

Questions

1. Please summarize the main claim(s) of this paper in two or three sentences.
The paper proposes a method for inducing non-uniform sparsity across different layers of a neural network. It involves using a soft threshold for the weights where the threshold is learned per layer. This results in different sparsity budgets for different layers and hence a better allocation of model capacity across the layers to obtain the best performance. The paper also shows how the method can be applied for structured sparsity as well which is useful for deployment on commodity hardware.

2. Merits of the Paper. What would be the main benefits to the machine learning community if this paper were presented at the conference? Please list at least one.
* It is a very well written paper with a good literature review, explanation and analysis of the proposed method.
* The proposed non-uniform sparsity method manifests as a greater reduction in total FLOPs compared to the other sparsity inducing methods at the same level of sparsity. This is a serendipitous artifact even though the method is not directly trying to address reduction of overall FLOPs.
* The extension to structure sparsity is also simple as it can be applied on eigenvalues of a parameter tensor, hence resulting in low rank tensors.
* The authors have done extensive experiments on Imagenet, Google Speech commands and Human Activity Recognition dataset comparing the proposed method with other state of the art methods.

3. Please provide an overall evaluation for this submission.
Very good paper, I would like to see it accepted.

4. Score Justification Beyond what you've written above as "merits", what were the major considerations that led you to your overall score for this paper?
The only concern I have is that the amount of sparsity can only be controlled indirectly through the weight decay parameter and the initialization. The authors do mention this as a drawback in the discussion. I hope future work on related ideas can overcome this drawback.

5. Detailed Comments for Authors Please comment on the following, as relevant: - The significance and novelty of the paper's contributions. - The paper's potential impact on the field of machine learning. - The degree to which the paper substantiates its main claims. - Constructive criticism and feedback that could help improve the work or its presentation. - The degree to which the results in the paper are reproducible. - Missing references, presentation suggestions, and typos or grammar improvements.
Is there a thresholding schedule since the threshold value is almost zero till epoch 20 something or that is indeed the behavior of the method? Could you provide some thoughts on that behavior?

6. Please rate your expertise on the topic of this submission, picking the closest match.
I have seen talks or skimmed a few papers on this topic, and have not published in this area.

7. Please rate your confidence in your evaluation of this paper, picking the closest match.
I tried to check the important points carefully. It is unlikely, though possible, that I missed something that could affect my ratings.

8. Datasets If this paper introduces a new dataset, which of the following norms are addressed? (For ICML 2020, lack of adherence is not grounds for rejection and should not affect your score; however, we have encouraged authors to follow these suggestions.)
This paper does not introduce a new dataset (skip the remainder of this question).

12. I agree to keep the paper and supplementary materials (including code submissions and Latex source) confidential, and delete any submitted code at the end of the review cycle to comply with the confidentiality requirements.
Agreement accepted

13. I acknowledge that my review accords with the ICML code of conduct (see https://icml.cc/public/CodeOfConduct).
Agreement accepted