Localizing Task Performance in LMs

To localize a task, we find a circuit: the minimal computational subgraph that preserves LM behavior when corrupting all other edges.

Prior Circuit Finding Methods

Past approaches found a singular circuit using patching. Newer approaches score edge importance, then find a circuit using those scores; you can choose the circuit size. But how to get scores?

- **Activation Patching** [1] patches an edge’s corrupted activation into a clean forward pass. This requires \(O(edges)\) passes.
- **EAP** [2]: approximates the impact of ablating an edge \((u,v)\) as
  \[
  (z'_u - z_u)^+ \nabla_v L(s),
  \]
  where \(z_u/z'_u\) are u’s clean / corrupted acts; \(L(s)\) is the loss when running the model on a clean example. This requires \(O(1)\) passes.

Integrated-Gradients-Based Methods

Integrated gradients [3] is a technique like EAP that attempts to find important parts of the inputs. It improves inputs \(x\) gradients by interpolating between inputs when computing gradients.

\[
(z'_u - z_u) \int_{a=0}^{1} \frac{\partial M(z'_u + a(z' - z_u))}{\partial a} = (z'_u - z_u) \int_{a=0}^{1} \frac{\partial M(z'_u + a(z' - z_u))}{\partial z'_u} \frac{\partial z'_u}{\partial a}.
\]

How can we use this technique for circuit finding?

- **EAP-IG (Inputs, ours)**: Average grads over input interpolation; \(O(m)\)
  \[
  (z'_u - z_u) \int_{a=0}^{1} \frac{\partial M(z'_u + a(z' - z_u))}{\partial a} = (z'_u - z_u) \int_{a=0}^{1} \frac{\partial M(z'_u + a(z' - z_u))}{\partial z'_u} \frac{\partial z'_u}{\partial a}
  \]
- **EAP-IG (Activations)** [4]: Average grads over interpolation between activations—in \(O(m^2)\) passes.
  \[
  (z'_u - z_u) \int_{a=0}^{1} \frac{\partial M(z'_u + a(z' - z_u))}{\partial a} = (z'_u - z_u) \int_{a=0}^{1} \frac{\partial M(z'_u + a(z' - z_u))}{\partial z'_u} \frac{\partial z'_u}{\partial a}
  \]
- **EAP-IG (Activations)** [5]: Try to do the former in \(O(m)\) passes.

- **Clean-Corrupted**: Average grads on clean / corrupted inputs; \(O(1)\)

### Tasks

**Indirect Object Identification (IOI, 7)**: When John and Mary went to the store, \(<John/Alice> gave a drink to [Mary/John]
**Greater-Than** [8]: The war lasted from the year \(<1741/1701>\) to the year \(1702/42\)
**Country-Capital**: [France/Italy], whose capital, [Paris/Rome]
**Gendered Pronouns**: The <nurse/doctor> said that [she/he]
**Hypernymy**: <Roses/diamonds> and other [flowers/gems]

### Comparing Methods

We find circuits of varying sizes for these tasks in GPT-2 small. EAP-IG (in input or activation space) outperforms EAP. The Clean-Corrupted method is a strong baseline, often as good as EAP-IG.

### Inter-Task Comparison

We measure intersection over union (# overlapping nodes + total # nodes) & cross-task faithfulness (run one task on another’s circuit)

### Conclusions and Open Questions

- Alternative patching methods, like EAP-IG and Clean-Corrupted, outperform vanilla EAP at little to no extra cost
- Overlap and cross-task faithfulness disagree on circuit similarity
Still, questions remain:
- Which metric, if any, best quantifies how similar circuits are?
- Can we judge mechanistic similarity via component circuits alone?
- Are these techniques yielding complete circuits?

### References