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  \([\text{Nipkow+2015}]\)
- seL4: 49K lemmas, 400K LoC  \([\text{Klein+2014}]\)
- Flyspeck: 27K lemmas, 2B intermediate steps  \([\text{Hales+2016}]\)

Problems handled by ATPs

- Avatar  \([\text{Voronkow 2015}]\)
- E-prover history mining  \([\text{Schulz 2016}]\)
- 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 $G : D \rightarrow P$

$f : D \times M \rightarrow P$

$\sigma : P \times P \rightarrow \mathbb{R}$

$S \subset D \times P$

Ground truth $G$

Model architecture $f$

Prediction Metric $\sigma$

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

- Hand crafted Features
- Data
- Predictor

Deep Learning

- Learned Features
- Data
- Predictor

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

Mostly NP-hard
- General purpose solvers
- Hierarchical models
Deep Learning vs Shallow Learning

Traditional machine learning

- Hand crafted Features
- Data

Deep Learning

- Learned Features
- Data
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
- 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

- Encoded in FOF

32,521 Mizar theorems with $\geq 1$ 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

<table>
<thead>
<tr>
<th>Model</th>
<th>Embedding Size</th>
<th>Accuracy on 50-50% split</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree-RNN-256×2</td>
<td>256</td>
<td>77.5%</td>
</tr>
<tr>
<td>Tree-RNN-512×1</td>
<td>256</td>
<td>78.1%</td>
</tr>
<tr>
<td>Tree-LSTM-256×2</td>
<td>256</td>
<td>77.0%</td>
</tr>
<tr>
<td>Tree-LSTM-256×3</td>
<td>256</td>
<td>77.0%</td>
</tr>
<tr>
<td>Tree-LSTM-512×2</td>
<td>256</td>
<td>77.9%</td>
</tr>
<tr>
<td>CNN-1024×3</td>
<td>256</td>
<td>80.3%</td>
</tr>
<tr>
<td>⋆CNN-1024×3</td>
<td>256</td>
<td>78.7%</td>
</tr>
<tr>
<td>CNN-1024×3</td>
<td>512</td>
<td>79.7%</td>
</tr>
<tr>
<td>CNN-1024×3</td>
<td>1024</td>
<td>79.8%</td>
</tr>
<tr>
<td>WaveNet-256×3×7</td>
<td>256</td>
<td>79.9%</td>
</tr>
<tr>
<td>⋆WaveNet-256×3×7</td>
<td>256</td>
<td>79.9%</td>
</tr>
<tr>
<td>WaveNet-1024×3×7</td>
<td>1024</td>
<td>81.0%</td>
</tr>
<tr>
<td>WaveNet-640×3×7(20%)</td>
<td>640</td>
<td>81.5%</td>
</tr>
<tr>
<td>⋆WaveNet-640×3×7(20%)</td>
<td>640</td>
<td>79.9%</td>
</tr>
</tbody>
</table>

⋆ = train on unprocessed clauses as negative examples
Hybrid Heuristic

Already on proved statements performance requires modifications:

![](image1)

![](image2)
Harder Mizar top-level statements

<table>
<thead>
<tr>
<th>Model</th>
<th>DeepMath 1</th>
<th>DeepMath 2</th>
<th>Union of 1 and 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auto</td>
<td>578</td>
<td>581</td>
<td>674</td>
</tr>
<tr>
<td>✭ WaveNet 640</td>
<td>644</td>
<td>612</td>
<td>767</td>
</tr>
<tr>
<td>✭ WaveNet 256</td>
<td>692</td>
<td>712</td>
<td>864</td>
</tr>
<tr>
<td>WaveNet 640</td>
<td>629</td>
<td>685</td>
<td>997</td>
</tr>
<tr>
<td>✭ CNN</td>
<td>905</td>
<td>812</td>
<td>1,057</td>
</tr>
<tr>
<td>CNN</td>
<td>839</td>
<td>935</td>
<td>1,101</td>
</tr>
<tr>
<td>Total (unique)</td>
<td>1,451</td>
<td>1,458</td>
<td>1,712</td>
</tr>
</tbody>
</table>

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

Mizar: State-of-the-art and beyond.

Hammering towards QED.

C. Kaliszyk and J. Urban.
FEMaLeCoP: Fairly efficient machine learning connection prover.

C. Kaliszyk and J. Urban.
MizAR 40 for Mizar 40.

C. Kaliszyk, J. Urban, and J. Vyskocil.
Efficient semantic features for automated reasoning over large theories.

D. Whalen.
Holophram: a neural automated theorem prover for higher-order logic.