View Reviews

Paper ID

4265

Paper Title

Rotated Binary Neural Network

Reviewer #1

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.

This paper focuses on solving the quantization error in binary neural network. Their contributions are:

1、This paper for the first time talks about the quantization error from the angular bias between the weight vector and its binarization.

2、To solve the angular bias, this paper introduces a weight rotation scheme in the beginning of each training epoch. A bi-rotation is further devised to reduce the optimization complexity. The optimization objective is intuitive and the derivations are elaborated. Experiments have verified the effectiveness of closing angular bias.

3、 A training-aware approximation is proposed to compensate the “no gradient” problem in the sign function. Its superiority over existing methods is experimentally validated.

4、 Extensive experiments on CIFAR-10 and ImageNet show better performance of the proposed method.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.

Soundness of the claims:

To my knowledge, this is the first work that talks about the quantization error from angular bias in BNN. I think the angular bias indeed exists from a mathematical view and it has been validated by the authors. The discovery is insightful. Their solution of weight rotation is very interesting and attractive. I have carefully checked all the optimization formulations and confirmed their correctness. The training-aware approximation may not be a pioneer but shows significant difference and superiorities over the old methods. Experiments have been made to show its better performance.

Significance and novelty of the contribution:

In general, I agree with the novelty and contributions claimed in this paper. This paper is well written and easy to follow. The discussion on angular bias is the first time. Their solution of weight rotation is original and attracts my attention. Experimental results provide comparable accuracy performance and the available code is a bonus to the community.

Relevance to the NeurIPS community:

It is highly relevant. I have seen many related works accepted by NeurIPS each year.

3. Weaknesses: Explain the limitations of this work along the same axes as above.

I still need some clarifications from the authors. The concerns are listed below.

1、 It is unclear why ${\alpha}^i$ in Eq. (11) is limited within [0, 1]. In my opinion, the final weight vector is not necessary to be among the original weight and rotated one. Please clarify.

2、 The author performs rotations in the beginning of each training epoch. What if the rotation is applied each training iteration? Intuitively, this may feed back a better performance.

3、 There seems a contradiction between figure 2(b) and figure 5. In figure 2(b), the quantization error of the

proposed method is almost zero for most layers. Obviously, figure 5 provides larger quantization error though the two-mode distributions of weight values are centered around -1/+1.
4、 As an analogy, the activation quantization also suffers from the angular bias. However, the authors simply binarize the activations using sign function (see line 99 – line 100). This seems counter-intuitive.
5、 Typo and grammar errors:
Figure 2: and its binarizaton – and its binarization
Line 59: with significantly reduced complexity – with a significantly reduced complexity

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?
Yes.

5. Clarity: Is the paper well written?
Yes. A pleasure to read though some unclear parts in the weakness box need clarifying.

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?
Yes.

7. Reproducibility: Are there enough details to reproduce the major results of this work?
Yes

8. Additional feedback, comments, suggestions for improvement and questions for the authors:
See the box for weakness.

Post-rebuttal comments:
The author has addressed my questions.

9. Please provide an "overall score" for this submission.
8: Top 50% of accepted NeurIPS papers. A very good submission; a clear accept.

10. Please provide a "confidence score" for your assessment of this submission.
4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?
Yes

Reviewer #2

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.
The paper proposes to learn rotation matrices to rotate the floating-point vectors in each layer before projecting them to binary vectors while learning binary neural networks. This approach reduces the angular bias in the float to binary projection. The results on cifar and imagenet show improved results compared to the previous methods when both weights and activations are binarized.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.
1. The idea of learning rotation matrices to reduce the angular bias when projecting from float to binary is new and interesting.
2. The paper is clearly written and the method is sufficiently explained.
3. Extensive experiments show improvements over the compared baselines.

3. Weaknesses: Explain the limitations of this work along the same axes as above.

Some weaknesses/comments/questions of the proposed method in my opinion are as follows:

1) The improvement due to rotation seems small (ref. table 4) and the most of the improvement seems to be due to the gradient approximation. Please clarify.

2) How many additional learnable parameters are introduced while learning rotation matrices? Are they binary or floating point? Please provide a summary of new learnable parameters and actual (theoretical) improvement in memory and FLOPS. This is important to see as there are some parameters in the network is kept in floating point and additional parameters are introduced compared to the baselines.

