1. Motivation

- Internal representations in deep neural network (NN) models are not well understood.
- Previous studies analyze full vector representations and do not inspect individual dimensions.
- We study individual neurons in neural machine translation (NMT) and neural language modeling (NLM) via three questions:
  1. Do they contain interpretable linguistic information?
  2. Do they play an important role for obtaining high-quality translations?
  3. Can we manipulate the translation in desired ways by modifying specific neurons?
- Potential applications in model distillation and mitigating model bias.

2. Linguistic Correlation Method

Goal: Identify linguistically motivated neurons in deep NLP models through auxiliary tasks.
Example: Morphological or Semantic tagging

Approach:
- Extract neuron activations from the model for every input word.
- Train a classifier on extracted activations against some supervised task.
- Extract a ranking of the neurons using the trained weights.
- Learned weights are representatives of which neurons are important for a property

3. Evaluation via Ablation

<table>
<thead>
<tr>
<th>Task</th>
<th>ALL</th>
<th>10% Top</th>
<th>15% Top</th>
<th>20% Top</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT</td>
<td>FR (POS)</td>
<td>93.2</td>
<td>63.2</td>
<td>23.8</td>
</tr>
<tr>
<td></td>
<td>EN (POS)</td>
<td>93.5</td>
<td>60.3</td>
<td>21.5</td>
</tr>
<tr>
<td></td>
<td>DE (POS)</td>
<td>90.1</td>
<td>51.5</td>
<td>16.3</td>
</tr>
<tr>
<td>NLM</td>
<td>FR (POS)</td>
<td>92.4</td>
<td>41.6</td>
<td>23.8</td>
</tr>
<tr>
<td></td>
<td>EN (POS)</td>
<td>92.9</td>
<td>54.6</td>
<td>26.4</td>
</tr>
<tr>
<td></td>
<td>DE (POS)</td>
<td>92.0</td>
<td>49.7</td>
<td>21.9</td>
</tr>
</tbody>
</table>

Classifier accuracy on different tasks using all neurons (ALL).

4. Focussed vs. Distributed

- Cross model correlation method

5. Visualizations

6. Cross Model Correlation Method

7. Ablation Studies

8. Controlling Translations

- Hypothesis: If a neuron matters to the model, then we can manipulate the translation by modifying its activations.
- Method:
  1. Encode the source sentence as usual
  2. Before decoding, replace the activation of a particular neuron with a value of \( \alpha \)
  3. Observe how translation changes with different \( \alpha \) values
- Tense manipulation
  1. Changing tense (past->present / present->past) in several languages
  2. Example: Morphological or Semantic tagging
- Hypothesis: Different NMT models learn similar properties, and therefore should have similar neurons.
- Approach: Rank neurons by strength of their correlations with neurons from other networks, on several levels.

Effect of neuron on language model perplexity when removing top or bottom neurons based on Cross correlation analysis reordering.

Effect of neuron on language model perplexity when removing top or bottom neurons based on Cross correlation analysis reordering.

Ablating top neurons is more damaging than ablating bottom neurons.

MaxCorr/MinCorr/LinReg are similar; SVCCA has very important top directions

This research was carried out in collaboration between the HBKU Qatar Computing Research Institute (QCRI) and the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) — * denotes equal contribution

Acknowledgements: This research is supported by the Qatar National Research Fund (QNRF) under grant number (NPRP 8-590-2-143).

Also See: Identifying and Controlling Important Neurons in Neural Machine Translation. Accepted at ICLR’19 NeuroX: A Toolkit for Analyzing Individual Neurons in Neural Networks, AAAI’19 Demonstration

We were able to control tense (up to 67%), but gender and number are harder (21% and 37%).

Fahim Dalvi* Nadir Durrani* Hassan Sajjad* Yonatan Belinkov Anthony Bau James Glass {faimaduddin,nadurrani,hsajjad}@qf.org.qa {belinkov,abau,glass}@mit.edu