Evaluating Neuron Interpretation Methods of NLP Models

Yimin Fan fanyimin@link.cuhk.edu.hk
Fahim Dalvi {faimaduddin,ndurrani}@hbku.edu.qa
Nadir Durrani {faimaduddin,ndurrani}@hbku.edu.qa
Hassan Sajjad hsajjad@dal.ca

NEURAL INFORMATION PROCESSING SYSTEMS

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

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

Numerous Approaches Exist for Neuron Interpretation

Probeless IoU Probe Lasso Probe Ridge Probe
LCA Probe Gaussian Probe Mean Select

Goal is to discover **salient neurons** of a network with respect to any given concept

Problem: Evaluating and Comparing Methods is Difficult!

Salient neurons are usually evaluated manually, or by training classifiers using selected neurons

	Salient Neurons			
Probeless	17	53	1	...
LCA Probe	53	77	17	...

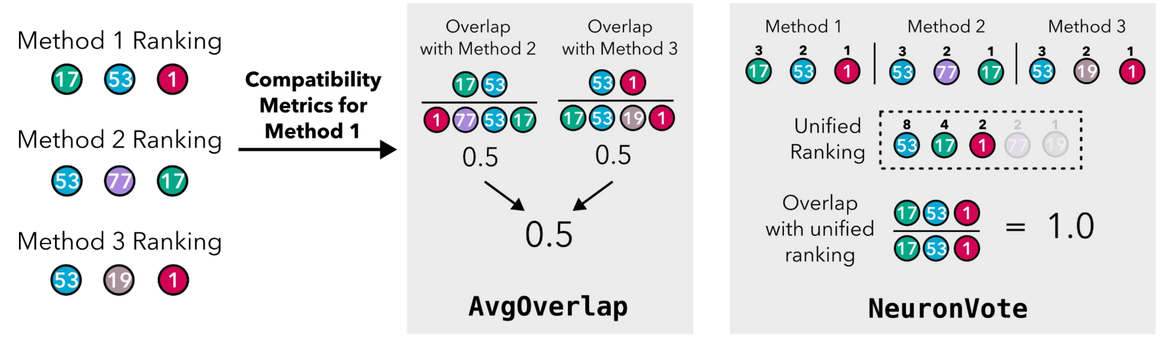
Which ranking is more accurate?

Standard benchmarks and metrics do not exist

Creating a benchmark for "correct neurons" is challenging and infeasible

Can Voting Be Used as an Evaluation Metric?

A viable approach can be to assess the **compatibility** among the methods using voting. Neurons that are commonly discovered by different interpretation methods should be more informative than others!

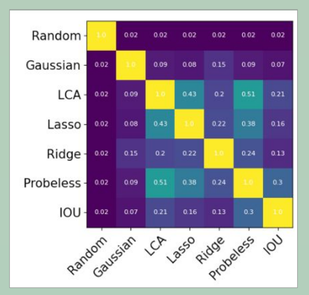


Results

	AvgOverlap	NeuronVote
Random	0.021	0.021
Gaussian	0.086	0.169
LCA	0.258	0.514
Lasso	0.240	0.473
Ridge	0.177	0.362
Probeless	**0.269**	**0.532**
IoU	0.156	0.365

Both Metrics show similar patterns. **Probeless** is most compatible with other methods

LCA and Lasso both showed a substantial drop in their compatibility scores for the 12th layer



Pairwise analysis shows Probeless and LCA agree with each other a lot, even with very different underlying theoretical mechanics

Compatibility trends are consistent across models

Toolkit

NeuroX

香港中文大學 The Chinese University of Hong Kong

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

DALHOUSIE UNIVERSITY

Paper