Skoltech
Skolkovo Institute of Science and Technology

Thesis Changes Log

Name of Candidate: Vadim Lebedev

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Algorithms for speeding up convolutional neural networks

Supervisor: Prof. Victor Lempitsky

Chair of PhD defense Jury: Prof. Andrzej Cichocki ***Email**:* a.cichocki@skoltech.ru

Date of Thesis Defense: 30.10.2018

The thesis document includes the following changes in answer to the external review process.

Multiple reviewers have noted that the thesis should be written in plural (we vs. I). I have changed to plural (we, our) throughout the text of the thesis.

Multiple reviewers pointed out inconsistency in the beginning of the section 3.1, where the approaches [Jaderberg et al., 2014b] and [Denton et al, 2014] are used in the text without proper introduction and explanation of details. I have added a new section 3.1.1, containing description of these works, to address this issue.

Reviewer: Anh-Huy Phan

1. The title “Fine-tuned CP-decomposition” of Chapter 3 is not consistent with its content. The process is to apply the low-rank CPD to the weight tensors in the convolutional layers and fine-tune the other layers of the network.
   Answer:
   I agree that finetuning is not a defining part of the method. The title is changed to "CP-decomposition of convolutional weights"
2. [Denton et al, 2014] decomposed the kernel tensor into rank-1 tensor composed by three factor componentswhile the proposed method decomposes the kernel tensor into four factor matrices.
   Answer:
   The correct description of method [Denton et al, 2014] is now provided on page 45. The complexity analysis is also corrected on page 49.
3. Reference of the CPD should be corrected. [Kolda and Bader, 2009] is an excelent review paper but not the appropriate papers introduced CPD.
   Answer:
   I have changed the reference to [Hitchcock, 1927] (also mentioning the history of rediscoveries an citing [Kolda and Bader, 2009])
4. The NLS algorithm in the Tensorlab minimises the Frobenius norm of the residual tensor, not the L2 norm. Moreover, it implements the conjugate gradient algorithm. For the Gauss-Newton algorithm for CPD, the reference can be “Low Complexity Damped Gauss–Newton Algorithms for CANDECOMP/PARAFAC”, SIMAX, 2013
   Answer:

The mistake is corrected, and the reference is added (page 45)

5. Size of the weight tensors should be mentioned. For example, the weight tensors in the layers 2 and 3 of CharNet are of size $9 \times 9 \times 48 \times 128$ and $8 \times 8 \times 64 \times 512$, respectively. However, for theAlexNet, dimensions of the weight tensor are not given.
   Answer:
   Information on AlexNet weight sizes is added
6. For the CharNet, Table 3.1 shows that a tensor decomposition with rank R = 64 using NLS yielded an accuracy drop of 0.09%, but in Fig. 3.2-(2-left), the drop was about 0.2%.
   Answer:
   I have changed the plot (Fig 3.2.) for consistency.
7. For the decomposition with rank R = 256, the accuracy drop was negative, -0.31 and -0.52. The author should discuss this result.
   Answer:
   Paragraph added on page 50.
8. It is not clear that the performance for the Alexnet shown in Table 3.1 was obtained with or with- out fine-tuning. In addition, it is also not clear that the greedy method compared in Table 3.1 reflected the algorithm by Denton et al, 2014, or the order-4 tensor decomposition.
   Answer:
   The clarification is added in the description.
9. On page 56, the kernel tensor W is matricized to a filter matrix F of size $N \times d\ 2\ C$. Different from matricization, reshaping keeps the order of elements of W in their vectorization.
   Answer:
   I changed an inaccurate reference to tensor reshape into the description of the transformation in question. (page 57 in the final version)
10. On page 55, “it’s implication to the sparsity structure are discussed”
    Answer:
    Corrected
11. Through the whole thesis, “3-D tensor”, “4-D tensor” should be corrected to order-3 or order-4 tensor.
    Answer:
    I have changed “3D tensor” to “3D arrays”

Reviewer: Pavlo Molchanov

1. Explaining details behind parameter rescaling of group regularization will improve the presentation of the method
   Answer:
   Description of rescaling is added on page 62.
2. The motivation in the case of smartphones should use privacy and low capability arguments […] Power is a huge concern in autonomous driving.
   Answer:
   I have edited the motivation parts for style and clarity and added arguments.
3. Page 6, ?? sign
   Corrected.
4. Page 20, Figure 2.2 caption: “MobilNet” -> “MobileNet”
   Corrected
5. BinaryConnect. It would be worth to explain how the probability of Bernoulli sampling is learned.
   Answer
   Details on weight sampling in evalution time added
6. Motivation behind introducing Section 2.8 is not clear. There are might be many other applications such as image generation, NLP, translation, audio synthesis, segmentation.

Why object detection was selected for more detailed overview is not clear.
Answer:
I was deeply impressed by the progress in the area of object detection (concerning evaluation speed of the models), so I decided it is an appropriate example, partly because of personal preference. Image generation is also mentioned. I have added the clarification in the beginning of the section.

7. Page 54: “speed-up o convolutional layers”:
   corrected
8. Section 5.2. On page 77 has an unfinished section “Separability”.
   The corrupted section is now removed
9. Term “bi-clustering” appears first time on the page 47. Reader might be not familiar with this term and having a description in related work section, or in this chapter will make it easier to follow the text.
   Answer:
   I have added the description of biclustering on the page 45

Reviewer: Stamatios Lefkimmiatis

Author should elaborate more on the CP decomposition using the Greedy versus the Non-linear squares method.
Answer:
New section on page 45 addresses the topic.

Reviewer: Stefan Roth

The following additional changes were made on the advice of Prof. Roth:

1. Equation 2.1: Wrong index corrected (s → c)
2. page26(page 27 in the final version): “recoup” changed to “compensate”.
3. Figure 2.4: Are the random matrices representative?
   Answer:
   The purpose of the plot is to demonstrate the overhead of sparse matrix multiplication. Since the sparsity structure is not introduced at this point in the text, I have no better option then to use random data for the experiment.
4. Page36(37): Incorrect singular number referring to the authors of the paper [Bucila et al., 2006] is fixed.
5. Typo in Equation 2.20 is fixed ($q^T \rightarrow p^T$)
6. page41(42): spelling corrected: spacial → spatial
7. page41(42): 'virtually all decompositions' changed to 'many decompositions'
8. page54(55) Typo is fixed: then → when
9. page55(56) The shift is removed from Equation 4.1
10. Notation in Figure 4.1 is made consistent with the rest of the thesis. The description is also updated
11. page56(57): error in notation is corrected (t → k)
12. page57(58) Equation number is added to the equation 4.2
13. page58(59): The description of Figure 4.2 is edited to provide more clarity and reference to the non-existent error bars is removed.
14. page63(64): Inconsistent marking in the table 4.1 is fixed (8.33->8.33x)
15. Page70(72): “much fewer” -> “many fewer”
16. page72(74): The typo is corrected. (images to → images are resized to)
17. page74(75) Architecture names in the Table 5.1 are properly capitalized
18. page75(77): “We have performed preliminary experiments on this Fungi 2018 [...]” The sentence is reformulated.
19. page77(79) Corrupted reference to Figure 5.3 is corrected.
20. page77(79) The corrupted section on separability is removed in the final version.

21. page86(88) Contradicting claims on the affordability of training CNNs in the future are removed.

I wish to thank all the reviewers for a valuable feedback and I am looking forward to the discussion of the provided comments.