Semantic Analysis with Deep Neural Networks
Deep Learning Overview
Why Deep Learning?

- ImageNet Error Rate
  - Using deep learning
  - Human performance

- Deep Learning Popularity
  - Year

- Images of cats with dates
"[Machine Learning] gives computers the ability to learn without being explicitly programmed"

— Arthur Samuel, 1959
What is Machine Learning?

- Task: Part-of-Speech tagging
- Performance: e.g. accuracy
- Data: Annotated corpus

Demo!
What does the computer learn from?

**Features!**

- Hand-coded
- Time consuming
- Not necessarily effective
- Word and character n-grams
- Relevant linguistic properties (e.g. affixes, capitalisation, root form)

![Diagram showing model training and prediction with features like colours and number of corners]
What is a neural network?

- Biologically inspired
  1. Take an input
  2. Learn feature representations
  3. Predict output
  4. Self-correct if output is wrong
  5. Repeat!
What is Deep Learning?

❖ Deep Neural Networks
❖ Automatically combine simple features into complex features
❖ Deeper is (often) better

Demo
playground.tensorflow.org
What can Deep Learning do?

❖ Make psychedelic images!
❖ Google QuickDraw: https://quickdraw.withgoogle.com
❖ Text-to-speech: https://deepmind.com/blog/wavenet-generative-model-raw-audio/
❖ Generate hand-writing: http://www.cs.toronto.edu/~graves/handwriting.cgi
❖ Currently the most successful approach to many NLP problems
Deep Learning for everything?

- Not a silver bullet
- Simple problems do not require fancy methods
Deep Learning and the Human Brain

- In Computational Linguistics: Not an attempt to model the brain
- Some inspiration is useful (ReLU)
Why are neural networks back?

- More computational power (GPUs)
- More data
- Better algorithms/architectures
Semantic Analysis with Deep Neural Networks
Semantic Analysis

- Parallel Meaning Bank
  http://pmb.let.rug.nl/explorer/explore.php

- English, Dutch, German, Italian

- **Goal:**
  Parallel corpus with Discourse Representation Structures for all languages

- About 11 million tokens
Chapter I: Semantic Tagging

- Multilingual Semantic Parsing
- Experimenting with different Neural Network architectures
Semantic Tags — Motivation

- POS tags: insufficient and irrelevant information

- Insufficient:
  - *every* (DT / univ. quant.)
  - *no* (DT / neg.)
  - *some* (DT / exist. quant.)

- Irrelevant:
  - *walks* (VBZ / pres. simpl.)
  - *walk* (VBP / pres. simpl.)
Semantic Tags — Example

Tokens:  These cats live in that house .

Sem-tags: PRX  CON  ENS  REL  DST  CON  NIL

UD-POS:  DET  NOUN  VERB  ADP  DET  NOUN  PUNCT
Semantic Tags — Overview

- About 75 tags
- Abstract over POS and NE tags
- Includes categories for negation, modality and quantification
- Generalises over languages (en, de, nl, it)
Auxiliary tasks

- Giving the NN more work to do
- Informing the NN of what additional task might be helpful to learn
- Word frequencies for POS tagging
- This work: Semantic tags for POS tagging

[Plank et al., 2016]
## Results

### Table 1: Experiment results on semtag (ST) test sets (% accuracy).

<table>
<thead>
<tr>
<th>Baselines</th>
<th>MFC</th>
<th>TNT</th>
<th>Bi-LSTM</th>
<th>Bi-GRU</th>
<th>$\tilde{c}$</th>
<th>$\tilde{c} \wedge \tilde{w}$</th>
<th>+AUX</th>
<th>$\tilde{c}$</th>
<th>$\tilde{c} \wedge \tilde{w}$</th>
<th>+AUX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semtag Silver</td>
<td>84.64</td>
<td>92.09</td>
<td>94.98</td>
<td>94.26</td>
<td>91.39</td>
<td>94.63</td>
<td>94.53</td>
<td>94.39</td>
<td><strong>95.14</strong></td>
<td>94.23</td>
</tr>
<tr>
<td>Semtag Gold</td>
<td>77.39</td>
<td>80.73</td>
<td>82.96</td>
<td>80.26</td>
<td>69.21</td>
<td>76.83</td>
<td>80.73</td>
<td>76.89</td>
<td><strong>83.64</strong></td>
<td>74.84</td>
</tr>
</tbody>
</table>

