Optimal ablation for interpretability Maximilian Li and Lucas Janson

NeurIPS 2024 (Spotlight)

November 13, 2024





・ロト・西ト・西ト・西・ うらの

Maximilian Li and Lucas Janson Optimal ablation for interpretability

Interpreting neural networks

How important is a model component?



Maximilian Li and Lucas Janson

◆□▶ ◆□▶ ◆三▶ ◆三▶ 三三 シののや

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 D the model M could have achieved without component A?

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 D the model M could have achieved without component A?

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

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 D the model M could have achieved without component A?

I. <u>Performance on subtask \mathcal{D} </u> 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. <u>Model \mathcal{M} could have achieved</u>: $\mathcal{M}^{\setminus \mathcal{A}}$ is constructed solely by changing the value of $\mathcal{A}(X)$.

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 D the model M could have achieved without component A?

I. <u>Performance on subtask \mathcal{D} </u> 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. <u>Model \mathcal{M} could have achieved</u>: $\mathcal{M}^{\setminus \mathcal{A}}$ 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



◆□ > ◆□ > ◆臣 > ◆臣 > ○臣 - のへぐ

Maximilian Li and Lucas Janson

Example



◆□ > ◆□ > ◆臣 > ◆臣 > ○臣 - のへぐ

Maximilian Li and Lucas Janson

Example



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

Maximilian Li and Lucas Janson

イロン イロン イヨン イヨン

э

Example



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

Maximilian Li and Lucas Janson

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 D the model M could have achieved without component A?

IV. <u>"Best" performance</u>: we want to understand how much performance degrades *because* we had to ablate 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:

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

Proposition

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

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

Maximilian Li and Lucas Janson

イロト 不得 トイヨト イヨト

э

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.



Maximilian Li and Lucas Janson

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.

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

Table 1: Comparison of ablation loss gap Δ on IOI

▲□▶▲□▶▲≡▶▲≡▶ ≡ のQ@

Maximilian Li and Lucas Janson

Circuit discovery

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



Maximilian Li and Lucas Janson

Circuit discovery results

5 Optimal Mean 2 Resample CF Ablation loss gap **A** 1 0.5 0.2 0.1 0.05 ò 250 500 750 1000 1250 1500 1750 Circuit edge count $|\tilde{E}|$

IOI circuits, ablation comparison

Greater-Than circuits, ablation comparison



▲□▶▲御▶▲臣▶▲臣▶ 臣 のへで

Maximilian Li and Lucas Janson

Causal tracing





▲□▶ ▲□▶ ▲ 臣▶ ▲ 臣▶ 三臣 - のへで

Maximilian Li and Lucas Janson

Causal tracing results



・ロト・日本・ キョン・ ヨー うえつ

Maximilian Li and Lucas Janson

Latent prediction



▲□▶ ▲□▶ ▲目▶ ▲目▶ 三日 - つへで

Maximilian Li and Lucas Janson

Latent prediction



▲□▶ ▲□▶ ▲目▶ ▲目▶ 三目 - のへで

Maximilian Li and Lucas Janson

Latent prediction: tuned lens



▲□▶ ▲□▶ ▲目▶ ▲目▶ 三日 - 20ペ

Maximilian Li and Lucas Janson

Latent prediction: Optimal Constant Attention (OCA lens)



Maximilian Li and Lucas Janson

Optimal ablation for interpretability

▲日 ▶ ▲圖 ▶ ▲ 画 ▶ ▲ 画 ▶ ④ ④ ◎

Latent prediction results



Maximilian Li and Lucas Janson

イロト イボト イヨト イヨト

Latent prediction: causal faithfulness



Maximilian Li and Lucas Janson

< ロ > < 回 > < 回 > < 回 > < 回 >

Latent prediction: truthful elicitation



Elicitation accuracy on selected datasets with 10 demos, GPT-2-XL

Maximilian Li and Lucas Janson

・ロト・日本・日本・日本・日本・日本

Maximilian Li and Lucas Janson