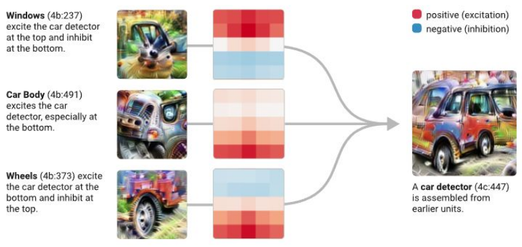
Automatic Discovery of Visual Circuits

Q: How are intermediate computations conducted in vision models?

Previous Work: Manually Constructed Circuits that Compose Concepts

(Figure from Olah et. al 2020)



Conmy et al. 2023 proposed **ACDC** for circuit detection in language models

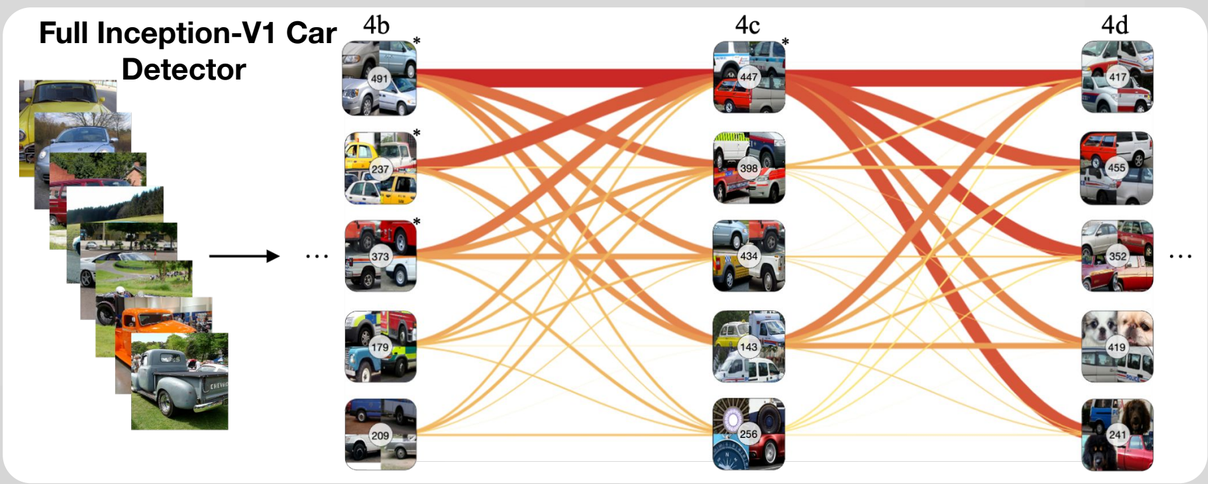
How do we automatically detect circuits in vision models?

Ablations on Circuits Allow for Targeted, Causal Interventions

- Circuits allow for removal of entire **paths of influence**
- We introduce two variants:
  - **Edge Pruning**: corrupting all connections between the first and second layers of the circuit
  - **Circuit Pruning:** removing all neurons in a circuit

Building Circuits from Connectivity Graphs

1. **Select** initial neurons **layer-by-layer,** maximizing the sum of their attribution scores to adjacent layers
2. **Refine** the neurons, maximizing the sum of attribution scores **within the circuit**



Redefining Functional Connectivity: Cross-Layer Attribution (CLA)

1. Select circuit by specifying an **input distribution** of images
2. Compute **attribution matrix** from input distribution