3) What is the motivation for the specific form of gradient approximation (sec. 3.4) used in this paper? Why not simply use an existing function such as tanh and anneal using a temperature parameter to approximate the step function similar to [a]?

4) Is there any reason for explicitly learning the rotation matrices rather than penalizing the angle between floating-point and the binary vectors while learning? I believe, this approach will not have any additional learnable parameters and encourage the float vectors to align with the binary vectors. Please comment.

5) Is there any loss due to the introduction of the bi-rotation formulation?

6) What is the justification for Eq. 11?

7) Any overhead at test time?

[a] Ajanthan, T., Gupta, K., Torr, P.H., Hartley, R. and Dokania, P.K., 2019. Mirror descent view for neural network quantization. arXiv preprint arXiv:1910.08237.

Post rebuttal update:
The rebuttal clarifies most of my initial concerns and authors have promised to add discussions in certain parts. The main one, improvement due to rotation on xnor-net (without gradient approximation) is interesting and I recommend including that in the paper. I increase the rating.

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?

The method and the claims are correct and the experimental setup is valid.

5. Clarity: Is the paper well written?

Overall the paper is well written and sufficient details are provided. In addtion, please consider giving more details on the number of additional learnable parameters. See weakness.

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?

Prior works are sufficiently discussed.

7. Reproducibility: Are there enough details to reproduce the major results of this work?

Yes

8. Additional feedback, comments, suggestions for improvement and questions for the authors:

Please release the code to ensure reproducibility.

9. Please provide an "overall score" for this submission.

7: A good submission; accept.

10. Please provide a "confidence score" for your assessment of this submission.

4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?

Yes

Reviewer #3

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.
The authors focus propose a method for training binary neural networks. They focus on an overlooked issue in the binarization process, namely the rotation bias, which degrades performance and propose a solution to it. They also propose a smooth gradient approximation function

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.
The paper is innovative in that it identifies something that no one has looked into before.
The methodology is interesting. The insight is hard to realize, but the authors propose several interesting technical tricks to solve it.
There is a good awareness of the literature.

3. Weaknesses: Explain the limitations of this work along the same axes as above.
Context is sorely missing in the early parts of the paper. Intuition of what the paper is focusing on is very low in the introduction.

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?
It seems correct to me, although a more intuitive explanation in the introduction would have helped. The validation is correct.

A doubt I am left with is what happens with the rotation - you cannot just rotate the weights without using the inverse rotation elsewhere, otherwise the network will give totally wrong predictions. So how do you compensate for rotating the weights? does it have any implication in terms of computational cost? because if you have to rotate the input tensor at every step to compensate for the weight rotation, then it is non-trivial. Please clarify in the rebuttal this point - or what I am misunderstanding.

Two small things:
- line 37: the scaling factor is on a per-channel basis, which is not clear in the text
- line 86: Bi-real net does not increase the weights
- line 104: "Besides, this..." I do not understand what the authors mean, even after reading the full text

5. Clarity: Is the paper well written?
The paper has some issues early on, where the context and an intuitive explanation of what the authors are talking about would have helped a lot. Once they focus on the methodological part it is easier to follow. For example:
- what the angular bias is is not explained at all. There is a graphical abstract (Fig 1.a) with very little information on it that I was unable to decode. Only after reading the rest of the paper I understood.
- Th "weight flips" is equally impossible to understand early on. l50: "possibility of flips achieves 50%". First, I would say it is "probability", not possibility. But what changes for weights to flip? are we talking during iterations of training? across different images? it seems to me that what actually flips is the activations, not the weights themselves? anyway, I'm just hypothesizing, in any case this really needs a clarification early on.
- The self-adjustment in eq. 11 has to be justified. Where does this formula come from? why does it work?

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?
In general, the work shows good awareness of the literature and often discusses relevant work.
Nit: Section 3.4, would be nice to include an explanation that several approximations do exist in the literature.

