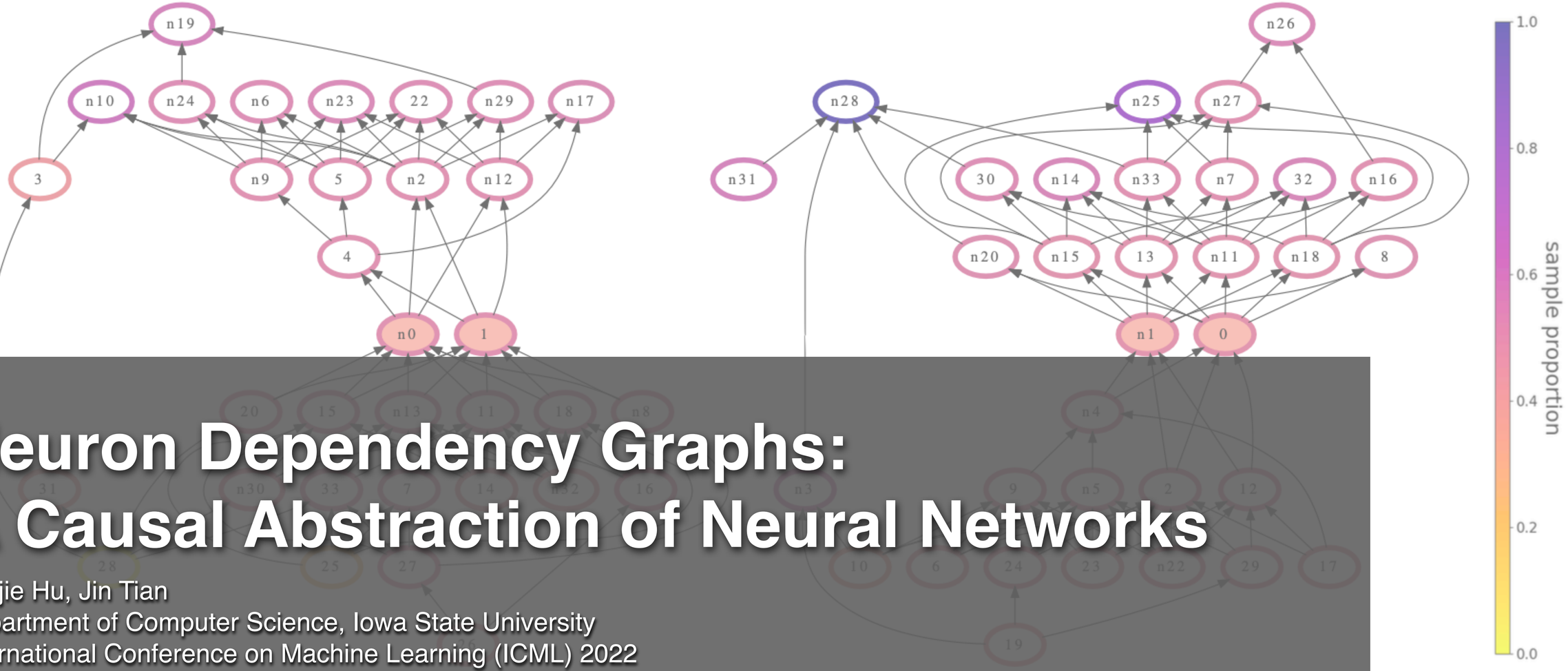
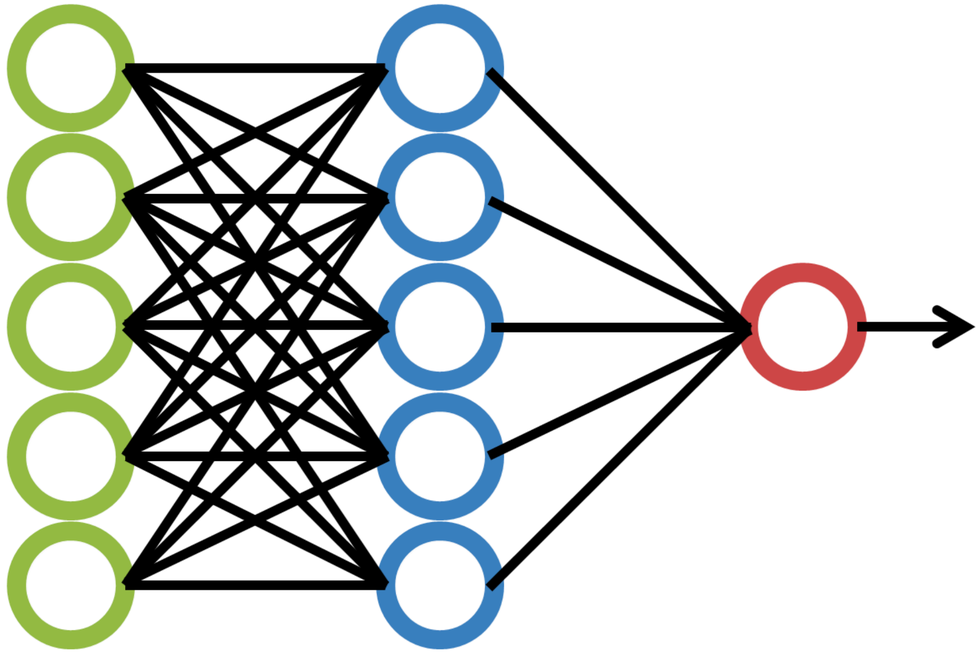


Neuron Dependency Graphs: A Causal Abstraction of Neural Networks

Yaojie Hu, Jin Tian
Department of Computer Science, Iowa State University
International Conference on Machine Learning (ICML) 2022

Problem Background & Goal

- Difficult to interpret
  - What is the meaning of each neuron?
  - How do neurons relate to each other?
  - Which neurons would cause the final output of the network to predict a certain class c?
- NDGs can (approximately) answer e.g.
  - A neuron N is some sufficient or necessary condition for a certain class c.
  - A neuron N_i logically implies neuron N_j
  - The necessary neurons of sufficient neurons would cause the network to predict class c



An illustration of a neural network (source: *Wikimedia)*

Neuron Dependency Graph. Hu and Tian. ICML 2022.

ICML International Conference On Machine Learning

Neuron Dependency Graphs (NDG)

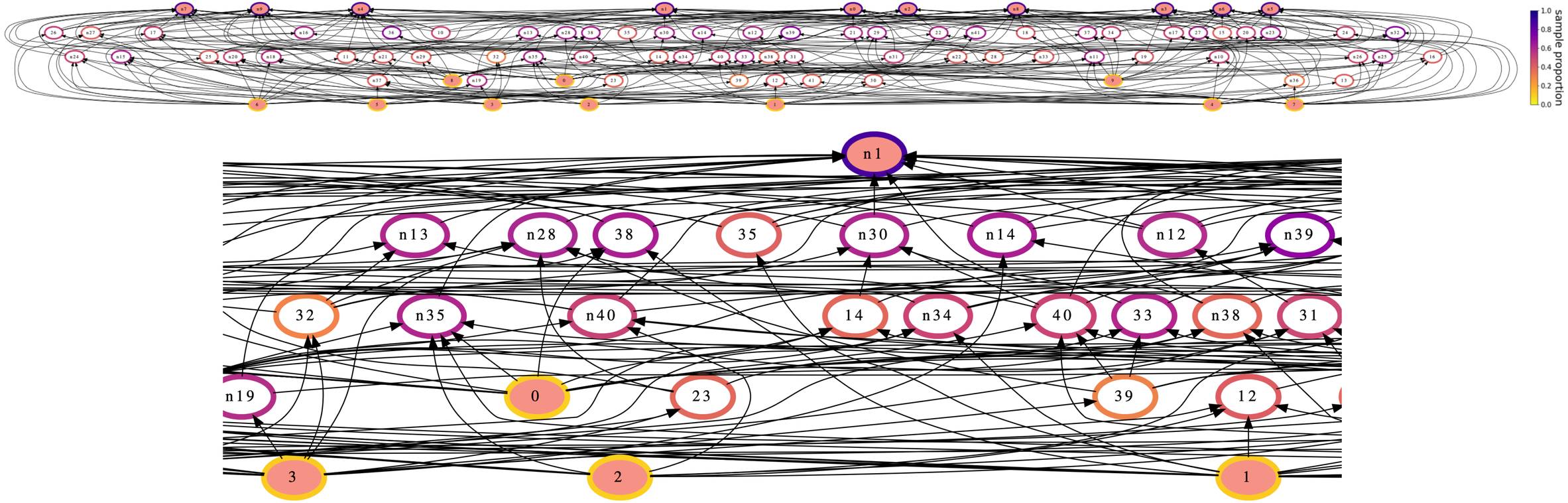


Figure 1 in paper. A neuron dependency graph for a convolutional neural network (CNN) trained on MNIST dataset.

- We discover that neural networks exhibit approximate **logical dependencies** among neurons, and we introduce Neuron Dependency Graphs (NDG) that extract and present them as directed graphs.

Neuron Dependency Graph. Hu and Tian. ICML 2022.

Neuron Dependency Graphs (NDG)