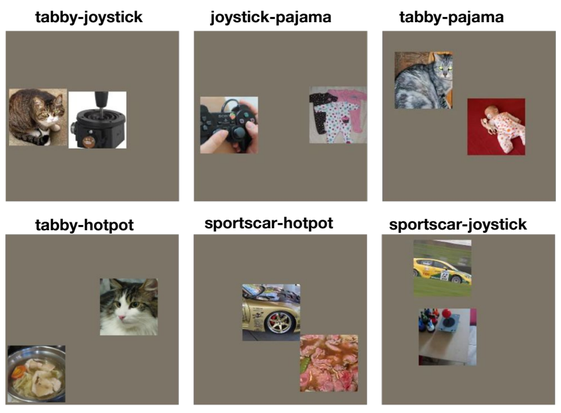
Cross-Layer Attribution Matrix:

$$attr[m,n] \leftarrow |a_{i,m}| \cdot \frac{\partial |a_{i+1,n}|}{\partial a_{i,m}}$$

Attribution Score · **Activation (relevance)** · **Gradient (effect size)**

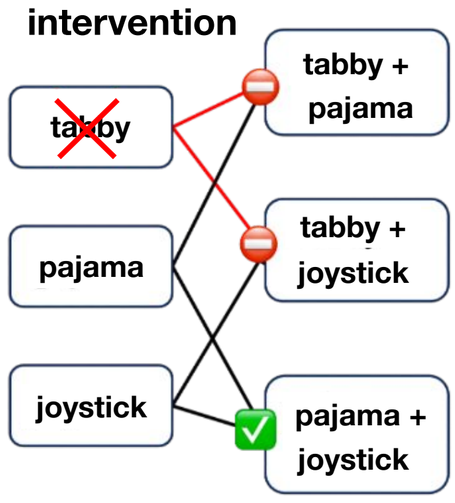
Circuits Implement Visual Feature Hierarchy

Q: How do you train models with known intermediate features?



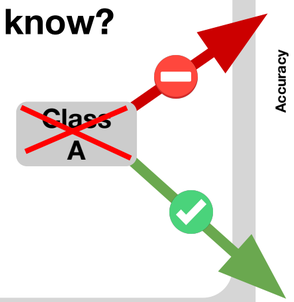
A: The CatFish Dataset!

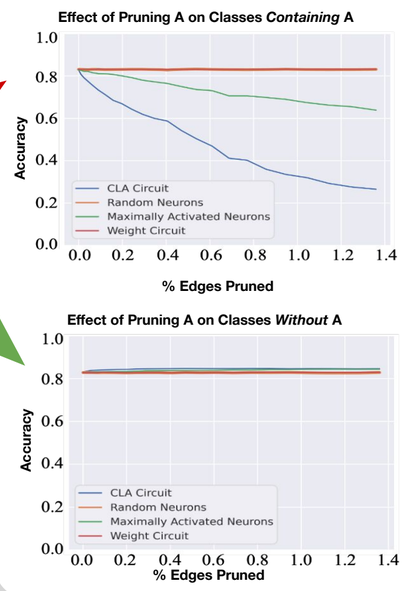
Subclass-level intervention



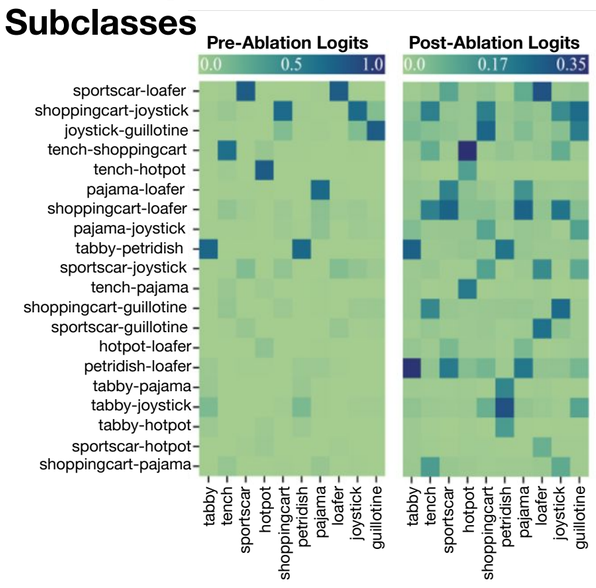
Q: Are circuits actually used for model predictions? A: Yes! But how do we know?

