Deep Network Guided Proof Search

Sarah Loos **Geoffrey Irving** **Christian Szegedy** **Cezary Kaliszyk**

LPAR 2017, Maun

May 8, 2017

Progress in ITP and ATP

Large Formalizations

- AFP: 64K lemmas, 593K LoC [*Nipkow+2015*]
- seL4: 49K lemmas, 400K LoC [*Klein+2014*]
- Flyspeck: 27K lemmas, 2B intermediate steps [*Hales+2016*]

Problems handled by ATPs

- Avatar [*Voronkow 2015*]
- E-prover history mining [*Schulz 2016*]
- SAT traces are big data

Little use of machine learning

Fast progress in machine learning

Tasks involving logical inference

- Natural language question answering [*Sukhbaatar+2015*]
- Knowledge base completion [*Socher+2013*]
- Automated translation [*Wu+2016*]

Games

AlphaGo problems similar to proving [*Silver+2016*]

- Node evaluation
- Policy decisions

Computer Vision

Better than human performance on some tasks [*Russakovsky+2015*]

Machine Learning in Theorem Proving so far

Predict Statement Dependencies

- Premise selection and relevance in ATPs
- Heuristics, learning and deep learning useful

Estimate Statement Usefulness

- Heuristics and simple learning methods

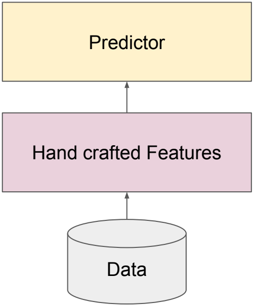
Propose Useful Conjectures

Supervised Learning Task

Assume $G : D \to P$	Ground truth G
$f : D \times M \to P$	Model architecture f
$\sigma : P \times P \to \mathbb{R}$	Prediction Metric σ
$S \subset D \times P$	Training Samples S

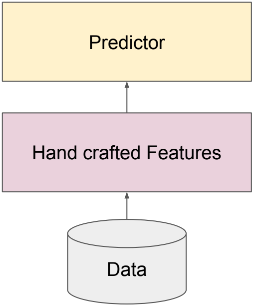
Find model parameters $m \in M$ such that the expected $\mathbb{E}(\sigma(f(d, m), G(d)))$ is minimized.

Deep Learning vs Shallow Learning

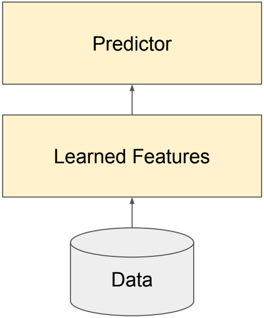


Traditional machine learning

Deep Learning vs Shallow Learning

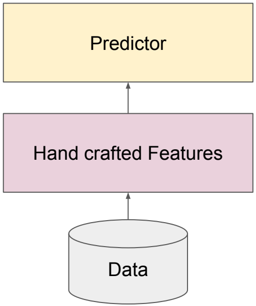


Traditional machine learning



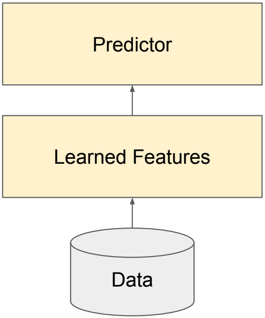
Deep Learning

Deep Learning vs Shallow Learning



Traditional machine learning

- Mostly convex, provably tractable
- Special purpose solvers
- Non-layered architectures



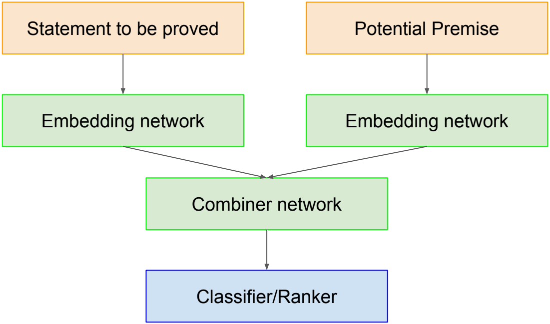
Deep Learning

- Mostly NP-Hard
- General purpose solvers
- Hierarchical models

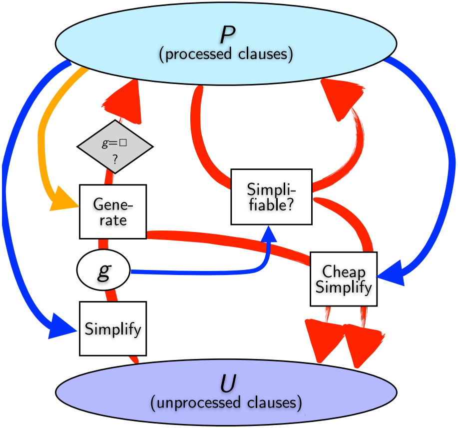
Deep Learning for Mizar Lemma Selection

[Alemi+2016]

- Embed all lemmas into $\mathbb{R}^n$ using an LSTM
- Embed conjecture into $\mathbb{R}^n$ using an LSTM
- Simple classifier on top of concatenated embeddings
- Trained to estimate usefulness on positive and negative examples



E-Prover given-clause loop



Most important choice: unprocessed clause selection [Schulz 2015]

Data Collection

Mizar top-level theorems

[*Urban 2006*]

