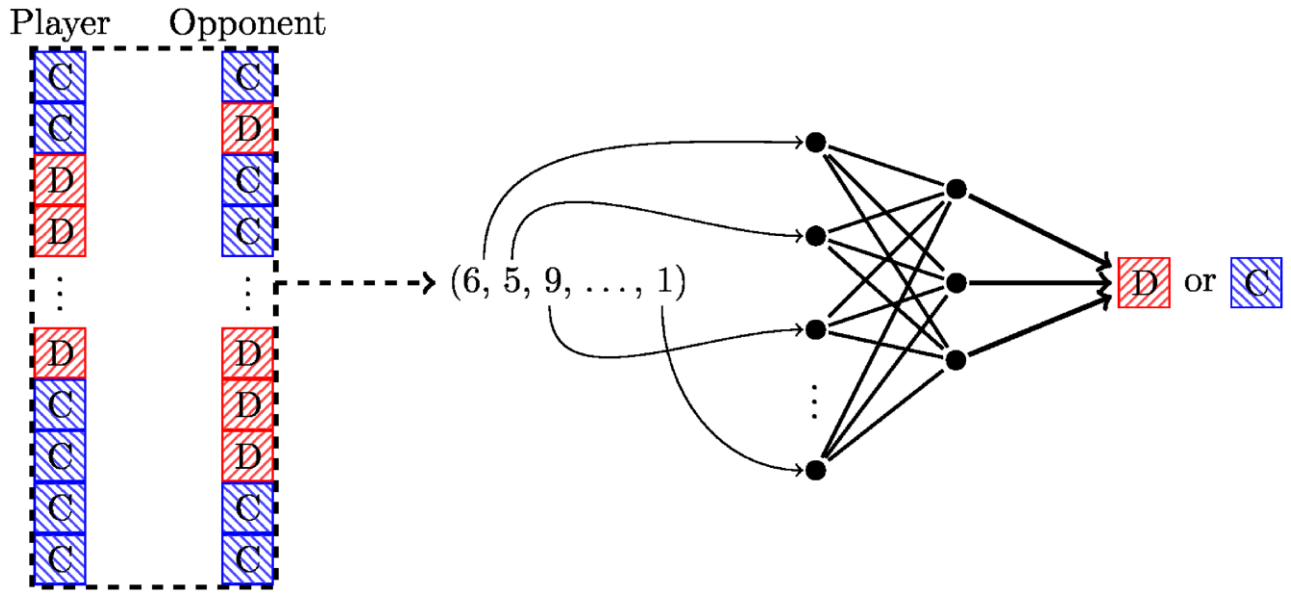
Interpret strategies based on neural nets



- Forgivingness
- Generosity
- Competitiveness
  (Winning is everything or not)
- Extortion
- Anger
- Contrite
- Emotions

https://openai.com/research/language-models-can-explain-neurons-in-language-models

Interpreting IPD Models with Large Language Models

Objectives, Claims, Experiments

Objective

- We want to develop a method that can *mechanistically* interpret *learned features* of a neural network using large language models.
- We seek to demonstrate interpretability at two drastically different scales:
  - The neuronal level: summarizing the *feature* that each neuron captures.
  - The network level: provides a rationale for the decision of the network as a whole.
- We demonstrate our method on neural networks that solve the iterated prisoner's dilemma (IPD).
  - *Why prisoner's dilemma?*
    - Games like the iterated prisoner's dilemma are well-defined and well-analyzed tasks which have many known interpretable strategies for game play.
    - Moreover, any strategy can be quantitatively and qualitatively analyzed (by humans) more easily compared to CV tasks or NLP tasks.
    - Note that for a game like IPD, in many cases, there is "only one interpretation" which makes it easy to qualitatively verify our interpretations.
  - *Why a large language model?*
    - See related literature. Also, LLMs may capture hidden dependencies between layers (sequence data, maybe strong auto-regressive properties?) well.
    - LLMs' emergent properties may allow for more interesting reasoning about the relationships between various features. This more aptly characterizes the "critical thinking" that is required interpret.
  - *What type of neural networks are we interpreting?*
    - Linear forward feeding networks that are trained using reinforcement learning (or genetic algorithm) to play the iterated prisoner's dilemma.

Related Literature

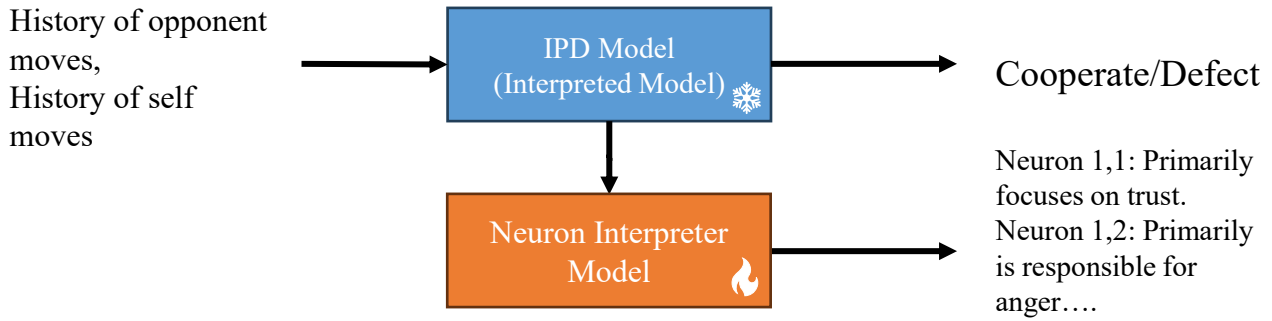
- *Mechanistic Interpretability.* (existing statistical methods to interpret neural networks)

