i-Algebra: Towards Interactive Interpretability of Deep Neural Networks

Zhang et al. AAAI 2021.

WKD, Hubert Baniecki, 07.2021

Outline

1.

The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)

i-Algebra: Towards Interactive Interpretability of Deep Neural Networks

Xinyang Zhang,[1] Ren Pang,[1] Shouling Ji,[2] Fenglong Ma,[1] Ting Wang[1]
[1]Pennsylvania State University,
[2]Zhejiang University
{xqz5366, rbp5354, fenglong, ting}@psu.edu, sji@zju.edu.cn

2.
?

ICML 2021 Workshop on

Theoretic Foundation, Criticism, and Application Trend of Explainable AI

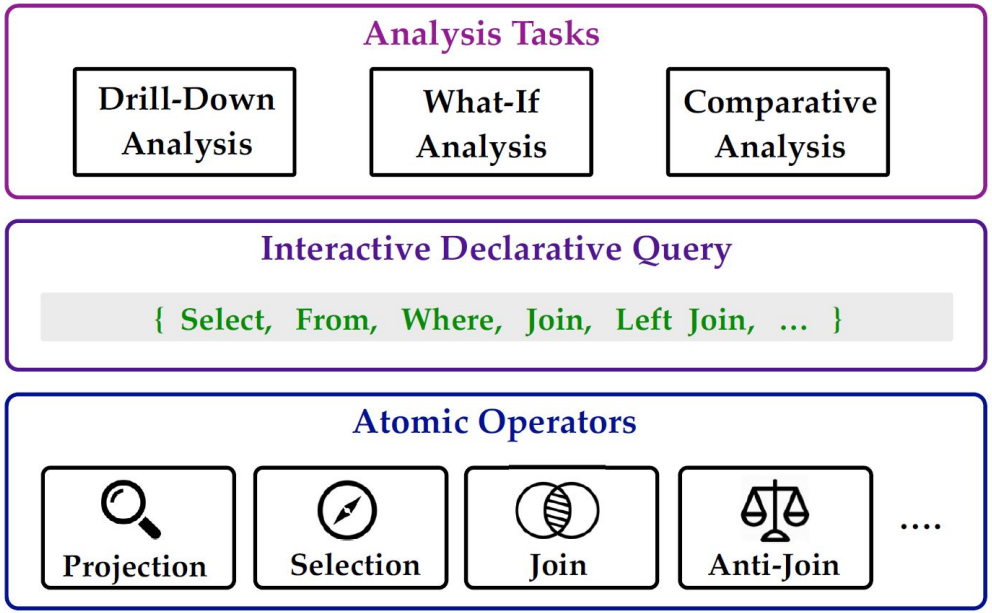
Interactive DNN Analysis

1. Single static explanations are impractical
2. “The answer to one question may trigger follow-up questions”
3. Interactive human-model language
4. Evaluate its usefulness on user-studies (solve tasks with interpretability)

- How does the feature importance change if some other features are present/absent?
- How does the feature importance evolve over different stages of the DNN model?
- What are the common features of two inputs that lead to their similar predictions?
- What are the discriminative features of two inputs that result in their different predictions?

Zhang et al. *i-Algebra: Towards Interactive Interpretability of Deep Neural Networks*. AAAI. 2021.

i-Algebra framework



select - [explanation]
l - DNN layer
f - DNN prediction
x - input image
x’ - 2nd input image (to compare)
join - intersection of explanations

```
select l from f(x) left join (select l from f(x'))
```

Zhang et al. *i-Algebra: Towards Interactive Interpretability of Deep Neural Networks*. AAAI. 2021.

SHAP → Shapley Values

Lundberg & Lee. *A Unified Approach to Interpreting Model Predictions*. NeurIPS. 2017.

Ancona et al. *Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Value Approximation*. ICML. 2019.