- Encoded in FOF

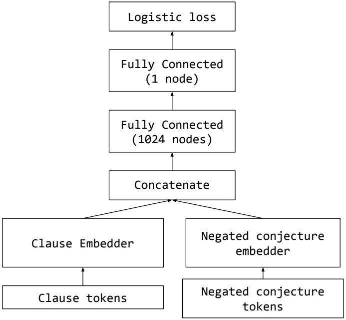
32,521 Mizar theorems with ≥ 1 proof

- training-validation split (90%-10%)
- replay with one strategy

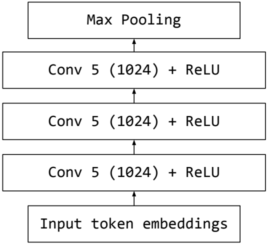
Collect all CNF intermediate steps

- and unprocessed clauses when proof is found

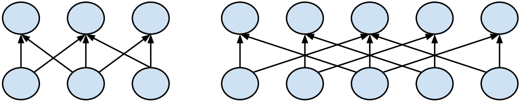
Deep Network Architectures



Overall network



Convolutional Embedding



Non-dilated and dilated convolutions

Recursive Neural Networks

- Curried representation of first-order statements
- Separate nodes for `apply`, `or`, `and`, `not`
- Layer weights learned jointly for the same formula
- Embeddings of symbols learned with rest of network
- Tree-RNN and Tree-LSTM models

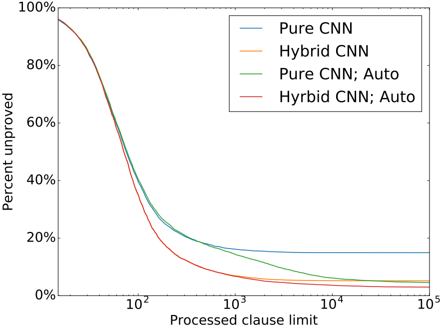
Model accuracy

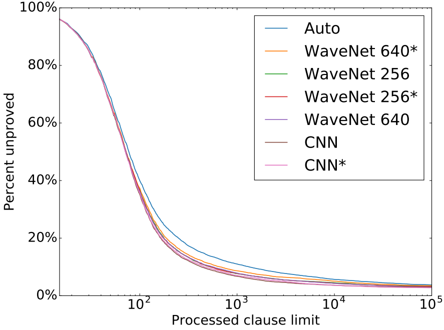
Model	Embedding Size	Accuracy on 50-50% split
Tree-RNN-256×2	256	77.5%
Tree-RNN-512×1	256	78.1%
Tree-LSTM-256×2	256	77.0%
Tree-LSTM-256×3	256	77.0%
Tree-LSTM-512×2	256	77.9%
CNN-1024×3	256	80.3%
⋆CNN-1024×3	256	78.7%
CNN-1024×3	512	79.7%
CNN-1024×3	1024	79.8%
WaveNet-256×3×7	256	79.9%
⋆WaveNet-256×3×7	256	79.9%
WaveNet-1024×3×7	1024	81.0%
WaveNet-640×3×7(20%)	640	**81.5**%
⋆WaveNet-640×3×7(20%)	640	79.9%

⋆ = train on unprocessed clauses as negative examples

Hybrid Heuristic

Already on proved statements performance requires modifications:





Harder Mizar top-level statements

Model	DeepMath 1	DeepMath 2	Union of 1 and 2
Auto	578	581	674
⋆WaveNet 640	644	612	767
⋆WaveNet 256	692	712	864
WaveNet 640	629	685	997
⋆CNN	905	812	1,057
CNN	839	935	1,101
Total (unique)	1,451	1,458	1,712

Overall proved 7.4% of the harder statements

Summary

Guiding superposition proof

- Deep network clause ranking

Performance

- Batching (evaluate clauses together)
- Hybrid heuristic
- Specialized hardware could help?

Deep network models

- Accuracy

References

A. A. Alemi, F. Chollet, G. Irving, N. Een, C. Szegedy, and J. Urban.
DeepMath-Deep sequence models for premise selection.
In *Advances in Neural Information Processing Systems*, pages 2235–2243, 2016.

G. Bancerek, C. Byliński, A. Grabowski, A. Korniłowicz, R. Matuszewski, A. Naumowicz, K. Pąk, and J. Urban.
Mizar: State-of-the-art and beyond.
In M. Kerber, J. Carette, C. Kaliszyk, F. Rabe, and V. Sorge, editors, *Intelligent Computer Mathematics - International Conference, CICM 2015, Washington, DC, USA, July 13-17, 2015, Proceedings*, volume 9150 of *LNCS*, pages 261–279. Springer, 2015.

J. C. Blanchette, C. Kaliszyk, L. C. Paulson, and J. Urban.
Hammering towards QED.
J. Formalized Reasoning, 9(1):101–148, 2016.

C. Kaliszyk and J. Urban.
FEMaLeCoP: Fairly efficient machine learning connection prover.
In M. Davis, A. Fehnker, A. McIver, and A. Voronkov, editors, *Logic for Programming, Artificial Intelligence, and Reasoning - 20th International Conference, LPAR-20 2015*, volume 9450 of *LNCS*, pages 88–96. Springer, 2015.

C. Kaliszyk and J. Urban.
MizAR 40 for Mizar 40.
J. Autom. Reasoning, 55(3):245–256, 2015.

C. Kaliszyk, J. Urban, and J. Vyskocil.
Efficient semantic features for automated reasoning over large theories.
In Q. Yang and M. Wooldridge, editors, *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 3084–3090. AAAI Press, 2015.

D. Whalen.
Holophrasm: a neural automated theorem prover for higher-order logic.
arXiv preprint arXiv:1608.02644, 2016.