- Superposition. (why this task is more non-trivial than it seems)

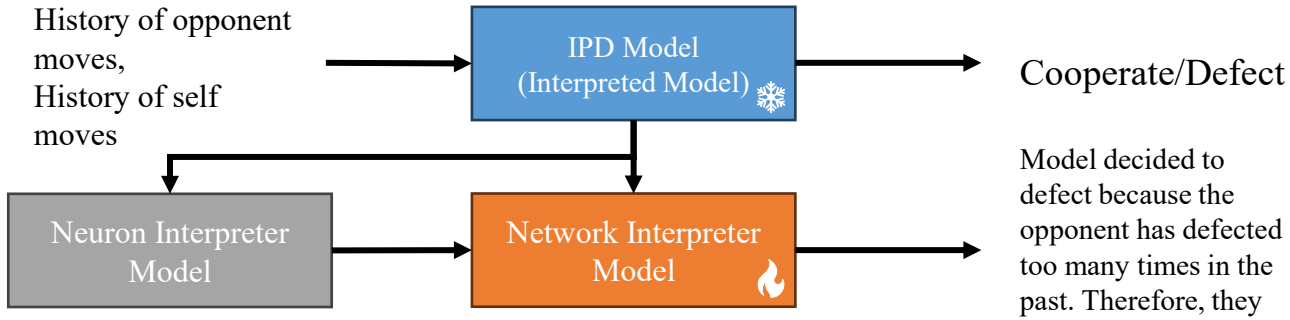
- *Interpretability using Large Language Models.* (mechanistic interpretability with LLMs)

Overview of Methods

Neuronal Level
(most existing methods)



Network Level
(our new method)



Takes in activations of individual neurons and the functions of these neurons created by the *neuron interpreter model* and outputs a holistic summary/decision about what the model is "thinking'

Neuronal Level

For a given neuron, can we mechanistically generate a natural language summary of the feature(s) it is capturing?

Experiment 1: Replicating Bills et. al (2023)

Assumptions.

Each neuron “looks at” exactly one *feature*.

https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html

Experimental Variables.

Independent	**Control**	**Dependent**
Training objective Layer size (1-layer neural networks) Training data Interpreter language model Aggregation method	Input data used for probing Interpretation granularity	Interpretation generated for each neuron

Methods.

Run the input data through the neural network and collect the activations at each layer. Generate an activation profile of each neuron. That is, perform a weighted average of the inputs based on the activation of the neuron. Perform few-shot prompting with the interpreter language model and ask the model to “characterize the neuron.”

Instead of performing weighted average over all of the inputs, first cluster the inputs, check which cluster has the strongest activation, then perform a weighted average in this class.

Use the maximum activated input instead a weighted average.

Experiment 1: Results

Overview.

Experiment 2: Inductively applying Bills et. al (2023)

Assumptions.

(1) Each neuron "looks at" exactly one *feature*. (2) Neurons have Markov property. That is, only depend on the features in the previous layer.

Experimental Variables.

Independent	**Control**	**Dependent**
Training objective Layer configurations Training data Interpreter language model Aggregation method	Input data used for probing Interpretation granularity Interpretations for neurons in the "previous layer."	Interpretation generated for each neuron

Methods.

Using the method of Bills et al. (2023) generate interpretations for the first layer. Then, use these interpretations as inputs and Bills et al. (2023) to generate interpretations for the next layer. Apply the same method as before.

The neuron in the last layer should be able to "summarize" the decision of the entire network. We expect that the interpretations to get more and more abstract.

Experiment 2: Results

Overview.

Experiment 3: Prototype/Modular Networks

Assumptions.

(1) Each neuron "looks at" exactly one *feature*. (2) Neurons have Markov property. That is, only depend on the features in the previous layer.

Experimental Variables.

Independent	**Control**	**Dependent**
Training objective Layer configurations Training data Interpreter language model Aggregation method	Input data used for probing Interpretation granularity	Interpretation generated for each neuron

Methods.

Run the input data through the neural network and collect the activations at each layer. Generate an activation profile of each neuron. That is, perform a weighted average of the inputs based on the activation of the neuron. Perform few-shot prompting with the interpreter language model and ask the model to "characterize the neuron."

Instead of performing weighted average over all of the inputs, first cluster the inputs, check which cluster has the strongest activation, then perform a weighted average in this class.

Use the maximum activated input instead a weighted average.

Experiment 3: Results

Overview.

Network Level

For a given network, can we mechanistically generate a natural language rationale for its decision?

Experiment 1: Multiclass Classification

Assumptions.

(1) There exists a consist rationale by which the network decides, across the data distribution. (2)

Experimental Variables.

Independent	**Control**	**Dependent**
Training objective Layer size (1-layer neural networks) Training data Interpreter language model Aggregation method	Input data used for probing Interpretation granularity	Interpretation generated for each neuron

Methods.

Run the input data through the neural network and collect the activations at each layer. Generate an activation profile of each neuron. That is, perform a weighted average of the inputs based on the activation of the neuron. Perform few-shot prompting with the interpreter language model and ask the model to “characterize the neuron.”

Instead of performing weighted average over all of the inputs, first cluster the inputs, check which cluster has the strongest activation, then perform a weighted average in this class.

Use the maximum activated input instead a weighted average.