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: Stamatios Lefkimmiatis

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**Signature:** S. Lefkimmiatis **Date: 15-09-2018**

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

Reviewers report should contain the following items:

- Brief evaluation of the thesis quality and overall structure of the dissertation.
- The relevance of the topic of dissertation work to its actual content
- The relevance of the methods used in the dissertation
- The scientific significance of the results obtained and their compliance with the international level and current state of the art
- The relevance of the obtained results to applications (if applicable)
- The quality of publications

The summary of issues to be addressed before/during the thesis defense

The focus of this thesis is on speeding up CNNs so that they can be deployed on devices that have memory storage or computation power constraints. This is a topic of great research interest since current deep learning approaches, while they lead to impressive performance in several computer vision tasks, are memory consuming and power hungry. This hinders their applicability to several domains, such as mobile applications.

The overall structure of this thesis is as follows:

In Chapter 2 a thorough overview of the field of network compression and network inference speed-up is provided. In particular the author covers in detail the research efforts that have been made in the following directions: 1) Tensor decompositions of the defining filter tensor, 2) Architecture designs that enable speed-up of the inference phase without a significant drop in the accuracy, 3) Automatic Neural Search and novel techniques for accelerating this search, 4) Quantization methods for the weights of the network, 5) Pruning approaches for reducing the complexity of the networks and 6) Teacher-student approaches where the underlying idea is to enable transfer knowledge between a deep network to a smaller model.

Chapter 3 focuses on speeding-up the execution time of CNNs during inference by approximating the convolution weights using the CP polyadic decomposition. The author assesses the performance of this approach by considering two networks, namely the CharNet (for character recognition) and AlexNet (for image classification). While the exposed idea is clear I believe that some technical details are missing. As a result the chapter is not self-contained which makes it difficult to follow by someone who is not expert in Tensor decompositions. For example, the author should elaborate more on the CP decomposition using the Greedy versus the Non-linear squares method.

In Chapter 4 the author describes a different approach to reduce the number of weights in a neural network through pruning and is motivated by the observation that current implementations of generalized convolutions compute generalized convolutions by reducing them to matrix multiplications. This opens the possibility of eliminating certain rows and columns from both factor matrices (according to selected criteria) so that the matrices will become thinner and the matrix multiplication becomes faster.

In Chapter 5 a different approach for running image classification in devices with limited storage and computational power is pursued. Specifically the author considers the design of a new architecture that involves a non-parametric classifier (RBF classifier). The advantage of this approach is that such classifiers do not require deep CNNs, which are computation and power hungry, to perform image classification and at the same time they do not compromise the recognition accuracy. This way the depth of the convolution layers of the impostor networks can be significantly reduced, leading to significant execution speedups.

Overall, the work in this thesis is of high quality and the author has done a very good job in describing in a clear way the current challenges in deploying modern deep learning networks in devices that are on budget either in memory storage or computational power. The author has pursued several different strategies to deal with some of these challenges, which has led to the development of novel techniques. Some of the results of this thesis have been published in top-tier international peer-reviewed conferences (ICLR and CVPR), which further supports the conclusion that this is a thesis of high quality.

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*