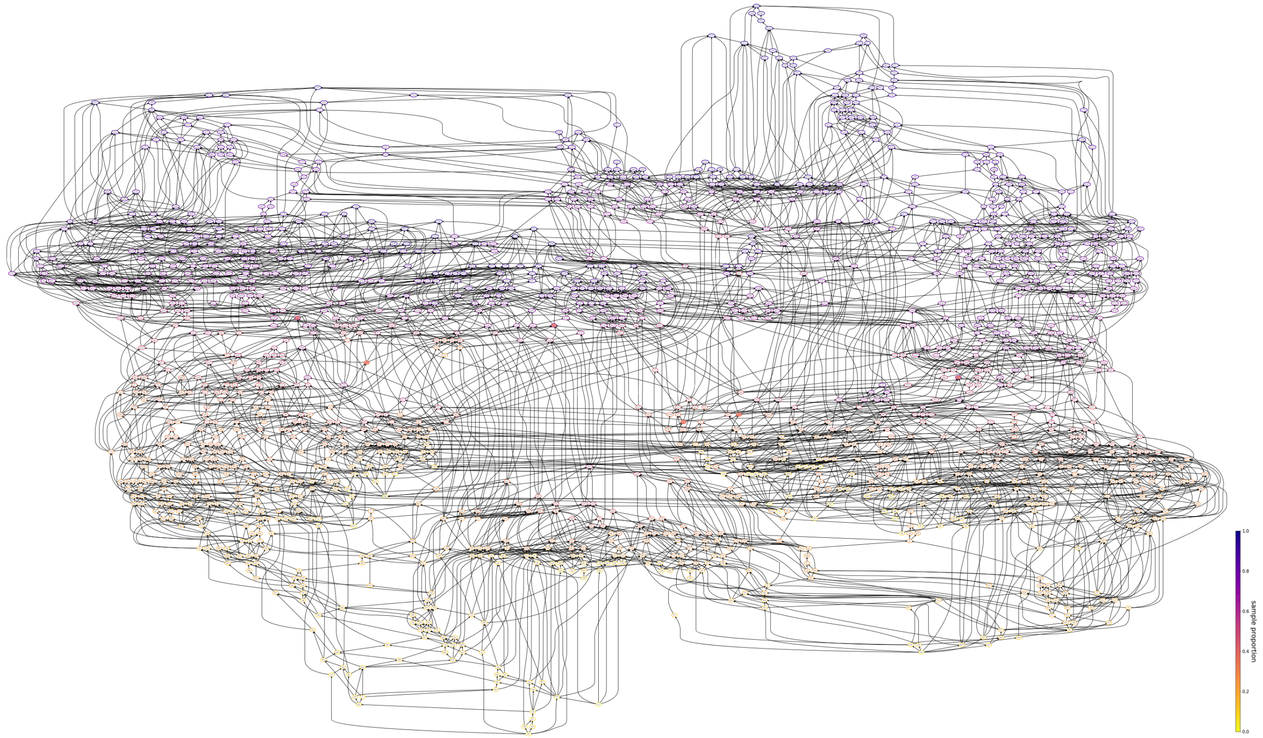


Figure 15 in paper. A neuron dependency graph for a Transformer (DistilRoBERTa) trained on AllNLI dataset.

- We discover that neural networks exhibit approximate **logical dependencies** among neurons, and we introduce Neuron Dependency Graphs (NDG) that extract and present them as directed graphs.

Neuron Dependency Graphs (NDG)

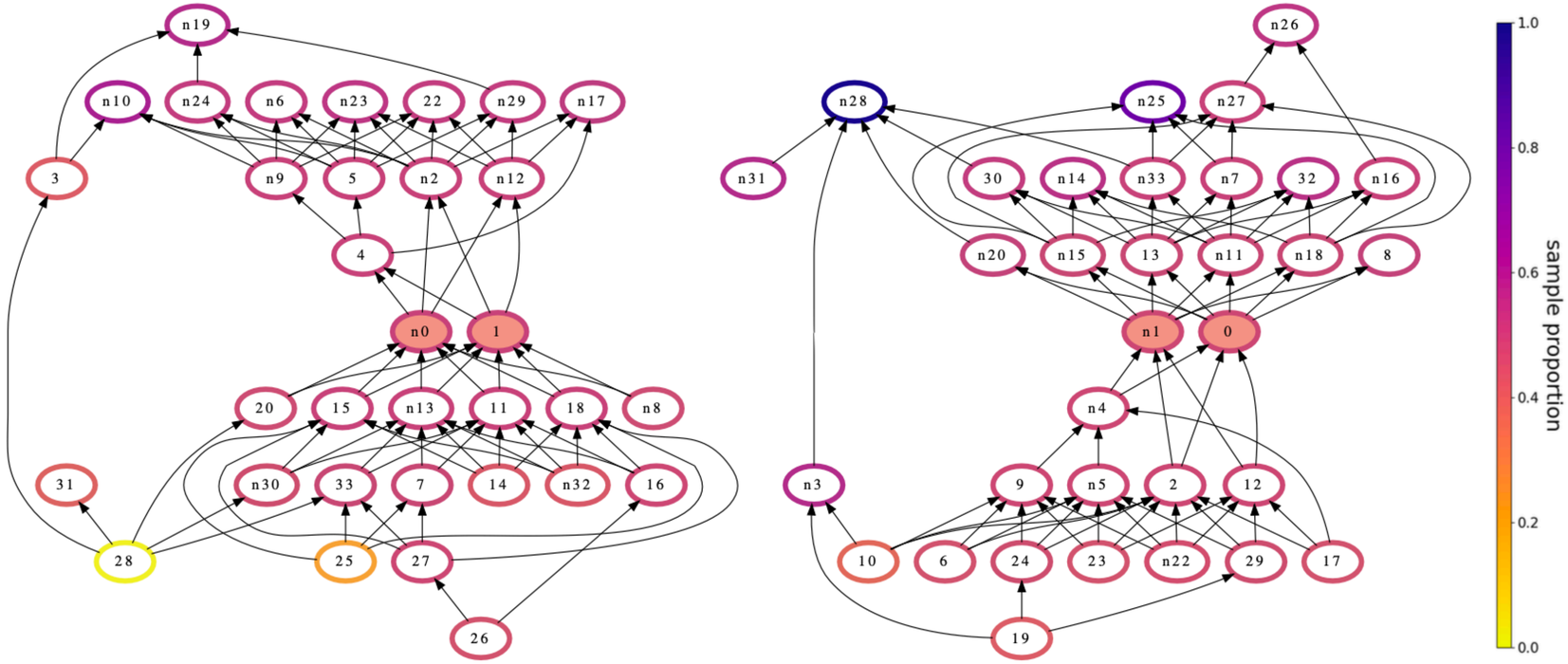


Figure 2 in paper. A neuron dependency graph for a convolutional neural network (CNN) trained on MNIST dataset to classify whether the digit is even or odd.

