

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

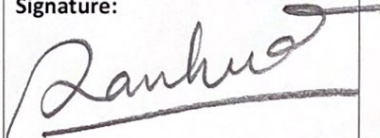
Supervisor: Prof. Victor Lempitsky

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)	
	Date: DD-MM-YYYY 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

Ph.D. candidate Vadim Lebedev investigated the problem of acceleration of Convolutional Neural Networks. More specifically, the doctoral thesis presents three proposed methods in Chapters 3, 4 and 5.

- The first algorithm presented in Chapter 3 is based on the Canonical Polyadic tensor decomposition (CPD) to represent the weight tensors in the convolutional layers by a smaller number of parameters, and thereby it reduces the computational complexity of the layer. The author applied the nonlinear least squares algorithm for CPD implemented in the Tensorlab toolbox. An important result in this chapter is that the fine-tuning can compensate the drop in the prediction accuracy due to the low rank approximation. A part of this chapter has been published in a conference paper [127] entitled "Speeding-up convolutional neural networks using fine-tuned CP-decomposition", ICLR (2015).
- The second algorithm presented in Chapter 4 relies on the group-sparse convolution. The filters are enforced group-sparsity by a regularizer. Similar to the low-rank approximation approach, while the sparsity patterns are fixed, the network can be fine-tuned to recover the original performance. This

acceleration method achieves the state-of-the-art performance. The result has been published in a conference paper "Fast ConvNets using group-wise brain damage". CVPE (2016).

- The third approach is based on the impostor network which has been recently presented in "Impostor Networks for Fast Fine-Grained Recognition", arXiv, 2018.

The Ph.D. thesis consists of six chapters and is well structured. The objective of each chapter is clearly outlined. The proposed algorithms are carefully implemented and verified for real-world data and with a comparison to commonly applied algorithms.

In my opinion, the thesis presents some novel contributions to the acceleration of CNNs. However, the author should explain the following remarks

- Regarding the acceleration method based on low-rank tensor approximation 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.

The author should present the related methods based on the low-rank tensor approximation, e.g., Denton et al, 2014 and Jaderberg et al. 2014b, in section 3.1.2, then highlight the difference between them and the proposed method. For example, Denton et al, 2014 decomposed the kernel tensor into rank-1 tensor composed by three factor components

$$W = \sum_r F_r \circ x_r \circ y_r$$

while the proposed method decomposes the kernel tensor into four factor matrices

$$W = \sum_r v_r \circ h_r \circ x_r \circ y_r$$

Reference of the CPD should be corrected. [Kolda and Bader, 2009] is an excellent review paper but not the appropriate paper for CPD.

The NLS algorithm in the Tensorlab minimizes the Frobenius norm of the residual tensor, not the ell-2 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.

Size of the weight tensors should be mentioned. For example, the weight tensors in the layers 2 and 3 of CharNet are of size 9×9×48×128 and 8×8×64×512, respectively. However, for the AlexNet, dimensions of the weight tensor are not given.

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%.

For the decomposition with rank R = 256, the accuracy drop was negative, -0.31 and -0.52. The author should discuss this result.

It is not clear that the performance for the Alexnet shown in Table 3.1 was obtained with or without 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.

- Regarding the group-spare convolution method

Some minor errors should be corrected. On page 56, the kernel tensor W is matricized to a filter matrix F of size $N \times d^2C$. Different from matricization, reshaping keeps the order of elements of W in their vectorization.

On the same page, "the t -th row corresponds to a sequence of C 2D filters $W(:, :, c, k)$...". The row index t should be k .

On page 55, "it's implication to the sparsity structure are discussed" \rightarrow ...

Through the whole thesis, "3-D tensor", "4-D tensor" should be corrected to order-3 or order-4 tensor.

- Regarding the impostor network

Similar to the presentation in Chapter 3, the author should first introduce the related work of Meyer et al. 2017, "Nearest Neighbour Radial Basis Function Solvers for Deep Neural Networks", which inspires the impostor net. The novelty of the proposed network is then explained.

The notation is not always consistent. For example, in (5.4), the author writes the mapping for the convolutional layer, $f_\theta(x_i, c_j)$, with two inputs, an image x_i and a reference point c_j , while its definition in (5.1) is $y = f_\theta(x)$. The error is also in (5.6).

My question is that if the imposter nets can be considered as a method to speed-up the CNN.

Finally, the author should use "we" in place of "I", "our method" instead of "my method", "proposed methods" for "novel methods". The references should follow the same style and format.

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