Skoltech
Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Talgat Daulbaev

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Applications of differential equations and reduced-order modeling for deep learning

Supervisor: Professor Ivan Oseledets
Co-supervisor: Professor Andrzej Cichocki

Name of the Reviewer: Rafael Ballester Ripoll

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**Date: 06-03-2023**

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

This dissertation explores three main themes and various connections between them: neural ODEs (chapters 2, 3, 4), adversarial attacks (chapters 4, 6) and model reduction using linear algebra techniques (chapters 5, 6). Rather than by publication date, the chapters have been ordered so as to make the transition between topics as smooth as possible.

Throughout the entire thesis, the downstream task is mostly supervised learning (classification) and some density estimation, for which standard architectures (convolutional, fully connected nets, sometimes autoencoders) undergo several modifications --mainly, replacement of ResNet by neural ODE blocks, variations on the normalization layers, and projection of the input/output of layers onto reduced subspaces that are computed via matrix factorization. All chapters (except chapter 5) include links to open-source code implementing the methods.

There is a fair amount of benchmarking where the proposed methods are compared against relevant state-of-the-art alternatives. Yet, there are some partial gaps in some evaluation sections that should be addressed (see my comments below).

Except chapter 3 which is a short benchmarking study, all chapters present novel methods: IRBM, MetaNODEs, RON, and ASNet, the quality of which is further attested by the level of their publication venues. While Talgat is often not the first author, his claimed contributions to each paper seem significant enough on their own. All in all, I am satisfied by the amount of technical depth of the thesis, which successfully and resourcefully combines methods and tricks from all the fields involved in the papers presented. I believe the candidate is ready to defend after the comments below have been taken into account.

Some minor comments:

- There are mentions to "supplementary materials", but the dissertation has no such materials. If possible, provide instead a URL to the supplementary materials of the respective papers.

- Pages "iii" and "iv" ("Publications"): add the publication years for the papers corresponding to chapters 6 and 4.

- Sometimes URLs are placed in the main text and sometimes as footnotes. It would be best to be consistent and always place them in footnotes.

- Figure in p. 16 lacks a number.

- There are a few typos: “contract” instead of “contrast”, “Lipchitz”, “IRMD” instead of “IRDM”, etc.

Next, for each chapter I list comments and concerns that should be addressed before the defense.

=== Chapter 1 ===

- NODEs play a major role in the thesis, and so it would be good to motivate them as much as possible early on. There are advantages of NODEs over ResNets that are only mentioned in section 3.1, and then again in 4.3.1 a study is cited that compares them favorably to CNNs. Such advantages should be mentioned in chapter 1 already.

- P. 13: instead of citing equation 2.1, cite 1.2 instead -- it is the same equation and it has been introduced already.

=== Chapter 2 ===

- Instability is mentioned as a shortcoming of RDM (p. 16) and as a motivation for the development of IRDM. However, the experiments do not demonstrate a significant gain in stability of IRDM over RDM, only some reduction of the loss, which is a different concept. Is there a definite example that conclusively shows that IRDM is more stable? I am thinking of, maybe, an example such as the one in the figure of p. 16. Stability is especially important since in 3.3 it is claimed that "ANODE is more robust than RDM", yet ANODE is not compared to IRDM in chapter 2.

- Related to the previous point: the main motivation of not using ANODE are the memory/time resources needed to keep the checkpoints. How would those requirements compare to IRDM for the models tested? I did not see this discussed.

=== Chapter 3 ===

- I miss some kind of take-away message here, as in what method is more effective when. In other words, some kind of useful recommendation to the reader on what normalization schemes and solvers are good "default options" in each situation --if such a recommendation can be made.

=== Chapter 4 ===

- There is no need to redefine the Runge-Kutta method and its order here (and with a slightly different notation) since it was already given earlier.

- I did not understand the motivation to go for gradient-free optimization of the RK parameter u (section 4.2, second paragraph). If u must be in (0, 1] and clamping to that interval is problematic, one could define u = 1/2 + arctan(x) / pi and optimize x. What am I missing?

- Please introduce the meaning of "robust accuracy". Without knowing that, it is hard to judge how significative the performance differences in tables 4.2, 4.3 and 4.4 are.

- Section 4.3.4 claims that “we consider how neural ODEs pre-trained in different regimes react to white-box attacks”, but I did not see that in the experiments --the table results only refer to black-box and gray-box.

=== Chapter 5 ===

- Broken sentence in p. 49: "urpose, we apply"

- In the second point of 5.3.2 ("Convolution" part), I did not understand what is meant by "can be compatible if the number of channels is not big". Similarly, I could not understand the "Batch normalization" part when it says that the normalization "can be merged" and the student can get rid of these layers. Could you rephrase/add more details?

- Figure 5.1: right now, one cannot tell which curves correspond to the deeper blocks. A sequential colormap would help. Also, the kind/name of the architecture is not mentioned here; is it one of the models from the repository https://github.com/SCUT-AILab/DCP/wiki/Model-Zoo mentioned two pages later?

- In 5.4.1, even if many singular values decay fast, rank truncation error will then propagate through subsequent layers, and it is not obvious that the final error will still be small. Section 5.4.1 and Fig. 5.1 would be more convincing if this were mentioned or countered.

=== Chapter 6 ===

- Fig. 6.5: when looking at singular value decay it is helpful to know the matrix size. For example, knowing that the singular values up to 200 capture 90% of a matrix's variance is more useful for compression if the matrix has 10^4 columns than if it has 250 columns only. What is the size of the layer matrices?

- A PCE degree of p = 2 is a bit underwhelming. Many applications of PCE for uncertainty quantification use a higher degree, e.g. 5 or even more, and are often combined with hyperbolic truncation to lower their memory/computational requirements (see for example "Data-driven polynomial chaos expansion for machine learning regression", by E. Torre et al.). Have higher degrees been tried?

=== Conclusions ===

It is claimed that the model reduction techniques from chapters 5 and 6 "allowed us to further explore the potential of neural ODEs". However, there is no mention to neural ODEs in these chapters.

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*