- We discover that neural networks exhibit approximate **logical dependencies** among neurons, and we introduce Neuron Dependency Graphs (NDG) that extract and present them as directed graphs.

Neuron Dependency Graph. Hu and Tian. ICML 2022.

Neuron Dependency Graphs (NDG)

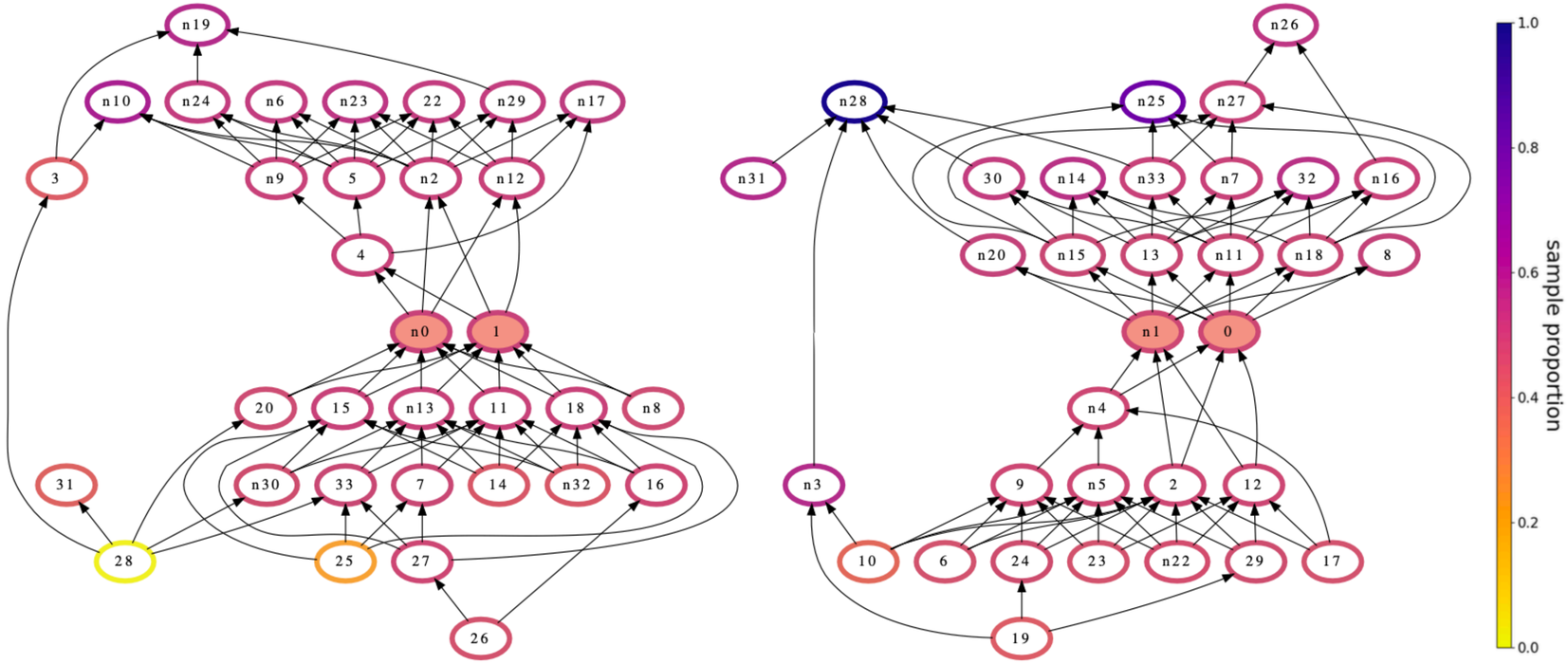


Figure 2 in paper. A neuron dependency graph for a convolutional neural network (CNN) trained on MNIST dataset to classify whether the digit is even or odd.

- In an NDG, each node corresponds to the Boolean activation value of a neuron, and each edge models an approximate logical implication from one node to another.

Neuron Dependency Graph. Hu and Tian. ICML 2022.

Neuron Dependency Graphs (NDG)

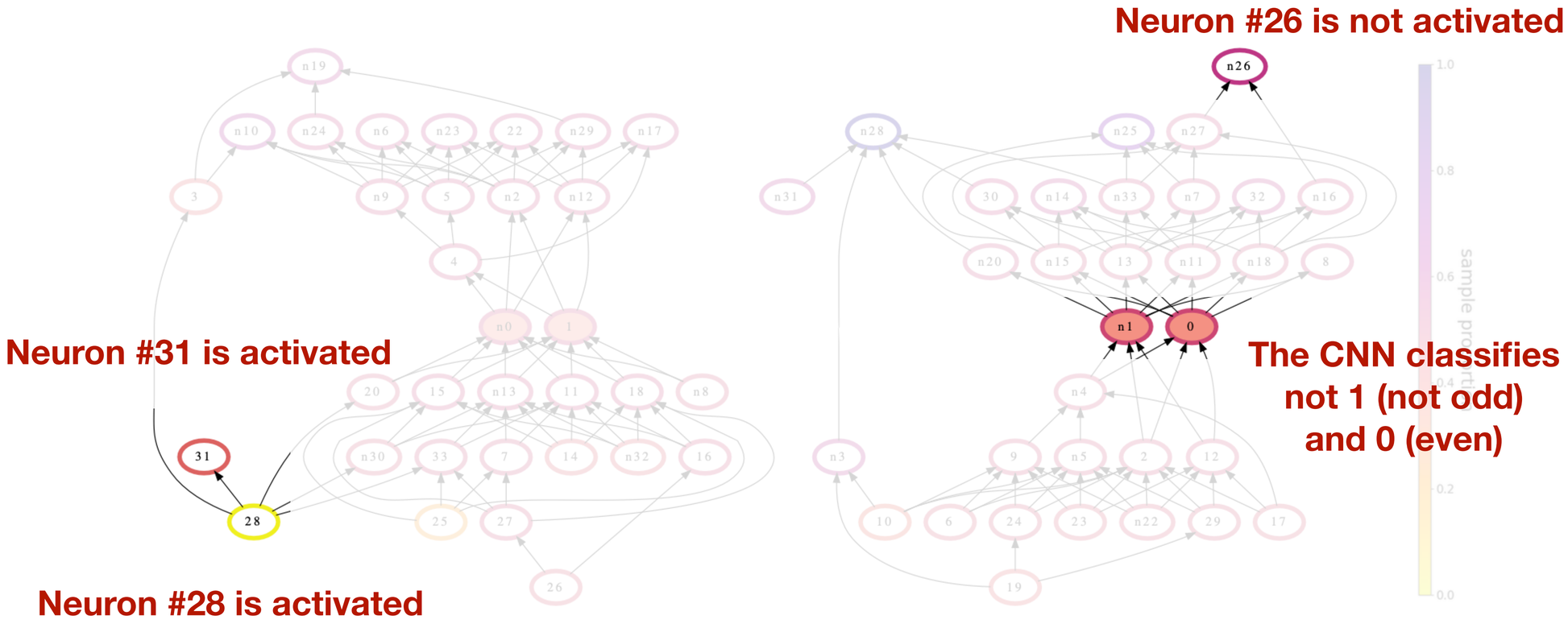


Figure 2 in paper. A neuron dependency graph for a convolutional neural network (CNN) trained on MNIST dataset to classify whether the digit is even or odd.

- In an NDG, each node corresponds to the Boolean activation value of a neuron, and each edge models an approximate logical implication from one node to another.

Neuron Dependency Graph. Hu and Tian. ICML 2022.

Neuron Dependency Graphs (NDG)

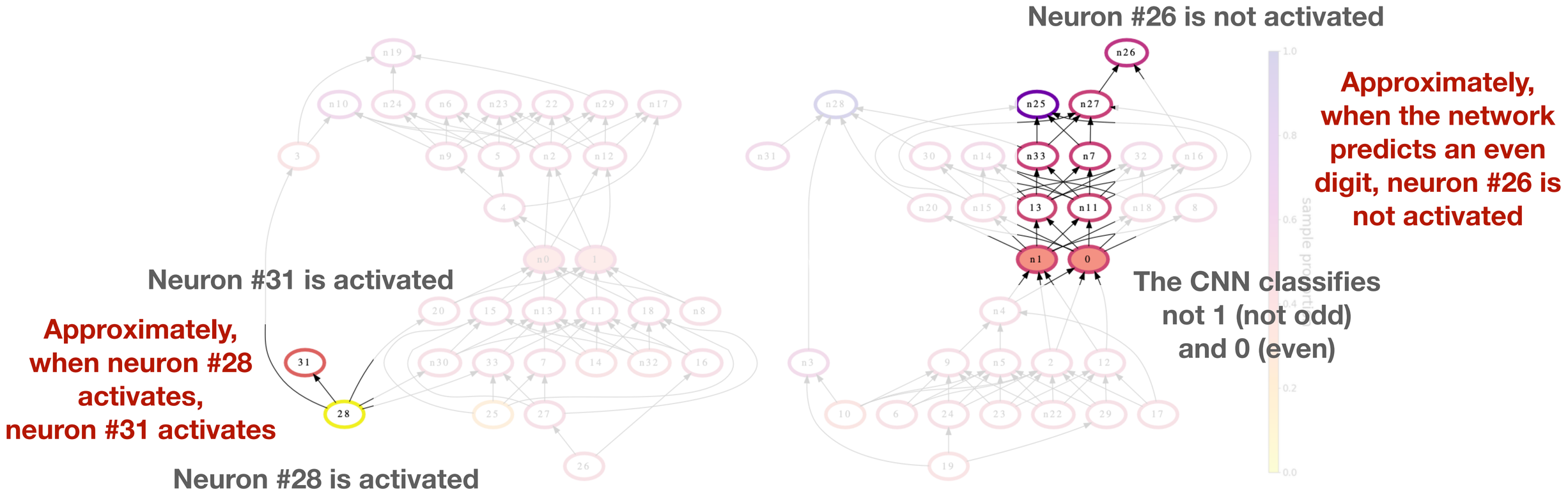


Figure 2 in paper. A neuron dependency graph for a convolutional neural network (CNN) trained on MNIST dataset to classify whether the digit is even or odd.