2^N → N^2

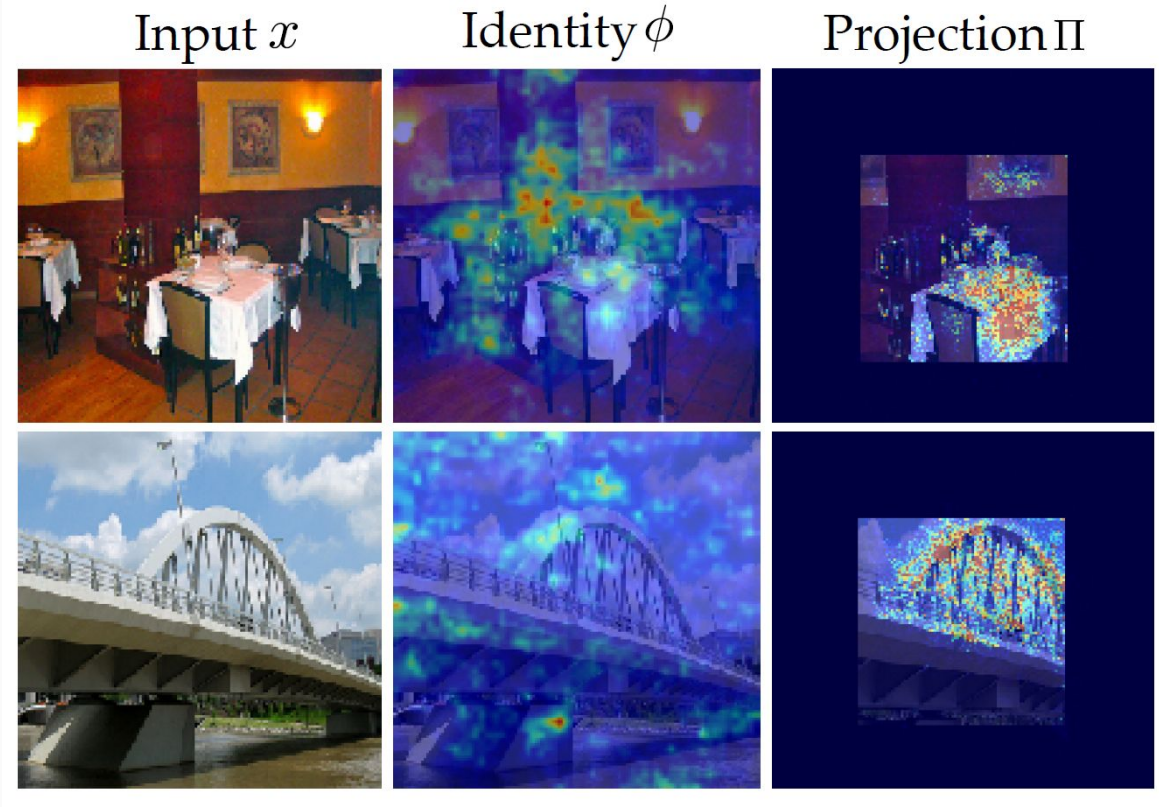
Algorithm 1 Deep Shapley algorithm (dense layers only)

Input: input $\mathbf{x}$, coalitions sizes $k_1, ..., k_K$, first layer weights $\mathbf{w}$, LPN without first linear layer $\hat{f}_c$
Initialize result vector $\mathbf{R}^c$ at zero
for $i = 1, ..., N$ **do**
  for $k = k_1, ..., k_K$ **do**
    $\bar{\mathbf{x}} = \mathbf{x}$
    $\bar{\mathbf{x}}[i] = 0$
    // Compute statistics of features excluding i
    $\boldsymbol{\mu} = \frac{1}{N-1}(\mathbf{W}\bar{\mathbf{x}})$
    $\boldsymbol{\sigma^2} = \frac{1}{N-1}(\mathbf{W}^2\bar{\mathbf{x}}^2) - \boldsymbol{\mu}^2$
    ...rrent coalition size
    ...oduced by i
    ...tions up to the output layer
    ...^{2})
    ...^{2})
    ... contribution of i to coalitions
    ...$\bar{\mu}^{(l)} - \mu^{(l)}$)
    ...apley values $\mathbf{R}^c$

In other previous works, the baseline value is sampled from the training set or a prior distribution [11, 2, 8]. Unfortunately, this approach is extremely slow. necessary samples (i.e.

