Optimal ablation for interpretability

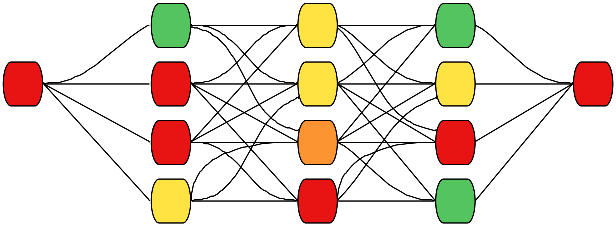
Maximilian Li and Lucas Janson

NeurIPS 2024 (Spotlight)

November 13, 2024

Interpreting neural networks

How important is a model component?



Motivating question

Define the *ablation loss gap* $\Delta(\mathcal{M}, \mathcal{A}) := \mathcal{P}(\mathcal{M}^{\setminus \mathcal{A}}) - \mathcal{P}(\mathcal{M})$.

What is the best performance on subtask $\mathcal{D}$ the model M could have achieved without component $\mathcal{A}$?

Motivating question

Define the *ablation loss gap* $\Delta(\mathcal{M}, \mathcal{A}) := \mathcal{P}(\mathcal{M}^{\setminus \mathcal{A}}) - \mathcal{P}(\mathcal{M})$.

What is the best performance on subtask $\mathcal{D}$ the model M could have achieved without component $\mathcal{A}$?

I. Performance on subtask $\mathcal{D}$ is measured via expected loss on the subtask, i.e. $\mathcal{P}(\tilde{\mathcal{M}}) = \mathbb{E}_{X \sim \mathcal{D}} \mathcal{L}(\tilde{\mathcal{M}}(X), \mathcal{M}(X))$.

Motivating question

Define the *ablation loss gap* $\Delta(\mathcal{M}, \mathcal{A}) := \mathcal{P}(\mathcal{M}^{\setminus \mathcal{A}}) - \mathcal{P}(\mathcal{M})$.

What is the best performance on subtask $\mathcal{D}$ the model M could have achieved without component $\mathcal{A}$?

I. Performance on subtask $\mathcal{D}$ is measured via expected loss on the subtask, i.e. $\mathcal{P}(\tilde{\mathcal{M}}) = \mathbb{E}_{X \sim \mathcal{D}} \mathcal{L}(\tilde{\mathcal{M}}(X), \mathcal{M}(X))$.

II. Model $\mathcal{M}$ could have achieved: $\mathcal{M}^{\setminus \mathcal{A}}$ is constructed solely by changing the value of $\mathcal{A}(X)$.

Motivating question

Define the *ablation loss gap* $\Delta(\mathcal{M}, \mathcal{A}) := \mathcal{P}(\mathcal{M}^{\setminus\mathcal{A}}) - \mathcal{P}(\mathcal{M})$.

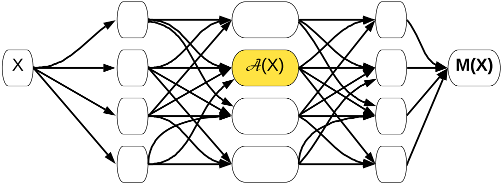
What is the best performance on subtask $\mathcal{D}$ the model M could have achieved without component $\mathcal{A}$?

I. Performance on subtask $\mathcal{D}$ is measured via expected loss on the subtask, i.e. $\mathcal{P}(\tilde{\mathcal{M}}) = \mathbb{E}_{X\sim\mathcal{D}}\, \mathcal{L}(\tilde{\mathcal{M}}(X), \mathcal{M}(X))$.

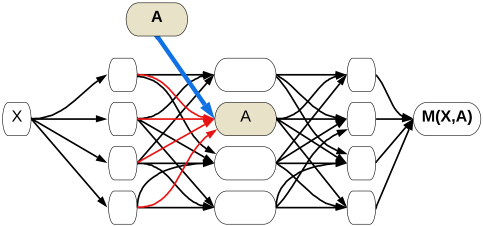
II. Model $\mathcal{M}$ could have achieved: $\mathcal{M}^{\setminus\mathcal{A}}$ is constructed solely by changing the value of $\mathcal{A}(X)$.

III. Without component $\mathcal{A}$: $\mathcal{M}^{\setminus\mathcal{A}}(x)$ uses a value for $\mathcal{A}$ that conveys no information about x.

