Identifying and Controlling Important Neurons in Neural Machine Translation

Anthony Bau*, Yonatan Belinkov*, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, James Glass

{belinkov,abau,glass}@mit.edu {faimaduddin,ndurrani,hsajjad}@qf.org.qa

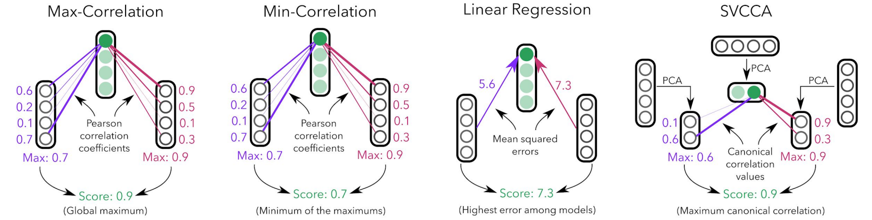
1. Motivation

- Internal representations in Neural Machine Translation (NMT) are not well understood.
- Previous work analyzed NMT at the level of whole vector representations. In contrast, computer vision work found meaningful individual units (Bau et al.; Zhou et al.).
- Previous work requires external supervision (linguistic annotation).
- We develop **unsupervised** methods for finding important neurons.
- **Key point**: different models learn similar patterns → similar important neurons should emerge in the models.
- We analyze their linguistic content using visualization and classification.
- We intervene in the representations at the neuron level and evaluate our success to control NMT output along linguistic properties.
- Potential applications: model distillation and mitigating model bias.

2. Experimental Setup

- **Data**:
  - The United Nations parallel corpus.
  - MT models from English to Arabic, Chinese, French, Russian, Spanish, and an English-to-English auto-encoder.
- **MT models**:
  - 500 dimensional 2-layer LSTM encoder-decoders with attention.
  - 3 models per language pair, on different training partitions.

3. Unsupervised Correlation Methods

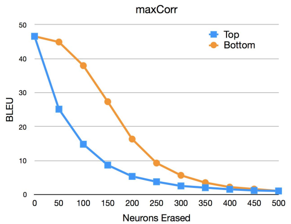


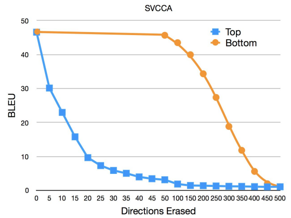
- **Hypothesis**: Different NMT models learn similar properties, and therefore should have similar neurons.
- **Approach**: Rank neurons by strength of their correlations with neurons from other networks, on several levels.

4. Ablation Studies

Results:

- Ablating top neurons is more damaging than ablating bottom neurons.
- This confirms the ranking correctness.
- MaxCorr/MinCorr/LinReg are similar; SVCCA has very important top directions
- These results are consistent across language pairs.





5. Analyzing Individual Neurons

Top ranked neurons:

MaxCorr			MinCorr			LinReg			SVCCA	
ID	Pos	Tok	ID	Pos	Tok	ID	Pos	Tok	Pos	Tok
464	**92%**	10%	342	**88%**	7.9%	464	**92%**	10%	**86%**	26%
342	**88%**	7.9%	464	**92%**	10%	260	0.71%	**94%**	1.6%	**90%**
260	0.71%	**94%**	260	0.71%	**94%**	139	0.86%	**93%**	7.5%	**85%**
49	11%	6.1%	383	**67%**	6.5%	494	3.5%	**96%**	20%	**79%**
124	**77%**	48%	250	**63%**	6.8%	342	**88%**	7.9%	1.1%	**89%**
394	0.38%	22%	124	**77%**	47%	228	0.38%	**96%**	10%	**76%**

Results:

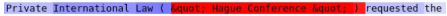
- Many top neurons capture position, especially with MinCorr and MaxCorr, showing this property arises in many models.
- Many top LinReg neurons or SVCCA directions capture token identity, implying that this information is more distributed.

Linguistic neurons:

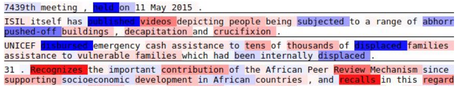
- Other top neurons capture various linguistic properties.

Visualizations:

- Parentheses


Private International Law (&quot; Hague Conference &quot;) requested the

- Tense

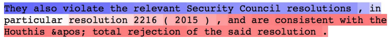

7439th meeting , held on 11 May 2015 .

ISIL itself has published videos depicting people being subjected to a range of abhorr
pushed-off buildings , decapitation and crucifixion .

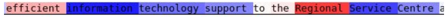
UNICEF disbursed emergency cash assistance to tens of thousands of displaced families
assistance to vulnerable families which had been internally displaced .

31 . Recognizes the important contribution of the African Peer Review Mechanism since
supporting socioeconomic development in African countries , and recalls in this regard

- Position


They also violate the relevant Security Council resolutions , in
particular resolution 2216 (2015) , and are consistent with the
Houthis &apos; total rejection of the said resolution .

- Noun phrase segmentation


efficient information technology support to the Regional Service Centre a

6. Controlling Translations

- Can we control translations by modifying activations of (source) neurons?
- Motivation: control sensitive attributes, such as handling gender bias.
- We were able to control tense (up to 67%), but gender and number are harder (21% and 37%).
- **Example**: change the translation of “The committee *supported* the efforts of the authorities” from past to present.

	α	Translation	Tense
Arabic	–/+10	وأيدت\وتؤيد اللجنة {جهود\الجهود التي تبذلها} السلطات	past/present
French	–/-20	Le Comité a appuyé/appuie les efforts des autorités	past/present
Spanish	–/-3/0	El Comité apoyó/apoyaba/apoya los esfuerzos de las autoridades	past/impf./present
Russian	–/-1	Комитет поддержал/поддерживает усилия властей	past/present
Chinese	–/-50	委员会 支持 当局 的 努力 / 委员会 正在 支持 当局 的 努力	untensed/present

- **Example**: change gender in the translation of “The interested *parties*”.
  - Notice agreement of determiner-noun-adjective

-0.5, -0.25	Los partidos interados	masculine
0, 0.25	Las partes interesadas	feminine

See Also

- Analyzing Individual Neurons in Deep NLP Models, AAAI
- NeuroX: Analysis toolkit

Acknowledgements: 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