We can “prune” circuits, but how do we automatically detect circuits in vision models?

A: Yes! But how do we know?

### Circuits Implement Visual Feature Hierarchy

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

**A:** We can “prune” circuits, blocking information flow to later layers. We introduce two variants:

- **Edge Pruning:** removing 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.

**Ablations on Circuits Allow for Targeted, Causal Interventions**

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

**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.
- **Multimodal Neurons in CLIP detect both images and text.**
- Can CLA Disentangle these Capabilities?

**CLIP Encoder**

**Red/ Green Traffic Light Circuit**

- **Adversarial Traffic Light:**
  - **Image Circuit:**
    - **Patching Probabilities**
      - Red Traffic Light: 0.974
      - Green Traffic Light: 0.026
  - **Original Probabilities**
    - Red Traffic Light: 0.036
    - Green Traffic Light: 0.964

**Adversarial Image Circuit**

**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.

**References**