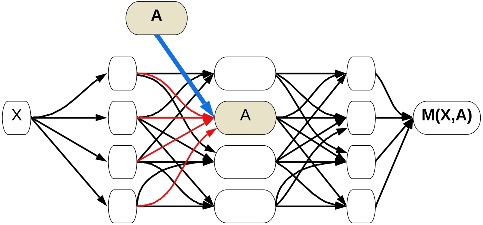
Example



Example

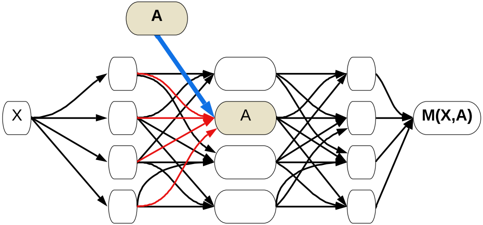


Example



Definition: A **total ablation method** satisfies
$\mathcal{M}^{\setminus\mathcal{A}}(X) = \mathcal{M}^{\setminus\mathcal{A}}(X, A)$ for $A \perp\!\!\!\perp X$.

Example



Zero ablation: $A = 0$.
Mean ablation: $A = \mathbb{E}_{X' \sim \mathcal{D}}[\mathcal{A}(X')]$
Resample ablation: $A = \mathcal{A}(X'), X' \perp\!\!\!\perp X$.

Motivating question

Define the *ablation loss gap* $\Delta(\mathcal{M}, \mathcal{A}) := \mathcal{P}(\mathcal{M}^{\setminus\mathcal{A}}) - \mathcal{P}(\mathcal{M})$.

What is the best performance on subtask $\mathcal{D}$ the model M could have achieved without component $\mathcal{A}$?

IV. "Best" performance: we want to understand how much performance degrades *because* we had to ablate $\mathcal{A}$.

Seeking best performance avoids interventions that "spoof" the model by causing it to confuse x for a different input, or treat x in a way that it never treated any training input.

Optimal ablation

Definition:

$$\mathcal{M}^{\backslash \mathcal{A}}_{(\text{opt})}(x) := \mathcal{M}_{\mathcal{A}}(x, a^*),$$
$$a^* := \arg\min_a \mathbb{E}_{X \sim \mathcal{D}} \, \mathcal{L}(\mathcal{M}_{\mathcal{A}}(X, a), \mathcal{M}(X))$$

Proposition

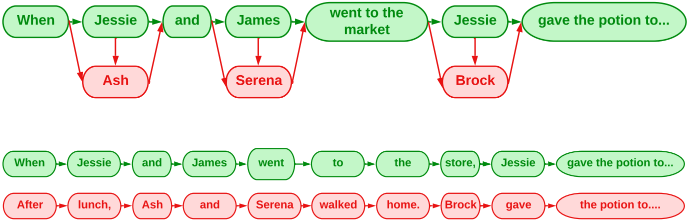
Let $\Delta(\mathcal{M}, \mathcal{A})$ be the ablation loss gap for some component $\mathcal{A}$ measured with any total ablation method. Then

$$\Delta_{\text{opt}}(\mathcal{M}, \mathcal{A}) \leq \Delta(\mathcal{M}, \mathcal{A})$$

Comparison to counterfactual ablation

Counterfactual ablation (CF) considers pairs of parallel inputs.

- CF requires manual effort for each subtask and may not be possible for complex subtasks. OA is more versatile than CF.



Comparison to counterfactual ablation

Counterfactual ablation (CF) considers pairs of parallel inputs.

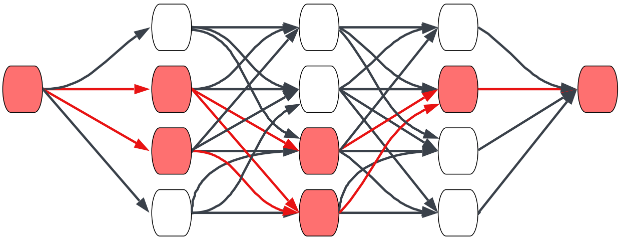
- CF removes *less* information than OA, yet still achieves higher loss, which is evidence that *most* loss can be attributed to spoofing.

Table 1: Comparison of ablation loss gap Δ on IOI

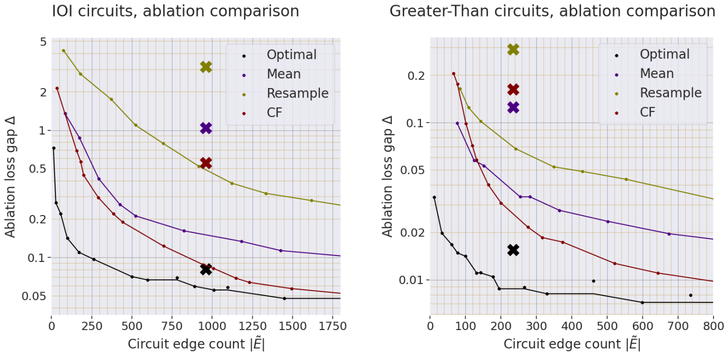
	Zero	**Mean**	**Resample**	**CF-Mean**	**Optimal**	**CF**
Rank correlation with CF	0.590	0.825	0.828	0.833	**0.907**	1
Median ratio of Δ_{opt} to Δ	11.1%	33.0%	17.7%	31.7%	100%	88.9%

