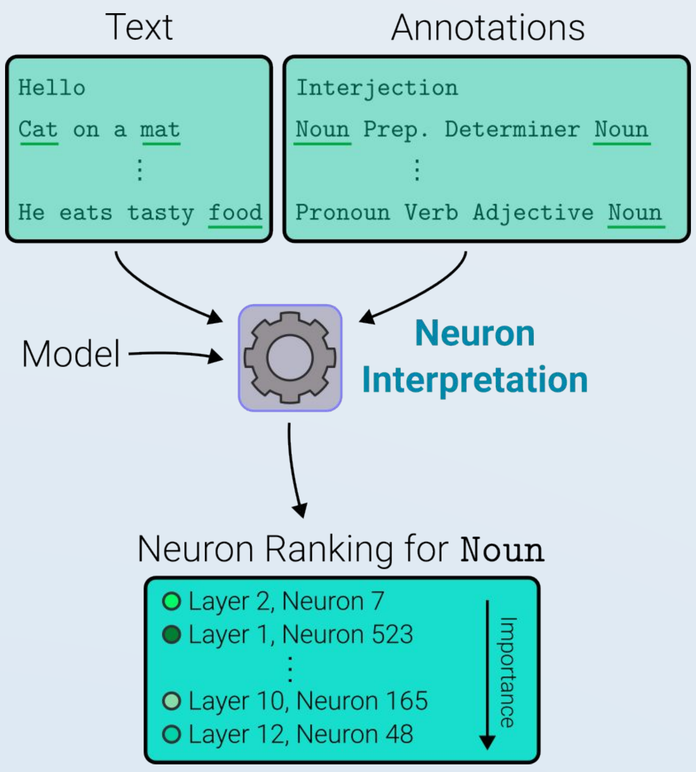
NeuroX

An *open-source* library for neuron interpretation

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

Interpretation of deep NLP models is important to build trustworthy systems and understand why and how they work



Neuron Interpretation Methods

Linear Probe Probeless IoU Probe
Gaussian Probe Mean Select

Supported Models

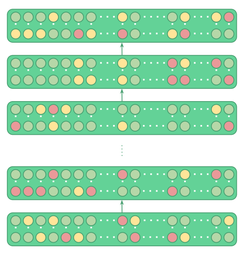
Hugging Face

Features

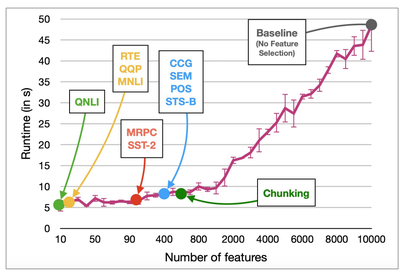
Annotation Helpers
Control tasks
Neuron visualizers

How does the library help?

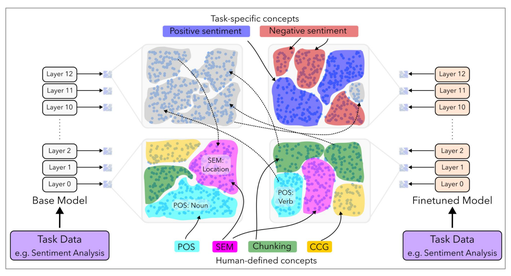
Neuron Activations Extraction



Downstream applications

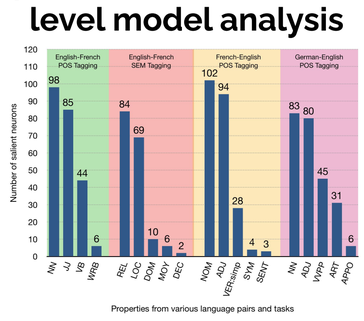


Efficient feature selection



Latent concept analysis

Representation and Neuron level model analysis



Qualitative analysis of neurons

```
It has been trying to improve its share of the residential
market .

He declined to comment on whether the company is
considering a dividend or is planning any acquisition .

The university is considering installing a $ 250,000 system
to store applications electronically .
```

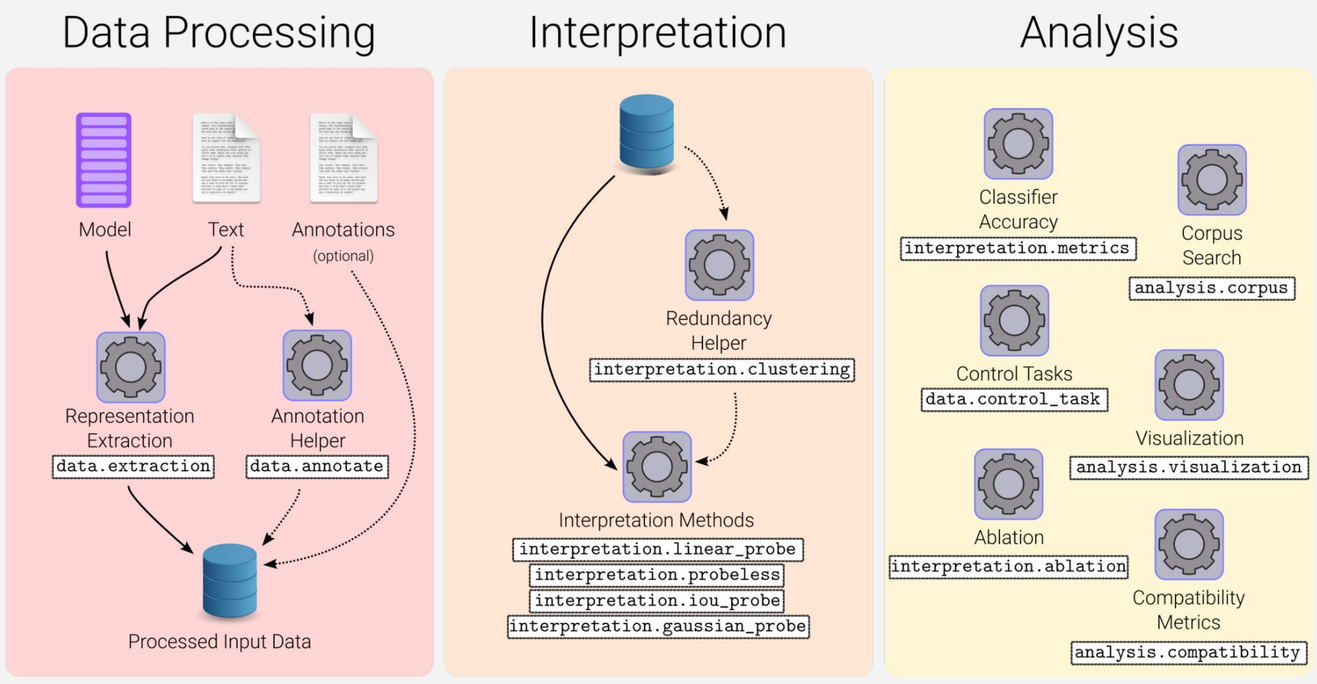
Getting started

```
pip install neurox
```

https://neurox.qcri.org

Check out the docs, examples, source code and more!

Library modules



Contributions are welcome!

10K+ installs so far. Help us support more interpretation methods and models! Thanks to Ahmed Abdelali, David Arps, Yimin Fan, Yifan Zhang for their contributions so far.

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