Skoltech

Skolkovo Institute of Science and Technology

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 (Alternatively, Reviewer can formulate a possible conflict)	**Signature:** ACichocki **Date: 14-09-2018**

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

The PhD thesis of **Vadim Lebedev,** entitled "**Algorithms for speeding up convolutional neural networks**", is devoted to the development of new algorithms that speed up evaluation of convolutional neural networks (CNNs). The thesis consists of 6 chapters: Introduction, literature review, three chapters describing specific new methods for speeding CNNs, and conclusion, which discusses applicability of proposed methods and outlines possible directions of the future works.

The modern CNNs have large computational costs, compared to the other machine learning methods used before in computer vision. The author of the thesis focus on challenging problem: How can we speed up the evaluation of CNNs, with minimal loss of performance metrics (classification accuracy).

High computation costs of CNNs occur for many practical applications, particularly when the computational resources are limited (e.g., laptops or mobiles), especially when the quick responses are require or when the very large amount of data have to be processed. Due to the importance of these areas, the research on the acceleration of CNNs is highly demanded a lot of research is already devoted to alleviate this problem, including low-rank tensor decompositions, fast architecture design, automatic architecture search, quantization and binarization, pruning, teacher-student approaches, and adaptive architectures.

The Main contributions of the thesis can be summarized as follows :

- Development of novel approach for speeding up convolutional layers of CNNs with CP?PARAFAC -decomposition of convolutional weights (chapter 3)

- Develop and test a novel pruning process that speeds up CNNs (chapter 4)

- Test the developed methods on image classification datasets (MNIST, ILSVRC2012, Caltech Birds 200), in the chapter 4 the image retrieval tests are also presented (INRIA Holidays, Oxford Buildings).

The author have used existing deep learning frameworks (Caffe and Pytorch) for the extensive experiments presented in the thesis and performed some modifications, using C++ and Python.

The author of the thesis has published quite impressive results in two top interntional conferences and the both papers are very highly cited (the first paper has, according to Google Scholar , more than 150 citations, which is really very impressive).

- **Vadim Lebedev**, Yaroslav Ganin, Maksim Rakhuba, Ivan Oseledets, and Victor Lempitsky. Speeding-up convolutional neural networks using fine-tuned CP-decomposition. International Conference on Learning Representations (ICLR), full conference paper, 2015.
- **Vadim Lebedev** and Victor Lempitsky. Fast ConvNets using group-wise brain damage. Computer Vision and Pattern Recognition (CVPR), 2016.

The results of the third chapter was published so far on arXiv, but publication in a conference is planned

- **Vadim Lebedev**, Artem Babenko and Victor Lempitsky. Impostor Networks for Fast Fine-Grained Recognition. Arxiv preprint, 2018.

The review /tutorial is also published in a peer-reviewed journal:

- **Vadim Lebedev** and Victor Lempitsky. Speeding-up Convolutional Neural Networks: A Survey. Bulletin of the Polish Academy of Sciences: Technical Sciences, 2018.

In my opinion publication are very good and convinced.

May be minor weak point of the thesis is that the developed of three different approaches h does not converge to a single point that is can not unified in one system. Another weak point is that author does not consider deeply other alternative and promising approaches like binarization and automatic architecture search, which are currently exploding in popularity and have a potential to completely change the landscape of deep learning, in general, especially speeding up CNNs in particular. Finally, there are some problems with consistency in the experiments presented in the thesis, as different frameworks and various datasets are used in different chapters.

However these are minor critical comments and I evaluate this PhD thesis very positive and high.

Provisional Recommendation

x□ *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*