- In an NDG, each node corresponds to the Boolean activation value of a neuron, and each edge models an approximate logical implication from one node to another.

High Test Accuracy

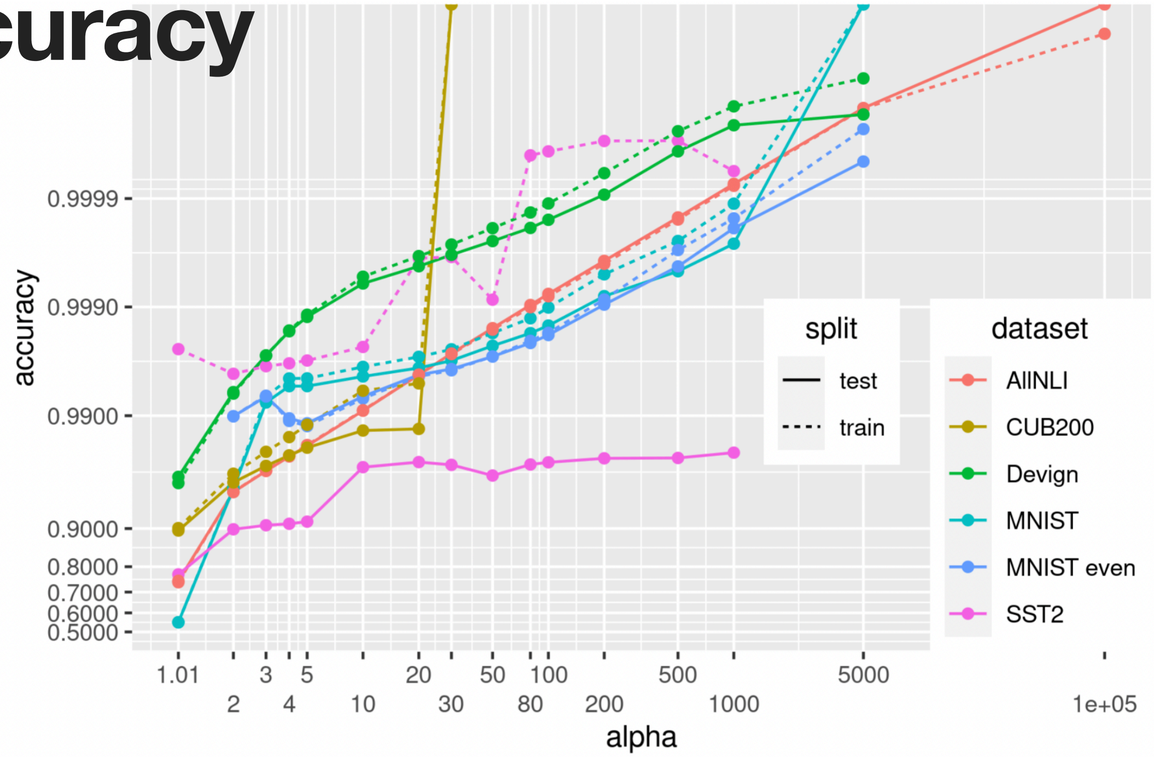


Figure 3. Average training and test accuracy for non-mutual-dependency edges versus various alpha thresholds. The x-axis is log transformed and y-axis is logistically transformed.

- We show that the logical dependencies extracted from the training set **generalize** well to the test set.
- Consistently for various models (including Transformer and CNN) and datasets (including image, natural language text, and programming language code).

Neuron Dependency Graph. Hu and Tian. ICML 2022.

Neuron Dependencies exist between two layers

Table 6. The number of edges, average training and test accuracy when neurons are selected from two layers in the same model.

	layer 1			layer 2			inter-layer		
	edges	train	test	edges	train	test	edges	train	test
MNIST	4172	99.93	99.90	10648	99.95	99.92	4986	99.92	99.89
MNIST even	790	99.94	99.95	7002	99.99	99.99	1088	99.88	99.90
SST2	529394	100.00	99.62	1877060	99.98	99.63	997980	100.00	99.74
AllNLI	747115	99.52	99.50	2186033	99.58	99.55	1087438	99.58	99.56
CUB200	900258	97.89	97.16	902210	97.89	97.16	900402	97.89	97.16
Devign	2191519	99.94	99.94	7952771	99.91	99.91	3430234	99.94	99.93

- Neuron Dependencies exist between **two layers of the same trained model**, and they generalize on test set.

Neuron Dependency Graph. Hu and Tian. ICML 2022.