- We can "prune" circuits, blocking **information flow** to later layers
- circuits have **selective effects** on only **relevant** prediction decisions





Probability Mass Spreads Across Subclasses



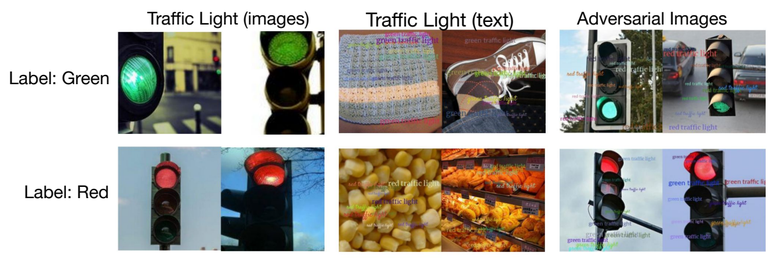
Q: What does a model "see" after circuit ablation?

A: Well, it certainly **<u>doesn't</u>** see the relevant subclass

Inception-CatFish assigns **approximately equal probability** to all classes containing the complement of the ablated subclass

Circuit Pruning Protects CLIP from Adversarial Textual Attacks

Benchmarking Textual Defense: the Traffic Light Dataset

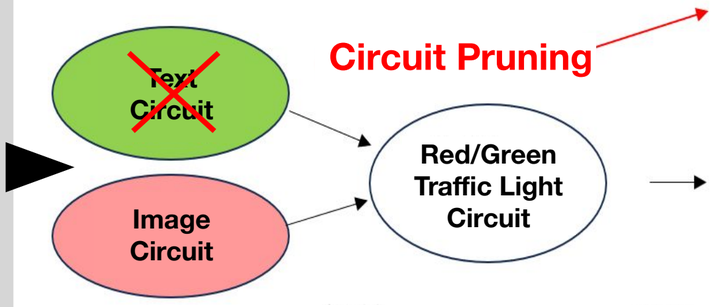


- Using **CLIP** to label traffic lights based on their color (**red**/**green**)
- **Multimodal Neurons** in CLIP detect **both images and text**

Q: Can CLA Disentangle these Capabilities?

- **Two proposed subcircuits:** an **image detector** that detects traffic lights, and a seperate **text detector**
- We find the **text detector** using **CLA**, and then use **circuit pruning** to remove it

CLIP Encoder



Patched Probabilities

Red Traffic Light	**0.874**
Green Traffic Light	0.126

Original Probabilities

Red Traffic Light	0.036
Green Traffic Light	**0.964**

Textual Defense with Circuit Pruning

- **Layer Choice Matters:** Multimodal Composition is **Localized**
- **Model Edits** needed for Defense are **Minimal**

CLIP **improves from 3% to 87% accuracy** on adversarial images, after pruning **only 6% of the edges** in layer 3

TLDR

1. **automatic visual circuit extraction**
2. neuron relevancy + downstream effect = **functional connectivity**
3. causal interventions on circuits → **predictable changes to model behaviour**

Q: What are the limitations?

- Only one allowed **circuit topology** —dense circuits with a set number of neurons per layer
- Requires a well-defined **input image distribution** to perform discovery

Q: Any next steps?

- Generalize CLA to **arbitrary circuit structures** (sparsification)
- Unsupervised **"dissection"** into several circuits
- **Text Based Detection**/Automatic Circuit Description

Q: What Could this be used for IRL?

- Locate and Remove Circuits corresponding to **Unwanted Behaviors**
- Understand **visual feature hierarchy** of large models
- End-to-End **model dissection**

References

Unless, otherwise noted, images are my own

[1] Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom In: An Introduction to Circuits.
[2] Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). Towards Automated Circuit Discovery for Mechanistic Interpretability
[3] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). Going Deeper with Convolutions
[4] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2014). ImageNet Large Scale Visual Recognition Challenge
[5] Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., & Olah, C. (2021). Multimodal Neurons in Artificial Neural Networks