Circuit discovery

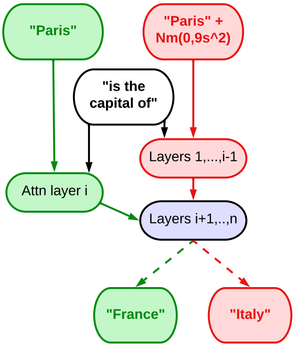
We introduce a *uniform gradient sampling* method to find circuits.



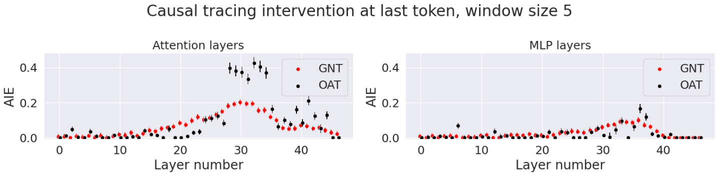
Circuit discovery results



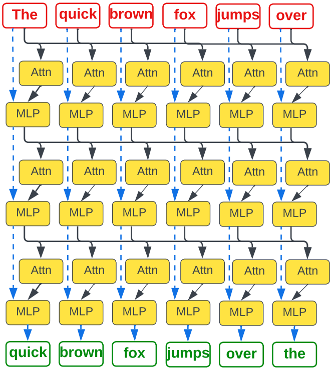
Causal tracing



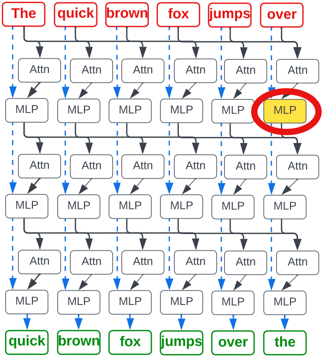
Causal tracing results



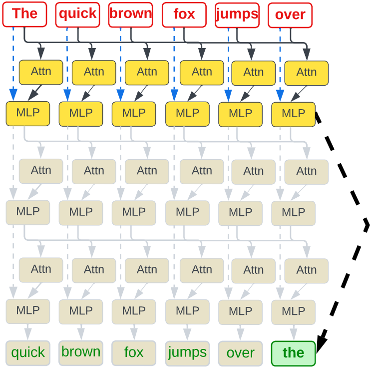
Latent prediction



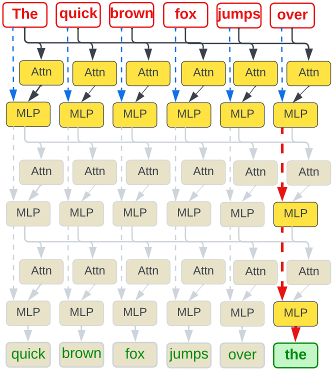
Latent prediction



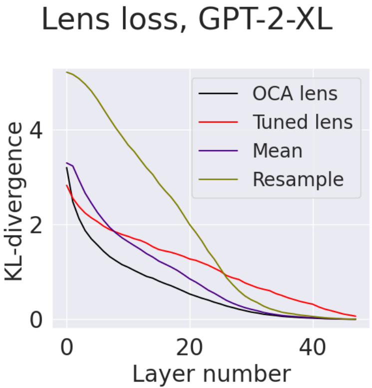
Latent prediction: tuned lens



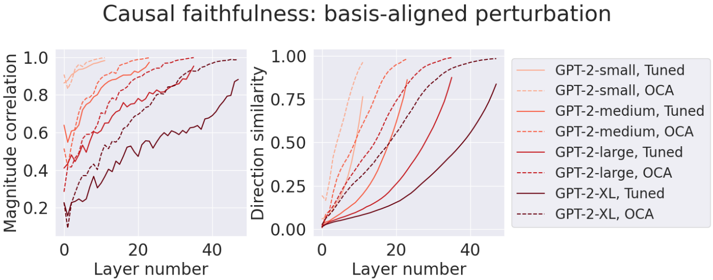
Latent prediction: Optimal Constant Attention (OCA lens)



Latent prediction results



Latent prediction: causal faithfulness



Latent prediction: truthful elicitation