What Is One Grain of Sand in the Desert? Analyzing Individual Neurons in Deep NLP Models

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

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

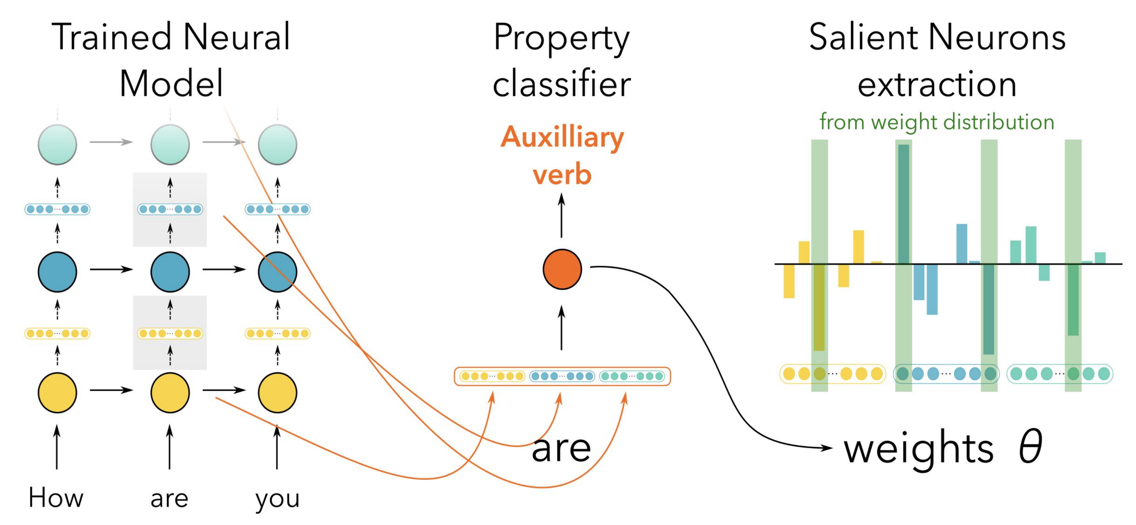
1. Motivation

- 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



$$\mathcal{L}(\theta) = -\sum_i \log P_\theta(\mathbf{l}_i|\mathbf{x}_i) + \lambda_1\|\theta\|_1 + \lambda_2\|\theta\|_2^2$$

3. Evaluation via Ablation

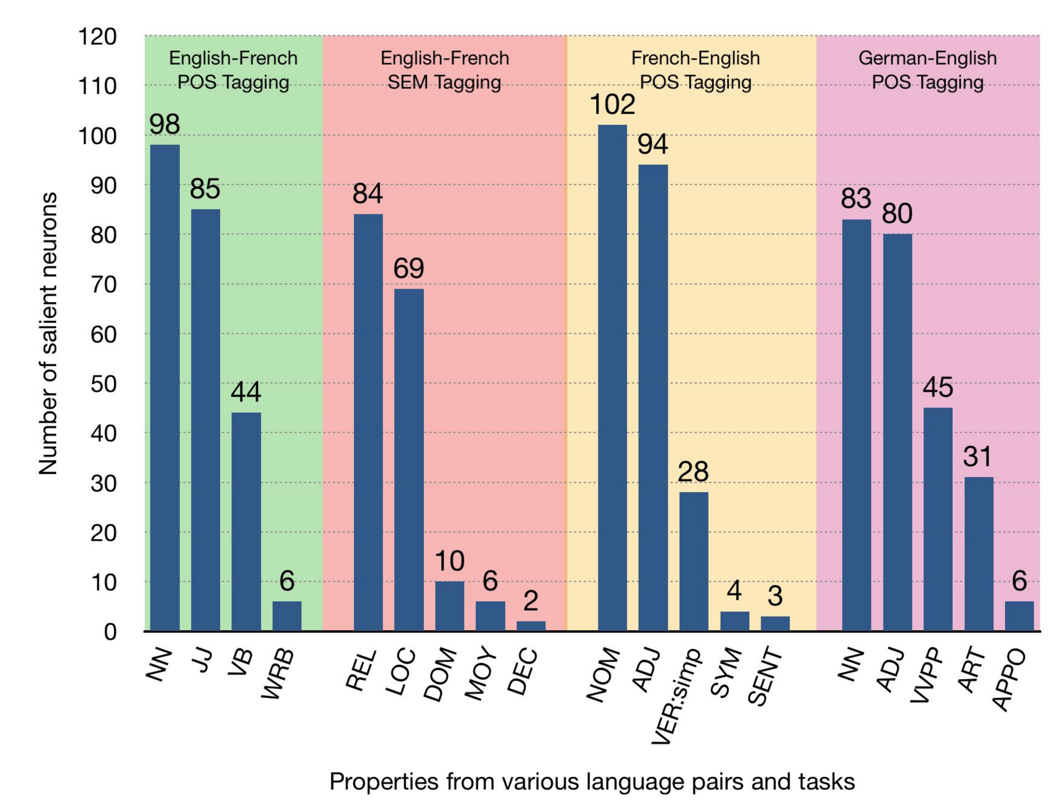
	French		English		German	
	POS	Morph	POS	SEM	POS	Morph
MAJ	92.8	89.5	91.6	84.2	89.3	83.7
NMT	93.2	88.0	93.5	90.1	93.6	87.3
NLM	92.4	90.1	92.9	86.0	92.3	86.5

Classifier accuracy when trained on activations of NMT and NML models. MAJ: local majority baseline

	Task	ALL	Masking-out 10% Top	10% Bot	15% Top	15% Bot	20% Top	20% Bot
NMT	FR (POS)	93.2	63.2	23.8	73.0	24.8	79.4	24.9
NMT	EN (POS)	93.5	69.8	15.8	78.3	17.9	84.1	21.5
NMT	EN (SEM)	90.1	51.5	16.3	65.3	18.9	74.2	20.7
NMT	DE (POS)	93.6	65.9	15.7	78.0	15.6	88.2	15.7
NLM	FR (POS)	92.4	41.6	23.8	53.6	23.8	59.6	24.0
NLM	EN (POS)	92.9	54.2	18.4	66.1	20.4	72.4	24.7
NLM	EN (SEM)	86.0	49.7	21.9	56.8	22.3	65.2	25.1
NLM	DE (POS)	92.3	39.7	16.7	51.7	16.7	67.2	16.9

Classifier accuracy on different tasks using all neurons (ALL).
Masking-out: all except top/bottom N% neurons

4. Focussed vs. Distributed



Properties from various language pairs and tasks

Results:
- Open class properties such as Nouns and Named Entities are much more distributed across the network compared to closed properties.
- The model recognizes hierarchy in language and distributes neurons based on it.
  - E.g multiple neurons for different verb forms

5. Visualizations

Supports the efforts of the Libyan authorities to recover funds misappropriated under the Qadhafi regime

(a) English Verb (#1902)

einige von Ihnen haben vielleicht davon gehört , dass ich vor ein paar Wochen eine Anzeige bei Ebay geschaltet habe .

(b) German Article (#590)

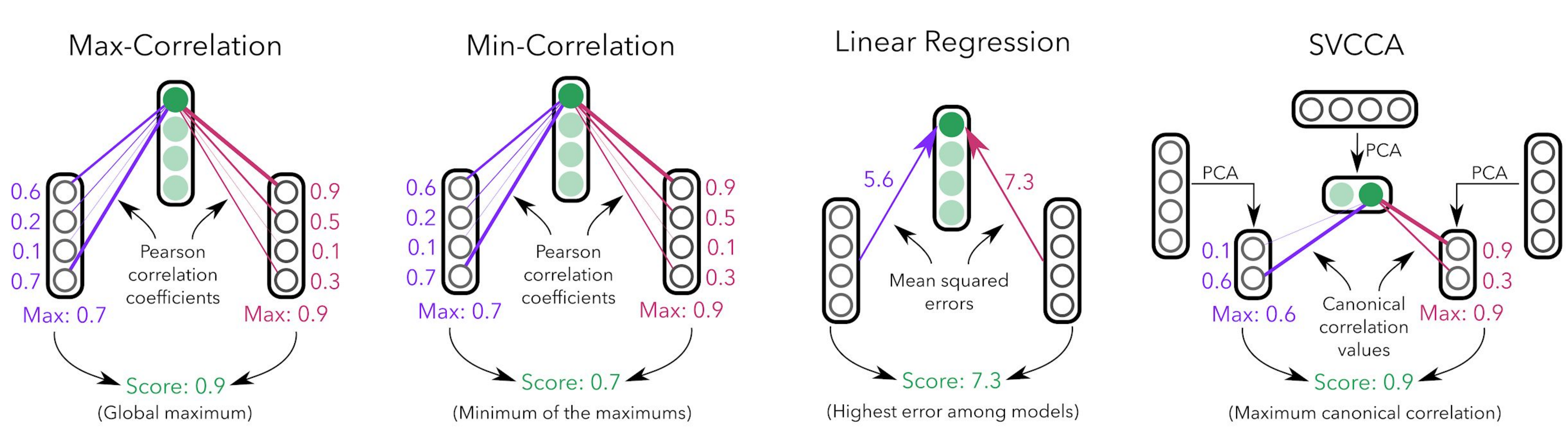
They also violate the relevant Security Council resolutions , in particular resolution 2216 (2015) , and are consistent with the Houthis ' total rejection of the said resolution .

(c) Position Neuron (#1903)

Neuron	Top 10 words
#1925 (Month)	August, July, January, September, October, presidential, April, May, February, December
#1960 (Negation)	no, No, not, nothing, nor, neither, or, none, whether, appeal
#1590 (Cardinality)	50, 10, 51, 61, 47, 37, 48, 33, 43, 49

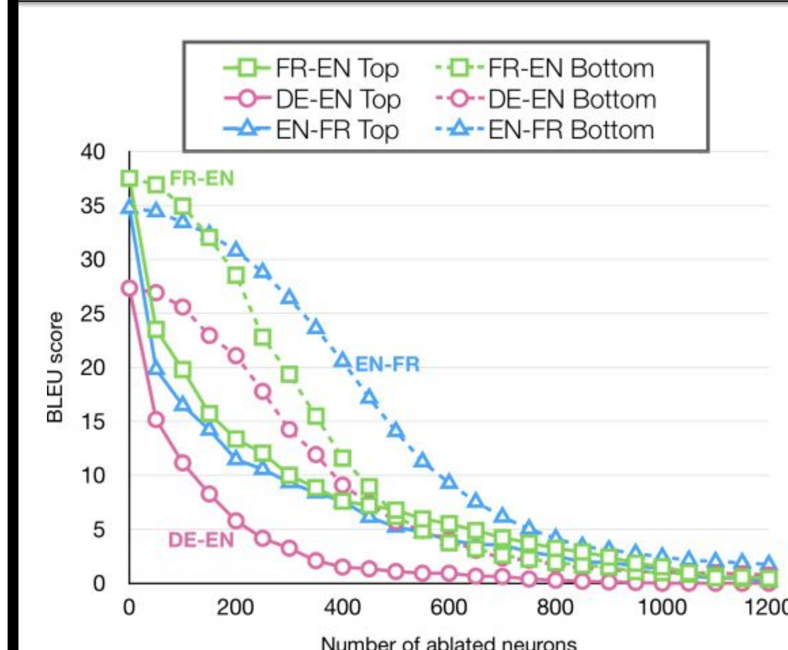
Ranked list of words for some individual neurons in the EN-FR NMT model

6. Cross Model Correlation Method

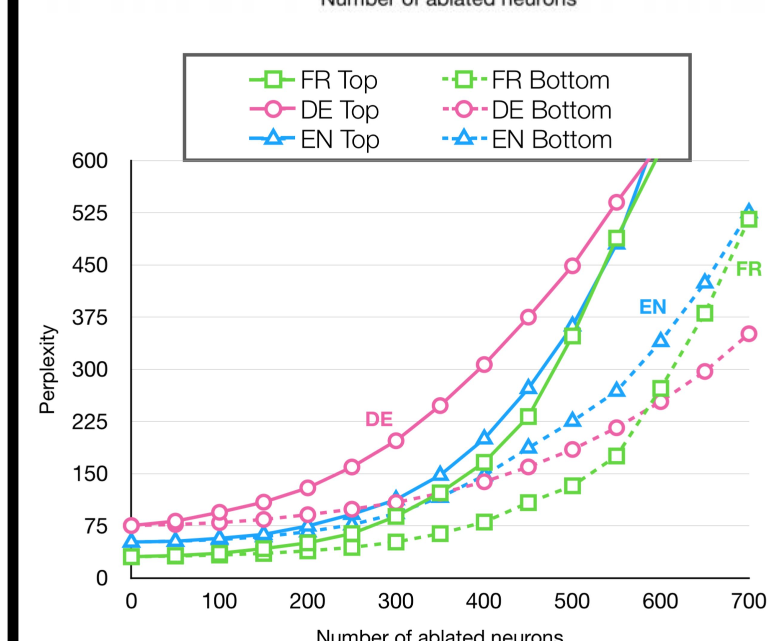


- **Motivation:** Linguistic correlation method may not be able to identify all the important neurons for the model itself
- **Hypothesis**: Different NMT models learn similar properties, and therefore should have similar neurons.
- **Approach**: Rank neurons by strength of their correlations with neurons from other networks, on several levels.

7. Ablation Studies

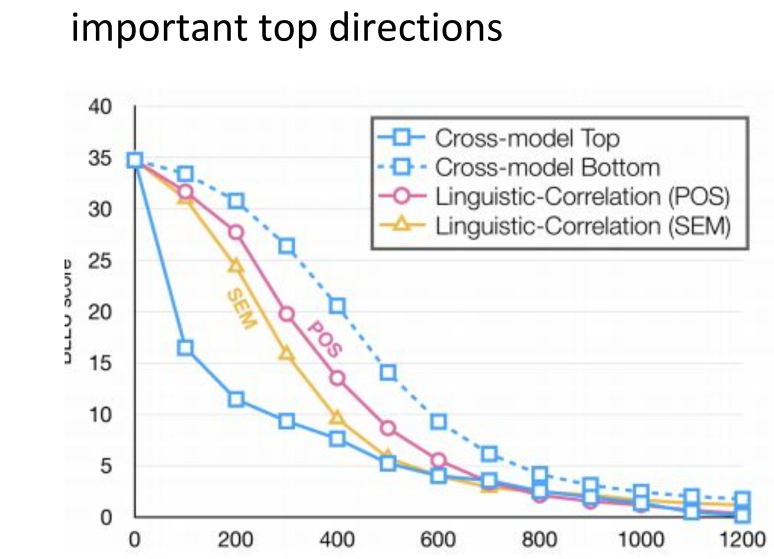


Effect of neuron on BLEU when removing top or bottom neurons based on Cross correlation analysis reordering



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



Effect on translation when ablating neurons in the order determined by both methods on the EN-FR model

8. Controlling Translations

- **Hypothesis**: if a neuron matters to the model, then we can manipulate the translation by modifying its activations
- **Method**
  - Encode the source sentence as usual
  - Before decoding, replace the activation of a particular neuron with a value of α
  - Observe how translation changes with different α values
- **Tense manipulation**
  - Changing tense (past->present / present->past) in several languages
  - Changing phrasing in Arabic translation beyond the verb
  - Chinese is hard to manipulate (needs high α), possibly because tense is usually not marked

	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

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

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

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