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: Prof. Stefan Roth, Ph.D.

I confirm the absence of any conflict of interest	**Signature:** Stefan Roth **Date: 12-10-2018**

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 presented dissertation *Algorithms for speeding up convolutional neural networks* of *Vadim Lebedev* is concerned with making deep neural networks more efficient at inference / test time in order to make them applicable also on devices where computational resources are limited. Examples for such settings include mobile phones or lightweight autonomous vehicles. This problem area is both important from an applications perspective as well as scientifically challenging.

To that end, three different approaches are proposed here. The first approach uses *tensor decomposition*, specifically CP decomposition, to make standard convolutional layers more efficient and derives a low-rank approximation of the 4D weight tensor. The computational benefit stems from factoring the 4D weight tensor into multiple combinations of 1D weight vectors, which are much cheaper to apply. Experiments show that this leads to a clear speed-up, especially when the filters are large (i.e. bigger than 3x3). Since modern neural networks often employ only small filters (typically 3x3), the second approach investigates an alternative approach, extending classic work on "brain damage" in neural networks such that the resulting sparsity pattern of the weights can be exploited computationally. The resulting *"group brain damage"* is very effective in reducing the computational overhead by automatically uncovering sparsity patterns in the filters than can be used to reduce the cost of convolutions. This not only leads to significant acceleration of the trained network with only little to no accuracy loss after fine-tuning, but also some interesting insights into which filter shapes are most important for high discriminative power. The third approach addresses the fine-grained classification task and proposes a hybrid CNN-kernel approach, termed *impostor network*, where a CNN is first used to provide a nonlinear embedding. Next, classification is performed by a non-parametric Gaussian-RBF classifier using impostors, which represent the training examples in the embedding space. The embedding and the impostor coordinates are learned jointly, which leads to a significant boost, especially of lightweight networks, in terms of their classification accuracy while incurring only an insignificant computational and memory overhead.

Each of these approaches forms a main chapter in the dissertation, which is completed using an introduction, which motivates the scenario, an extensive (and very well readable) review and taxonomy of related work, as well as a discussion section. The dissertation is well written and clear, with experiments carried out on standard networks employing widely established benchmark datasets.

The achieved results clearly extend the state of the art in the area of computationally efficient deep networks, outperforming leading prior approaches at the time of publication. This is also evidenced by the accompanying publications. The publications underlying the first two main chapters have been published in leading international conferences (CVPR and ICLR), which are very competitive with low acceptance rates. Both approaches have already enjoyed significant impact (currently 174 and 90 citations according to Google scholar), demonstrating the leading character of the contribution. The third main chapter is only published as a pre-print so far. Mr. Lebedev has further published two more papers in competitive international conferences (NIPS and ICML), but these are not included in the dissertation.

Overall, the present dissertation of Vadim Lebedev makes several significant contributions to the literature in computationally efficient neural networks, which have already influenced the scientific community in this area. Hence, I gladly recommend the acceptance of this thesis.

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*