MFC indicates the per-word most frequent class baseline, TNT indicates the TNT tagger, and Bi-LSTM indicates the system by Plank et al. (2016). Bi-GRU indicates the $\tilde{w}$ only baseline. $\tilde{w}$ indicates usage of word representations, $\tilde{c}$ indicates usage of character representations. The +AUX column indicates the usage of an auxiliary loss.

### Table 2: Experiment results on Universal Dependencies (UD) test sets (% accuracy).

<table>
<thead>
<tr>
<th>Baselines</th>
<th>MFC</th>
<th>TNT</th>
<th>Bi-LSTM</th>
<th>Bi-GRU</th>
<th>$\tilde{c}$</th>
<th>$\tilde{c} \wedge \tilde{w}$</th>
<th>+AUX</th>
<th>$\tilde{c}$</th>
<th>$\tilde{c} \wedge \tilde{w}$</th>
<th>+AUX</th>
</tr>
</thead>
<tbody>
<tr>
<td>UD v1.2</td>
<td>85.06</td>
<td>92.66</td>
<td>95.17</td>
<td>94.39</td>
<td>77.63</td>
<td>94.68</td>
<td>95.19</td>
<td>92.65</td>
<td><strong>94.92</strong></td>
<td>95.71</td>
</tr>
<tr>
<td>UD v1.3</td>
<td>85.07</td>
<td>92.69</td>
<td>95.04</td>
<td>94.32</td>
<td>77.51</td>
<td>94.89</td>
<td>95.34</td>
<td>92.63</td>
<td><strong>94.88</strong></td>
<td>95.67</td>
</tr>
</tbody>
</table>

Table 1: Experiment results on semtag (ST) test sets (% accuracy).
Chapter II: Multitask Learning

When and why does Multitask Learning help?
When does MTL help?

❖ “[…] when the label distribution is compact and uniform”
❖ —> High entropy, few labels
Is ‘high entropy’ sufficient?

Tokens: These cats live in that house.

Sem-tags: PRX CON ENS REL DST CON NIL

UD-POS: DET NOUN VERB ADP DET NOUN PUNCT
Is ‘high entropy’ sufficient?

Tokens: These cats live in that house.

Sem-tags: CON NIL DST PRX ENS REL CON

UD-POS: DET NOUN VERB ADP DET NOUN PUNCT
Tagset correlations

UD - SemTag correlation
Information-theoretic Measures

- Calculating tagset correlations:
  - Conditional Entropy
  - Mutual Information
Correlation with Auxiliary Task Effectivity

Conditional Entropy and Mutual Information both correlate far better than entropy!

| Auxiliary task          | \(\rho(\Delta_{\text{acc}}, H(Y))\) | \(\rho(\Delta_{\text{acc}}, H(Y|X))\) | \(\rho(\Delta_{\text{acc}}, I(X;Y))\) |
|-------------------------|-------------------------------------|-------------------------------------|-------------------------------------|
| Supertagging (Identity) | -0.06 (p=0.214)                     | 0.12 (p=0.013)                      | 0.08 (p=0.114)                      |
| Supertagging (Overlap)  | 0.07 (p=0.127)                      | 0.27 (p<0.001)                      | 0.43 (p<<0.001)                     |
| Supertagging (Disjunct) | 0.08 (p=0.101)                      | 0.25 (p<0.001)                      | 0.41 (p<<0.001)                     |
Change in accuracy (x) vs. Entropy (y)
Change in accuracy (x) vs. Mutual Information (y)
Remaining chapters

- Chapter III: Multilingual Learning
- Chapter IV: Semantic Similarity between Words and Sentences (SemEval Shared Tasks)
- Chapter V: Dataset Augmentation