THESIS REVIEWER'S REPORT

I. IDENTIFICATION DATA

Thesis title:	Extracting logic rules from neural networks with discrete weights
Author's name:	Armin Hadžić
Type of thesis :	master
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	CS
Thesis reviewer:	Gustav Šír
Reviewer's department:	CS

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment **challenging**

How demanding was the assigned project?

This is an interesting and challenging topic that required combining a range of diverse classic techniques.

Fulfilment of assignment **fulfilled**

How well does the thesis fulfil the assigned task? Have the primary goals been achieved? Which assigned tasks have been incompletely covered, and which parts of the thesis are overextended? Justify your answer.

Fulfilled in a very complete manner.

Methodology **correct**

Comment on the correctness of the approach and/or the solution methods.

The methodology seems very reasonable, the student provides a solid overview of the related areas, from rule learning to training quantized neural networks, and introduces a new NuLog rule extraction method that extends upon that foundation in a number of directions, which are described, analyzed, and experimented in solid detail.

Technical level **A - excellent.**

Is the thesis technically sound? How well did the student employ expertise in the field of his/her field of study? Does the student explain clearly what he/she has done?

I rate the technical level of the thesis very high for the sheer range of techniques covered, through custom tweaks introduced by the student, that clearly demonstrate his technical prowess in a number of areas explored during the studies, from neural networks, logic, and rule learning to integer linear programming.

Formal and language level, scope of thesis **C - good.**

Are formalisms and notations used properly? Is the thesis organized in a logical way? Is the thesis sufficiently extensive? Is the thesis well-presented? Is the language clear and understandable? Is the English satisfactory?

The thesis is sufficiently extensive, structured in a logical fashion, with nice typography and a reasonable use of formal notation (with only occasional typos). However, the language gluing the ideas together would benefit from some serious improvements – many sentences are broken and the wording is often rather unusual which, accompanied by some conceptual confusions, especially on the logical side (e.g. mismatching rule head<-body, term/predicate/literal), makes the thesis hard to comprehend at many places (it sometimes reads almost like a draft, as suggested by the footnote date).

Selection of sources, citation correctness **B - very good.**

Does the thesis make adequate reference to earlier work on the topic? Was the selection of sources adequate? Is the student's original work clearly distinguished from earlier work in the field? Do the bibliographic citations meet the standards?

While the related work is adequate and quite extensive (except for prior art on the very problem of rule extraction from NNs), I found the use of references throughout the thesis rather confusing, particularly when referencing parts of the thesis itself (e.g., "is shown here 0" etc.) – it would be largely impossible to follow when printed.

Additional commentary and evaluation (optional)
Comment on the overall quality of the thesis, its novelty and its impact on the field, its strengths and weaknesses, the utility of the solution that is presented, the theoretical/formal level, the student's skillfulness, etc.
See the overall evaluation.

III. OVERALL EVALUATION, QUESTIONS FOR THE PRESENTATION AND DEFENSE OF THE THESIS, SUGGESTED GRADE

Summarize your opinion on the thesis and explain your final grading. Pose questions that should be answered during the presentation and defense of the student's work.

The thesis presents a solid and innovative approach to extracting propositional rules from quantized neural networks. In the process, the student incorporates a wide range of techniques that allowed him to achieve some reasonable initial results in this challenging problem. The wording and textual flow of the thesis would benefit from some more (or first) rounds of editing, but this is outweighed by the amount of technical contribution.

Questions/opinions:

1) "Explanations provided are easy to interpret and convert to natural language…"
    a. It seems quite the opposite, which is why you need to use the rule visualizations?
    b. Actually, there isn't a single example of some logical/natural interpretation in the thesis
        i. Which is normally the main motivation for using logic in the first place…
2) "Interestingly, RIPPER shows rather poor accuracy in both training and testing data" [p.52]
    a. It actually seems second best to me?
    b. Moreover, the cohesive/symmetric regions (of the rule learners) look more interpretable to me
        i. Btw. these spatial/ordinal domains are classic showoffs for the (despised) fuzzy logic
3) If fidelity on test data is out of scope, why do you think the rules actually explain what the NN does?
    a. As opposed to some random rule learner trained black-box to produce the same training labels?
        i. Given that your extraction does quite some modifications and takes those into account…

The grade that I award for the thesis is **A - excellent.**

Date: **3.6.2024** Signature: