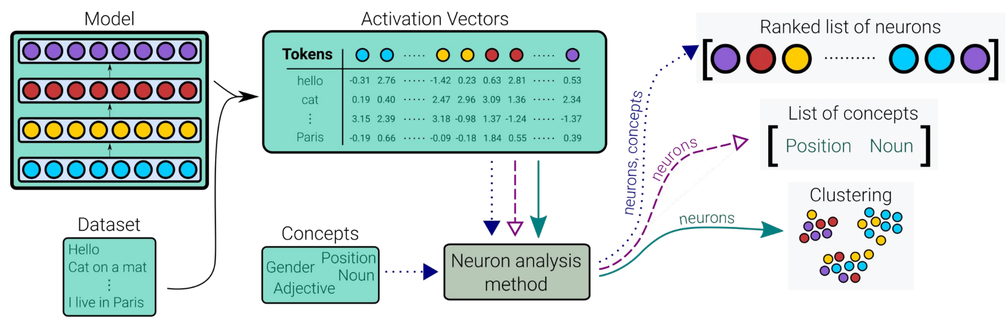
Neuron-level Interpretation of Deep NLP Models: A Survey

Hassan Sajjad Nadir Durrani Fahim Dalvi
hsajjad@dal.ca {ndurrani, faimaduddin}@hbku.edu.qa

Neuron Interpretation aims to understand
How is knowledge structured within neural network representations?

Methods



Definitions

- **Neuron:** Output of a single dimension from any neural network component (blocks, layers, attention)
- **Concept:** A group of words that are coherent w.r.t a linguistic property (Nouns, Professions, etc)

Classification Criteria:

- Does the method provide **global** or **local** interpretation?
- **Input** and **Output** e.g. a set of neurons or concepts
- **Scalability**: Can the method be scaled to a larger set of neurons?
- Does the method require **human-in-the-loop**?
- Does the method require **supervised data**?
- Is the interpretation connected to **model's prediction**?

	Scope	Input	Output	Scalability	HITL	Supervision	Causation
Visualization							
Karpathy et al. (2015)	local	neuron	concept	low	yes	no	no
Corpus-based methods							
Concept Search							
Kádár et al. (2017)	global	neuron	concept	low	yes	no	no
Na et al. (2019)	global	neuron	concept	high	no	no	no
Neuron Search							
Mu and Andreas (2020); Suau et al. (2020); Antverg and Belinkov (2022)	global	concept	neurons	high	no	yes	no
Probing-based methods							
Linear (Dalvi et al., 2019)	global	concept	neurons	high	no	yes	no
Gaussian (Hennigen et al., 2020)	global	concept	neurons	high	no	yes	no
Causation-based methods							
Ablation (Lakretz et al., 2019)	both	concept/ class	neurons	medium	no	no	yes
Knowledge attribution (Dai et al., 2021)	local	concept/ class	neurons	high	no	no	yes
Miscellaneous methods							
Corpus generation (Poerner et al., 2018)	global	neuron	concept	low	yes	no	no
Matrix factorization (Alammar, 2020)	local	neurons	neurons	low	yes	no	no
Clustering (Dalvi et al., 2020)	global	neurons	neurons	high	yes	no	no
Multi model search (Bau et al., 2019)	global	neurons	neurons	high	yes	no	no

Comparison of various neuron interpretation methods

Findings

Sentiment Intensification
"I like this movie **a lot**"
'the movie is **incredibly good**"

Negation
"... but **not** that ...."

Core Linguistic Concepts
Part of Speech Information

Polysemous Behavior
Tense switch neurons

Information Distribution
Syntax is captured at higher layers
Lower layers within an FFNN block house more salient neurons

Redundancy
Many neurons learn the same concept

Specific Semantics
Electronic items: "camera, laptops,cables"
Salad items: "broccoli, noodles"
Law: "law, legal, case"

Phrasal Neurons
Phrasal neurons: "horse racing"

Syntactic Concepts
Positional Information
Parenthesis
Alignment to Syntactic parses

Complex Semantic Concepts
Causativity neurons

Architectural Differences
Auto-encoder vs generative models

Finetuned Models
Fine-tuning forces core linguistic knowledge into neurons from lower layers

Applications

- **Model Control**
  - Changing tense in output translations from present to past-tense in an NMT model
- **Model Distillation**
  - Efficient Feature-based transfer learning
- **Domain Adaptation**
  - Prune unimportant neurons or finetune with frozen salient neurons
- **Compositional Explanations**

Open Challenges

- Gap between "what is learned" and "how is it used"
- Choice of "best" neuron interpretation method not clear
- Lack of evaluation benchmarks
- Reliance on pre-defined/annotated corpora

NeuroX
https://neurox.qcri.org

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY