Reg. No: 2018/29/N/ST6/02903; Principal Investigator: mgr Bartosz Paweł Piotrowski

Deep Neural Architectures for Automated Theorem Proving

~Abstract for general public~

Bartosz Piotrowski

The formal foundations of modern mathematics were laid at the turn of the 19th and the 20th centuries. Work of people like Frege, Russel and Whitehead brought us the notion of the mathematical proof as a formal derivation in a logical calculus. The subsequent advent of computers gave rise to the automated theorem proving (ATP) field. ATP systems, implementing complete proof calculi guided by manually designed search heuristics, can be in principle used to attack any formally stated mathematical problem. Nevertheless, they remain typically much weaker than human mathematicians.

Recently, *data-driven, machine learning* (ML) techniques are being incorporated into the ATP systems. This makes same parts of the ATPs closer to the human way of doing mathematics. Humans often develop *intuition* for choosing the appropriate reasoning steps – which is *learned* from the experience of proving other theorems. It seems to be wasteful to not use the benefits of learning from the past successes and failures also in ATP systems. The growing number of experiments following this idea suggests this is a promising line of research. Large formal proof corpora has been recently translated into ATP formalism providing training data sets. Various ML models are being employed to learn from previous proof attempts. In particular, deep learning methods have been applied recently.

Deep learning (DL), or synonymously, *deep neural networks* (DNN) is a branch of ML that is currently under very active development. DNN architectures are responsible for a number of recent breakthroughs in a variety of applications. In particular, DNNs have been successfully applied to computer vision tasks, natural language processing, and many others. The applicability of DL in the field of symbolic reasoning has so far received relatively little attention, although the interest in the subject is increasing and several results from the last couple of years are very encouraging.

There are distinguishable properties of DNN models that seem to be particularly relevant in the context of symbolic reasoning. DNNs realize hierarchical learning from raw, simple representations of training examples and is capable of internally representing more complex concepts via combinations of simpler ones. For example, in the context of computer vision it is representing some figurative object, like elephant, as some particular combinations of simple geometrical shapes, like lines. It means, that DNN models exhibit, roughly speaking, an ability to process raw, *syntactic* representation, into an internal abstract form which is more *semantic*. In case of logical formulae – objects in symbolic reasoning we operate on – the distinction between the syntax and semantics is crucial. Small syntactical change can completely change the meaning of the formula – on the other hand, very differently looking expressions can have the same semantics. It is very desirable to have ML models able to properly handle this gap between syntax and semantics of logical formulae.

The objective of the project, motivated by the above remarks, is **in-depth, systematic study of applicability of deep neural network models in the domain of automated theorem proving**.

The plan of the project distinguishes between two main phases. First, a thorough investigation supported by experiments in more *elementary, isolated settings* will be carried out. Its aim is to understand what are the capabilities and limitations of various neural architectures with respect to learning semantics of formulae and various logical relations involving them, such as entailment, subsumption, equivalence. The second phase, using the expertise gained in the first phase, will be integrating DL methods into existing ATP systems. In particular, we will investigate how these methods can be applied to (i) the *premise selection* task (selecting relevant facts from a large formal library for proving a new conjecture) and (ii) to the *internal guidance* of the proof-search in ATPs.

The proposed line of research will benefit both the ML and ATP fields. In case of the former, the results which we plan to produce will enable better understanding of capabilities of DNNs – the subject of a great interest nowadays. In case of the latter one, we hope it will enhance existing ATP methods and popularize data-driven paradigm for them. Any improvements here naturally translate into many related domains, such as interactive theorem proving (ITP) and formal verification – and big formalization projects associated with them, either mathematical or industrial ones. In these projects there often appear many smaller, tedious proof obligations where ATP systems are productive tools to discharge them. Strengthening ATP makes larger ITP projects tractable, which in turn produces more training data (proofs) for ML. This forms a positive feedback loop.