Applications of Deep Learning

Alpha Go – Google Translate – Data Center Optimisation

Robin Sigurdson, Yvonne Krumbeck, Henrik Arnelid

November 23, 2016
Applications of Deep Learning

Introduction
Applications of Deep Learning

Introduction

AlphaGo

Google’s Neural Machine Translation (GNMT)

Deep learning to control data centre cooling
Solving boardgames
Deep neural network
Supervised learning
Reinforcement learning from games of self-play
Deep neural network
Supervised learning
Reinforcement learning from games of self-play
Monte Carlo simulation
Policy network
Value network
Policy network

Value network
Policy network: classifies promising positions
Value Network: calculate estimates of winning
99.8% Winratio against other go programs
Google’s Neural Machine Translation (GNMT)
Google’s Neural Machine Translation (GNMT)

Introduction
Google’s Neural Machine Translation (GNMT)

Introduction

There are flaws, BUT...

- September 27, 2016 GNMT announced
- Error reduction by 60%
- “Bridging the Gap between Human and Machine Translation”

→ How?
Google’s Neural Machine Translation (GNMT)

Introduction

Overview

• Models used so far
• GNMT Model
• Experiments and Results
Google’s Neural Machine Translation (GNMT)

Models used so far

Overview

• Models used so far
• GNMT Model
• Experiments and Results
Google’s Neural Machine Translation (GNMT)
Models used so far

Phrase-Based Machine Translation (PBMT)
Google’s Neural Machine Translation (GNMT)
Models used so far

Phrase-Based Machine Translation (PBMT)

- Probability tables
- Linguistic properties

→ Neural Machine Translation (NMT)
Google’s Neural Machine Translation (GNMT)
Models used so far

Flaws of NMT

• Accuracy
• Speed / Computation
• Robustness
• Coverage
Google’s Neural Machine Translation (GNMT)

Overview

- Models used so far
- GNMT Model
- Experiments and Results
Google’s Neural Machine Translation (GNMT)

GNMT Model

How does GNMT handle these problems?

- Speed / Computation
- Robustness
How does GNMT handle these problems?

- Speed / Computation
- Robustness
Google’s Neural Machine Translation (GNMT)

GNMT Model

Architecture

→ Parallelism
Google’s Neural Machine Translation (GNMT)

GNMT Model

How does GNMT handle these problems?

- Speed / Computation
- Robustness
Google’s Neural Machine Translation (GNMT)

Segmentation: WordPiece Model (WPM)

Abwasserbehandlungsanlage

_Abwasser_ _behandlungs_ _anlage_

→ sewage water treatment plant
Google’s Neural Machine Translation (GNMT)
Experiments and Results

Overview

• Models used so far
• GNMT Model
• Experiments and Results
Google’s Neural Machine Translation (GNMT)
Experiments and Results

Tests on Benchmark Sentence Pairs

- Workshop on Machine Translation (WMT) data set
- BiLingual Evaluation Understudy (BLEU) metric
### Google’s Neural Machine Translation (GNMT)

Experiments and Results

#### Tests on Benchmark Sentence Pairs

<table>
<thead>
<tr>
<th></th>
<th>WMT En - Fr</th>
<th>WMT En - De</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model</td>
<td>BLEU</td>
</tr>
<tr>
<td>Word</td>
<td>37.90</td>
<td>0.2226</td>
</tr>
<tr>
<td>Character</td>
<td>38.01</td>
<td>1.0530</td>
</tr>
<tr>
<td>WPM-8K</td>
<td>38.27</td>
<td>0.1919</td>
</tr>
<tr>
<td>WPM-16K</td>
<td>37.60</td>
<td>0.1874</td>
</tr>
<tr>
<td>WPM-32K</td>
<td>38.95</td>
<td>0.1146</td>
</tr>
<tr>
<td>Mixed Word/Character</td>
<td>38.39</td>
<td>0.2774</td>
</tr>
<tr>
<td>PBMT</td>
<td>37.0</td>
<td></td>
</tr>
<tr>
<td>LSTM (6 layers)</td>
<td>31.5</td>
<td></td>
</tr>
<tr>
<td>LSTM (6 layers + PosUnk)</td>
<td>30</td>
<td>33.1</td>
</tr>
<tr>
<td>Deep-Att</td>
<td>37.7</td>
<td></td>
</tr>
<tr>
<td>Deep-Att + PosUnk</td>
<td>39.2</td>
<td></td>
</tr>
</tbody>
</table>
Google’s Neural Machine Translation (GNMT)

Experiments and Results

Human Evaluation

![Bar chart showing translation quality for different models and languages.]
Google’s Neural Machine Translation (GNMT)
Experiments and Results

“Bridging the Gap between Human and Machine Translation”
Deep learning to optimize cooling of data centres
Deep learning to control data centre cooling

Introduction

Facts and definitions

• 1 Google search = keep a lightbulb going for 25s
• 40.000 searches/s
• PUE = Power Usage Efficiency
Deep learning to control data centre cooling

Predicting PUE

- IT Load
- Pump Speed
- Chillers
- Wet Bulb

PUE
Deep learning to control data centre cooling

Predicting PUE

99.6% accuracy
Deep learning to control data centre cooling

Predicting PUE

Difficult to optimize efficiency

- Non-linear interactions between machines and environment
- Systems ability to adapt to operational changes
- Each facility has unique architecture
Deep learning to control data centre cooling

Controlling the data centre

Results: 40% reduction in energy usage for cooling
Conclusions

- Beating human intuition in board games
- Solving Language translation tasks
- Outperforming human engineering abilities
Thank you for listening!
Discussion Questions

• What are the limitations of deep learning? Are there tasks for which the technique cannot be applied?
• Are there areas where deep learning should be used, but isn’t?
• Who is responsible when a machine makes a critical error? For example: Who is responsible if an AI or machine causes a train to derail or fails to properly diagnose a patient?
References

- R. Sennrich, B. Haddow, A. Birch. *Neural Machine Translation of Rare Words with Subword Units.* (2016)
Applications of Deep Learning

References

- https://googleblog.blogspot.se/2014/05/better-data-centers-through-machine.html
- https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/
- https://googleblog.blogspot.se/2009/01/powering-google-search.html
- http://www.internetlivestats.com/google-search-statistics/