

Jury Member Report – Doctor of Philosophy thesis.

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: October 30, 2018

Name of the Reviewer:

I confirm the absence of any conflict of interest	Signature:
(Alternatively, Reviewer can formulate a possible conflict)	PM 
	Pavlo Molchanov
	Date: 30-09-2018

The purpose of this report is to obtain an independent review from the members of PhD defense Jury before the thesis defense. The members of PhD defense Jury are asked to submit signed copy of the report at least 30 days prior the thesis defense. The Reviewers are asked to bring a copy of the completed report to the thesis defense and to discuss the contents of each report with each other before the thesis defense.

If the reviewers have any queries about the thesis which they wish to raise in advance, please contact the Chair of the Jury.

Reviewer's Report

Review of the PhD thesis titled as "Algorithms for speeding up convolutional neural networks" by Vadim Lebedev.

Background and objectives

Deep neural networks dominate most of CV tasks related to functional mapping, including object recognition, pose estimation, segmentation. Such applications will benefit from improved inference speed and will be able to reduce costs and increase throughput.

Accelerating neural networks is important problem for machine learning community and has been studied since 1990. The goal of accelerating neural network is to reduce computational costs of the forward pass while remaining at the same or slightly lower accuracy level (within tolerable range). Considering continuous progression of neural network architectures, variability in tasks and cost functions there is no single approach previously proposed that will fit all of them. Different types of

hardware can benefit from architecture-aware accelerations therefore research in direction of developing new algorithms is promising and will benefit the community.

Thesis is focused on the problem of accelerating deep neural networks, particularly their inference time at deployment stage. Candidate proposed 3 approaches to deal with problem of speeding-up inference of neural networks. The first approach is based on tensor decomposition of the convolutional filter into 4 convolutions with smaller number of computations. The second approach is based on structural pruning that is performed together with structural regularization during training. The third approach uses a small back-bone network trained together with non-parametric classifier such as RBF to perform fine-grained classification under limited computational constraints. All three approaches advance research in corresponding areas, they sound technical and have sufficient academic novelty. First two approaches were published in peer reviewed high ranking international conferences (ICLR2015 and CVPR 2016) emphasizing contribution to the field. Author of the thesis is the main author of those papers.

Structure of the thesis and the contributions

Chapter 1 provides an overview of generalized convolution operator and motivation behind emerging area of their acceleration. Study of elapsed time per different layers of the neural network provides a clear background on why convolutional layers require more attention towards acceleration compared to other layers. Implementation details of how convolutions are executed from algorithmic point of view help to understand potential ways of acceleration. Finally, provided results of a tradeoff between network accuracy and inference speed for different architectures motivates to speed up architectures with high computational cost.

Chapter 2 revises related work for model acceleration and compression. In the section 6 main groups of algorithms are considered in depth with relevant references and analysis. Figure 2.2 compares different popular NN architectures considering their run time and model sizes. Although images are provided with results it would be worst to explain why some of the ideas do not result in faster inference on CPU or GPU. Additionally, explanation why reducing precision of weights and activations to fewer bits does not result in immediate speed-up in existing frameworks will be beneficial. This Chapter concludes with observation that none of the existing acceleration techniques have reached saturation point and have drawbacks associated with them. Improving methods for acceleration is still possible and potentially a combination of several approaches from different domains can lead to better results.

Chapter 3 proposes a new acceleration technique based on rank-R CP-decomposition and an efficient optimization approach for finding basis functions and the rank value. Main advantages of the technique include: a) simple implementation in existing frameworks as no new operations are required to be implemented, and b) efficiency due to possible fine tuning of the decomposed convolutions. Method is presented well with analysis of computation reduction as well as parameter reduction. Proposed method reaches superior performance when compared on relatively small models and outperforms state of the art in this area. Results on bigger network show a gap with existing methods. Although, most of details are presented it is not clear of how NLS optimization works and more details behind the algorithm will better explain the method, as significantly better results are observed versus greedy optimization.

Tensor decomposition method suffers with deeper networks and/or have small convolutional kernels. A new method of structural pruning is proposed in Chapter 4 to avoid these shortcomings. The new method is based on applying group sparsity regularization term on structurally combined parameters of the network removing which will result in inference speedup. Explored structural sparsity helps to execute convolution operator efficiently by eliminating groups that will be accounted in matrix-matrix

multiplication. Results of pruning small networks show superior results compared to l1 regularization and better pruning rates than for decomposition techniques. Several modifications are suggested to improve original method including gradual group-wise sparsification which demonstrates the best results and eliminates the need in search of meta-parameters. The approach demonstrates great pruning rates for networks trained on large scale datasets and even allows simultaneous pruning of multiple layers. Explaining details of the algorithm behind parameter rescaling of group regularization for layers will improve presentation of the method.

Chapter 5 introduces the third method proposed in the thesis called impostor networks. It starts with introduction of the main concept being a neural network that maps input image into d-dimensional space and non-parametric classifier (RBF) that is linked with impostor set. Three different types of impostor network are introduced and studied including a version where the anchor set (impostor set) is learner and named as "loose" that overcomes issues with the tied impostors. Evaluation of the proposed method is well written and results clearly demonstrate effectiveness towards fine-grained image classification. Additional evaluations with state of the art demonstrate several orders of magnitude speed up over other methods.

Finally, Chapter 6 summarizes main contributions of the work and discusses shortcomings and advantages of proposed methods.

Presentation of the results

The problem formulation, related work, previous attempts to solve the problem are well described and compared with sufficient details. Reader can clearly understand on how novel contributions presented in the thesis stand with respect to previous state of the art. Most important results and methods are explained in the thesis, also results are compared on relevant data making it easy to access importance of the work. Experimental rather than theoretical evaluation of speed ups on most dominant hardware used for neural networks(CPU and GPU) is a strong verification of effectiveness shown by proposed methods. The thesis is well-written and is very easy to follow. The thesis provides references and discussions about most of the relevant literature in the field.

Conclusions

In summary, the considered problem is important to the field of machine learning and its applications. Several solutions proposed in the dissertation clearly advance the research towards a better solution. Based on the above examination of the manuscript of Vadim Lebedev, I recommend that the candidate should defend the thesis by means of a formal thesis defense.

Provisional Recommendation

☒ *I recommend that the candidate should defend the thesis by means of a formal thesis defense*

☐ *I recommend that the candidate should defend the thesis by means of a formal thesis defense only after appropriate changes would be introduced in candidate's thesis according to the recommendations of the present report*

☐ *The thesis is not acceptable and I recommend that the candidate be exempt from the formal thesis defense*