7. Reproducibility: Are there enough details to reproduce the major results of this work?
Yes

8. Additional feedback, comments, suggestions for improvement and questions for the authors:

Some suggestions going forward

Please make a clear introduction of the concepts you're going to talk about. The paper would be so much more readable with a clear explanation of what's angular bias and where it comes from, and about the "flipping of weights". The graphical abstract should be way more intuitive too - a 2D sketch for example could do?

I would have liked to see some formatting of the validation like (I know the content is there, but parsing it is harder):
XNor-Net
XNor-Net + Ours
Bi-RealNet
Bi-RealNet + Ours
...

I think having a well thought-out optimization is particularly important to properly validate this method. It is mentioned by the authors that the angular bias could be corrected during optimization, but in practice it is shown that this is not the case. In my opinion, there has been a lot of improvement regarding optimization. Bi-Real net is a very good step forward, and it is great that it is included as a comparison. Yet, the current best baseline available is from "Training binary neural networks with real-to-binary convolutions", Martinez et al., ICLR'20 (what they call "strong baseline"), which has a large improvement over bi-real net just through refined optimization. In particular, I think that the two-stage training (proposed in another paper) is pivotal for a correct optimization. The question is, if you train the network "properly", will it correct the angular bias or still not manage?

Is it possible to train the full rotation matrix? it might be impractical, but if done at least once we would know how close the "practical approach" is to it.

Some of the notation used in tables (e.g. Table 4) is not explained in the caption. This means a reader has to go back and forth from table to text, which is annoying. It is always a nice touch to have "self-contained" tables if possible.

the Broader Impact is kind of an extended "conclusions" section. Maybe not the aim of the neurips organizers?

Rebuttal:
The rebuttal didn't add much from my perspective. Most of my remarks were not aimed to be tackled for the revision - the authors are indeed correct, unfortunately time and resources are limited. Also rewriting parts of the text is not possible for a rebuttal. Thus I simply leave the marks as they were. I would however encourage the authors to put some extra effort polishing the aspects I mentioned - especially making an effort to give better context and intuition early on - as it will make the paper more attractive and approachable.

9. Please provide an "overall score" for this submission.
6: Marginally above the acceptance threshold.

10. Please provide a "confidence score" for your assessment of this submission.
5: You are absolutely certain about your assessment. You are very familiar with the related work.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?
No

Reviewer #4

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.
This paper explores the influence of angular bias on the quantization error. A Rotated Binary Neural Network (RBNN) is introduced to reduce the angular bias. Compared with previous work, the proposed method can further reduce the quantization error theoretically considering the angular bias. Experimental results on CIFAR-10 and ImageNet using different architectures demonstrate the effectiveness.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.
1. The main idea is novel and reasonable. The previous work adopt scaling factors to lessen the quantization error. I agree that scaling factor can only partly mitigates the problem and it is a right way to solve this problem by reducing the angular bias.

2. The bi-rotation formulation is interesting. It is a clever approach to leverage the property of Kronecker product to reduce computation complexity. I appreciate this technique.

3. Experimental results on CIFAR-10 dataset and ImageNet are promising. The proposed method can achieve state-of-the-art performance on various architectures. The ablation studies verify the effectiveness of different modules.

3. Weaknesses: Explain the limitations of this work along the same axes as above.
1. My main concern is about self-adjust parameter \alpha. In my opinion, the ideal value for \alpha is 1, i.e. \tilde{w} is equal to the rotated weights. However, during training, \alpha varies from 0 to 1 and I can not understand its physical meaning. For example, when \alpha equals to 0.5, does the quantization error of \tilde{w} is smaller than that of original weight w ? I think authors should give some proof. Also, the figure of how \alpha update during training is not given. Does \alpha equal to 1 at the last epoch ?

2. There is a typo: the term “G” in table 4 is not consistent with the term “T” in line 217.

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?
The proposed method is correct and reasonable.

5. Clarity: Is the paper well written?
The paper is well writen and easy to understand.

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?
The author claim that they are the first to explore the influence of angular bias on quantization error in the field of BNN, which is the main difference from previous works.

7. Reproducibility: Are there enough details to reproduce the major results of this work?
Yes

9. Please provide an "overall score" for this submission.
8: Top 50% of accepted NeurIPS papers. A very good submission; a clear accept.

10. Please provide a "confidence score" for your assessment of this submission.
3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?

No