Supervisor's review of Filip Bartek's Doctoral Thesis

Machine Learning for Saturation-Based Theorem Proving

In my opinion, the submitted thesis fully satisfies the requirements for the Ph.D. degree. The findings reported therein are very significant and comprise a series of systems and experiments concerning the improvement of saturation-based automatic theorem provers (ATPs) with heuristics automatically discovered with the help of machine learning. Results are demonstrated using the award-winning ATP Vampire, which represents the state of the art in the field, but could be easily adapted to other saturation-based ATPs. The findings have been published in 5 peer-reviewed papers, 3 of those in established conferences in the field of Automated Reasoning: CADE 2021 (CORE A), LPAR 2023 (CORE A)[1], and IJCAR 2024 (CORE A), and two in the PAAR workshop (2020 and 2024).

The first paper (Learning Precedences from Simple Symbol Features, PAAR 2020) records a first attempt in the literature at using machine learning for proposing symbol precedences that work well for restricting the superposition calculus employed by an ATP and thus lead to manageable proof searches. It presents a system that learns pairwise symbol preferences from the performance of the Vampire theorem prover on problems with randomly sampled precedences. This system uses simple, human-engineered features to represent symbols and employs machine learning (specifically, linear regression with Lasso regularization) to learn a general preference model across different problems.

Building on experience from the first paper, the second paper (Neural Precedence Recommender, CADE 2021) uses a more powerful ML technology, namely a Graph Convolutional Neural Network (GCN) to propose symbol precedences. The system is designed to work in a signature agnostic way, so that it can generalize across problems coming from diverse domains. The GCN reads the input problem's clause normal form and produces symbol costs. By sorting problem symbols based on these costs to form the precedence, the prover can significantly reduce the search effort required to find a proof, outperforming a state-of-the-art heuristic by more than 4 % on unseen problems.

The third paper (How Much Should This Symbol Weigh? A GNN-Advised Clause Selection, LPAR 2023) explores automatically improving an ATP by learning to assign weights to symbols, thereby guiding the selection of clauses during the proof search process. It introduces a system using a GCN trained on successful proof attempts to predict effective symbol weights for clause selection in Vampire. The experiments demonstrate that the learned weighting scheme significantly enhances the prover's ability to solve first-order problems compared to uniform (by 6.6%) or goal-directed (by 2.1%) weighting strategies. The work also aims to explain what kind of knowledge the GCN learns, which, unlike with some other uses of ML in ATP guidance, is at least to a degree feasible thanks to the relative simplicity of how symbols weights influence the proof search.

---

[1]Only later in 2023, the ranking of LPAR has been updated: https://portal.core.edu.au/conf-ranks/1596/

The last two papers (Regularization in Spider-Style Strategy Discovery and Schedule Construction, IJCAR 2024, and Cautious Specialization of Strategy Schedules (Extended Abstract), PAAR 2024) explore methods for automatically discovering and scheduling diverse theorem-proving strategies, which is a well-known way to substantially improve the ATP performance in practice. Building on a large-scale experiment using a Spider-style approach to generate many strategies, a greedy algorithm for constructing schedules from these strategies is presented and its performance compared to an optimal cover method. Furthermore, the research investigates various regularization techniques applied to the greedy algorithm to improve the generalization of the resulting schedules on unseen problems. While the IJCAR paper focuses on monolithic schedules, the extended abstract presented at the PAAR workshop investigates how to further improve the ATP performance by creating branching strategy schedules that specialize based on problem features, while also addressing the crucial trade-off between training performance and generalization to unseen problems.

The thesis is framed by an introduction, which puts the contributions into a broader context, and short summaries of each of the papers. The thesis is well-structured, methodologically sound, and clearly presents an impact of machine learning in advancing ATPs. I strongly recommend the acceptance of Filip Bartek's thesis for the Ph.D. degree.

Prague, March 28th, 2025

Martin Suda, supervisor specialist