Neuron Dependencies exist between two models

Table 4. The number of inter-model neuron dependency edges and the average accuracy when two models independently trained on the same dataset are given the same inputs, and when the same model is given two different inputs.

	different models same input			same model different inputs		
	edges	train acc	test acc	edges	train acc	test acc
MNIST	0	-	-	0	-	-
MNIST even	27	99.71	99.75	0	-	-
SST2	35354	100.00	92.29	0	-	-
AllNLI	138406	99.48	99.43	0	-	-
CUB200	0	-	-	0	-	-
Devign	75597	99.90	99.92	0	-	-

- Generalizable Neuron Dependencies exist between **two independently trained models.**

Neuron Dependency Graph. Hu and Tian. ICML 2022.

Neuron Dependencies exist for trained models

Table 3. The number of edges, average training and test accuracy for graph extracted from random models with real inputs and trained model with random inputs.

	original			random model			random input		
	edges	train	test	edges	train	test	edges	train	test
MNIST	524	99.90	99.85	0	-	-	22	99.94	99.89
MNIST even	686	99.83	99.82	0	-	-	316	99.95	99.95
SST2	29132	100.0	97.35	0	-	-	464024	99.92	99.92
AllNLI	202250	99.58	99.58	2686	99.82	48.04	149116	99.68	99.55
CUB200	195998	97.87	97.15	237286	96.43	50.15	748	95.76	91.80
Devign	473058	99.97	99.96	4200	99.64	53.01	26786	99.75	99.99

- Generalizable Neuron Dependencies appear **only when the model is trained.**
  - Trained models have neuron dependencies even with random inputs

Neuron Dependency Graph. Hu and Tian. ICML 2022.

Causality

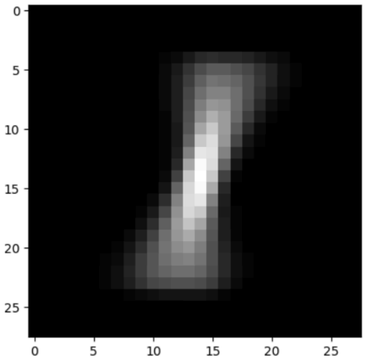
Table 5. After interchange intervention on the neural network, the percentage of aligned predictions $f^{i_{NN,c}}(\mathbf{x}) = c$, unchanged predictions $f^{i_{NN,c}}(\mathbf{x}) = y$, and unaligned predictions $f^{i_{NN,c}}(\mathbf{x}) \notin \{y, c\}$. Contradiction is when $N \leftarrow T_\phi$, $N \leftarrow F_\phi \in i_{NN,c}$ for some neuron N, and N is not intervened when it occurs. The average contradictions per input is reported. Interchange intervention empirically validates NDGs as a causal abstraction of neural networks.

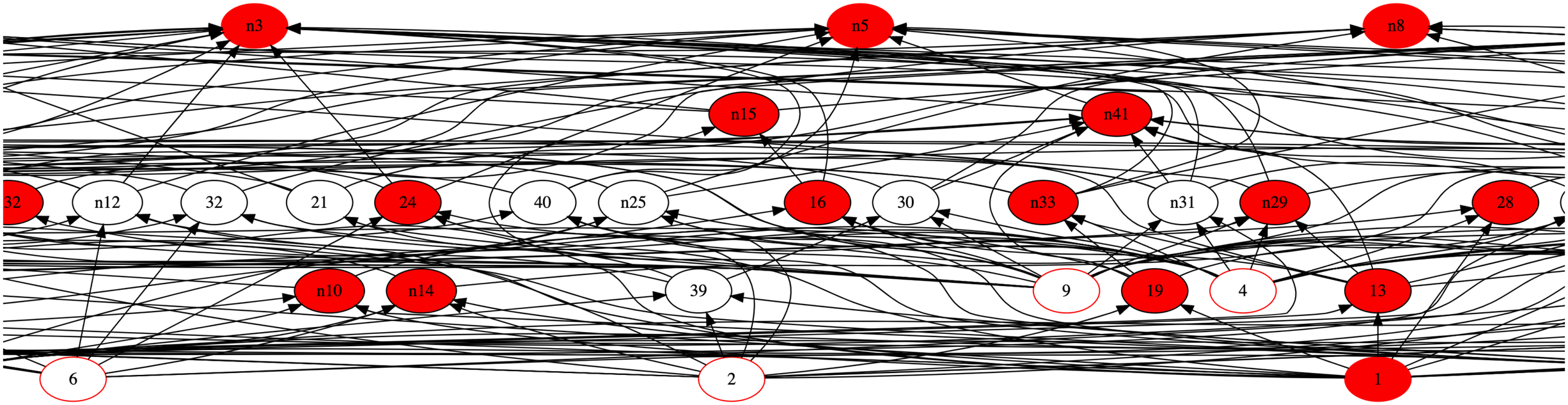
	aligned		unchanged		unaligned		model accuracy		contradictions	
	train	**test**	train	test	train	test	train	test	train	test
MNIST	**93.41**	**95.21**	5.53	3.78	1.05	1.01	99.48	98.93	0.00	0.00
MNIST even	**100.00**	**100.00**	0.00	0.00	0.00	0.00	99.28	99.17	0.00	0.00
SST2	**100.00**	**100.00**	0.00	0.00	0.00	0.00	100.00	90.00	0.00	0.00
AllNLI	**100.00**	**100.00**	0.00	0.00	0.00	0.00	76.28	72.05	14.02	14.12
CUB200	**89.38**	**88.39**	2.26	1.77	8.36	9.84	92.63	87.79	0.01	0.02
Devign	**100.00**	**100.00**	0.00	0.00	0.00	0.00	78.78	60.86	102.12	102.04

- Based on Neuron Dependency Graphs extracted, we intervene on the neural network's activations and successfully cause the network to produce a counterfactual classification c.

