Bachelor Thesis Review

Faculty of Mathematics and Physics, Charles University

Thesis author	Andrei Lupasco
Thesis title	Deep Neural Networks for Graph Data Processing
Year submitted	2024
Study program	Computer Science
Specialization	Artificial intelligence
Review author	doc. RNDr. Iveta Mrázová, CSc. Advisor
Department	Department of Theoretical Computer Science and Mathematical Logic

Overall

	good	OK	poor	insufficient
Assignment difficulty	X	X		
Assignment fulfilled		X	X	
Total size *... text and code, overall workload*		X	X	

The objective of the thesis was to discuss neural network paradigms relevant to deep learning and graph data processing. Further, the student had to develop a viable strategy to represent molecule graphs, implement the models, and test their performance on real-world data. This objective has been met, albeit with certain objections to be mentioned in the next section.

As molecules that can be captured by planar graphs strongly prevail in the considered type of data, the author chose to employ the KHC algorithm to avoid inefficient isomorphism testing for general graphs. Further, he proposed an extension of this approach to handle disconnected graphs, too, and outlined the main principle of a new Planar Graph Neural Network (PGNN) architecture independent of the traditional message-passing paradigm.

For thesis defense, the author might consider possible extensions of his approach towards general graphs (although non-planar molecule cases are rare).

Thesis Text

	good	OK	poor	insufficient
Form *... language, typography, references*		X	X	
Structure *... context, goals, analysis, design, evaluation, level of detail*		X	X	
Problem analysis			X	
Developer documentation		X	X	
User Documentation		X		

The text would deserve a thorough review and correction of numerous typos, grammatical, and formal errors, e.g., on the last line of p. 8 (’can not’) or on l. 12 of p. 34 (’ar’), problems with graphical formatting on p. 65 or 73. In Figure 2.8, the nodes’ color labeling should be explained. In Figure 2.10 on p. 26, there should be rather the label $1, \{2\}$ for node D instead of the copy-pasted code $2, \{3, 1\}$ for node C in Iteration 1, etc. Although appropriate information sources were used for the thesis, the references [9], [10], and [11] differ just in page numbers; [12] and [14] are identical. Throughout the work, original sources should be cited, e.g., for MLP or CNN on p. 11.

Unfortunately, some stylistic mistakes in the text may lead to misinterpretation, e.g., the first sentence on p. 28. The review provided in Chapters 2 and 3.1 contains quite many inconsistencies and imprecisions like incorrect or missing indices, lack of notation and choice of initial values, etc. – see, e.g., missing h_0 on p. 11 and $x^{(0)}$ for recurrent networks; there should be rather $\tanh(C_{t+1})$ instead of $\tanh(C_t)$ on p. 16; inconsistency between the text and Figure 2.6. The symbol σ is inconsistently used for the transfer function and for standard deviation. The reason for using $\exp(\log \sigma)$ to determine Z on p. 23 should be explained.

When discussing the message-passing models, it should be better distinguished between the updates performed during training and during recall. The definition of the contractive mapping on p. 19 lacks the parameter value q. In the ELBO optimization criterion on p. 23, the log term is missing; during training, this criterion should be rather maximized. The function of the algorithms outlined in Section 3.1.2 to build the canonical codes should be better explained and verified. Figures 3.1, 3.2, 3.4, and 3.5, the text lacks an explanation of the coding procedure, the final canonical code, and its form presented in Figure 3.7. Further, the meaning of ’(B’ and ’B)’ is explained on p. 40, but these symbols were used already on p. 33.

On line 3 of p. 57, the author states: “We have chosen three different graph datasets” (for the experiments. Anyway, I have found only the reports on the datasets NCI109 and ZINC in the text. Further, it is not clear how the actually used ZINC subset (less than 5% of the original data) was chosen and what was the model’s performance on the remaining data (not used for training). The achieved results should also be compared to those of the related (E-)BasePlanE approach introduced in [32].

Thesis Code	good	OK	poor	insufficient
Design *... architecture, algorithms, data structures, used technologies*		X		
Implementation *... naming conventions, formatting, comments, testing*		X	X	
Stability		X		

To test the performance of the proposed PGNN, the student used his own SW package, GraphMindKeras, which he implemented in Python 3 using the Keras framework and the Sage library. Testing performed on two datasets from the TUDataset benchmark collection yielded results comparable to several state-of-the-art methods. On the other hand, testing the model’s performance on more datasets (possibly also from other domains) with different model architectures might provide a better insight into the overall function of the model.

Overall grade Very Good (rather worse)
Award level thesis No

August 30, 2024 Signature