Neuron Dependency Graph. Hu and Tian. ICML 2022.

Causality: How to intervene?

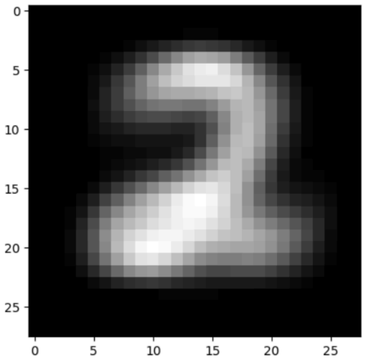


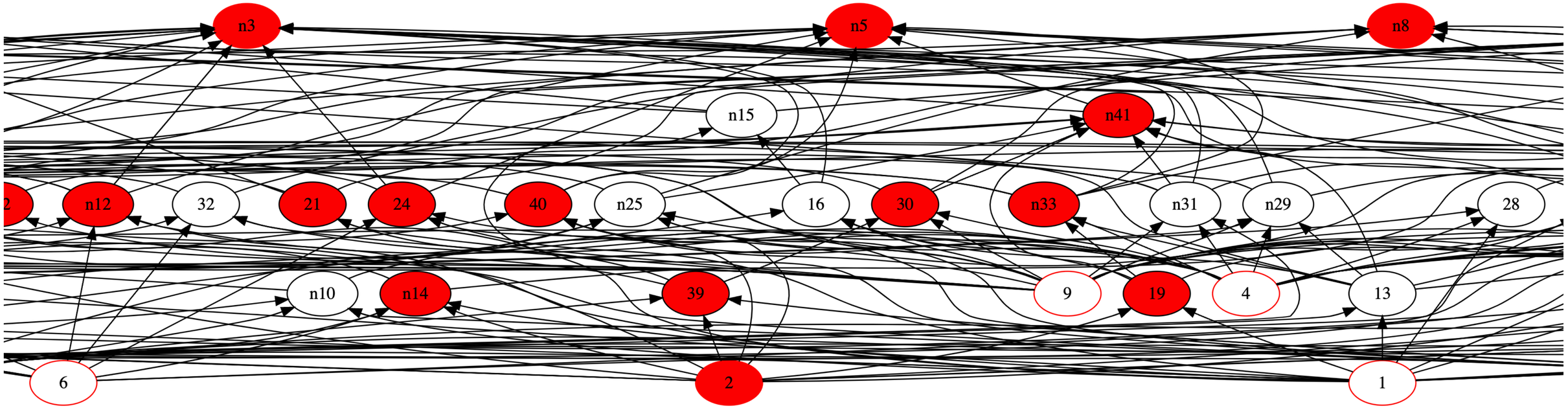


- How to change the activations to cause the neural network to predict 2?

Neuron Dependency Graph. Hu and Tian. ICML 2022.

Causality: How to intervene?





- Activate necessary neurons of sufficient neurons of class 2.

Neuron Dependency Graph. Hu and Tian. ICML 2022.

Causality

Table 5. After interchange intervention on the neural network, the percentage of aligned predictions $f^{i_{NN,c}}(\mathbf{x}) = c$, unchanged predictions $f^{i_{NN,c}}(\mathbf{x}) = y$, and unaligned predictions $f^{i_{NN,c}}(\mathbf{x}) \notin \{y, c\}$. Contradiction is when $N \leftarrow T_\phi$, $N \leftarrow F_\phi \in i_{NN,c}$ for some neuron N, and N is not intervened when it occurs. The average contradictions per input is reported. Interchange intervention empirically validates NDGs as a causal abstraction of neural networks.

	aligned		unchanged		unaligned		model accuracy		contradictions	
	train	**test**	train	test	train	test	train	test	train	test
MNIST	**93.41**	**95.21**	5.53	3.78	1.05	1.01	99.48	98.93	0.00	0.00
MNIST even	**100.00**	**100.00**	0.00	0.00	0.00	0.00	99.28	99.17	0.00	0.00
SST2	**100.00**	**100.00**	0.00	0.00	0.00	0.00	100.00	90.00	0.00	0.00
AllNLI	**100.00**	**100.00**	0.00	0.00	0.00	0.00	76.28	72.05	14.02	14.12
CUB200	**89.38**	**88.39**	2.26	1.77	8.36	9.84	92.63	87.79	0.01	0.02
Devign	**100.00**	**100.00**	0.00	0.00	0.00	0.00	78.78	60.86	102.12	102.04

- In addition to providing symbolic explanations to the neural network's internal structure, NDGs can represent a Structural Causal Model. We **empirically** show that an NDG is a **causal abstraction** of the corresponding neural network that “unfolds” the same way under causal interventions using the theory by (Geiger et al. 2021).