Instead, most literature on attribution methods for DNNs suggests the use a fixed baseline value. In this case, zero is the canonical choice [12, 15, 9]. Notice that Gradient × Input and LRP can also be interpreted as using a zero baseline implicitly. One possible justification relies on the observation that in network that implements a chain of operations of the form $x_j^{(1)} = \sigma(\sum_i (w_{ij}x_i) + b_j)$, the all-zero input is somehow neutral to the output (ie. $\forall c \in C : R_c(\mathbf{0}) \approx 0$). In fact, if all additive biases b_j in the network are zero and we only allow nonlinearities that cross the origin (e.g. ReLU or Tanh), the output for a zero input is exactly zero for all classes. Empirically, the output is often near zero even when biases have different values, which makes the choice of zero for the baseline reasonable, although arbitrary.

Identity & Projection



$$[\phi(x)]_i = \frac{1}{d}\sum_{k=0}^{d-1}\mathbb{E}_{I_k}[f(x_{I_k\cup\{i\}}) - f(x_{I_k})]$$

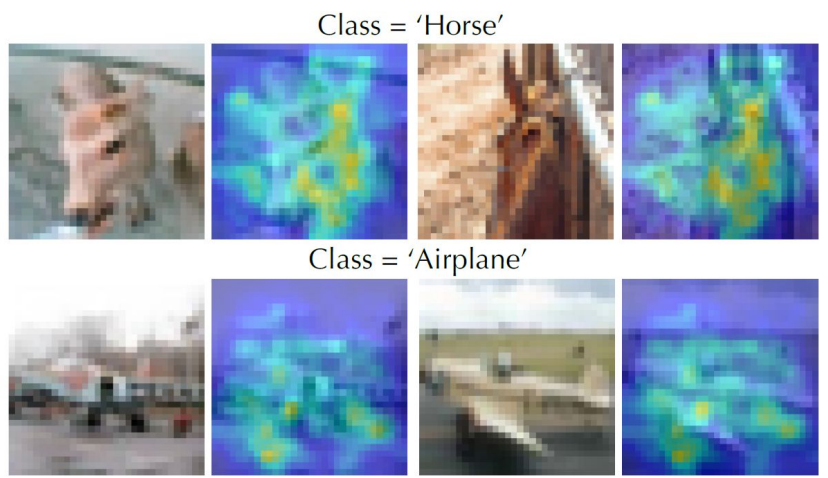
$$[\Pi_w(x)]_i = \begin{cases} \frac{1}{|w|}\sum_{k=0}^{|w|-1}\mathbb{E}_{I_k}[f(x_{I_k\cup\{i\}}) - f(x_{I_k})] & i \in w \\ 0 & i \notin w \end{cases}$$

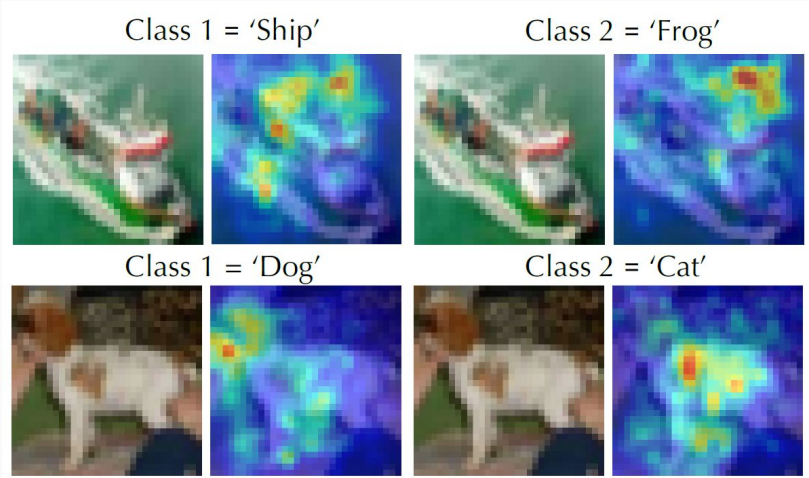
ImageNet and ResNet50

Zhang et al. *i-Algebra: Towards Interactive Interpretability of Deep Neural Networks*. AAAI. 2021.

Join & Anti-Join

adversarial input (changed)





$$[x \bowtie x']_i = \epsilon \cdot [\phi(x; \bar{x}, f)]_i + (1 - \epsilon) \cdot [\phi(x'; \bar{x}, f)]_i$$

$$[x \diamond x']_i = ([\phi(x; x', f)]_i, [\phi(x'; x, f)]_i)$$

CIFAR10 and VGG19

Note that the anti-join operator is extensible to the case of the same input x but different models f and f'. Specifically, to compute x's features that discriminate $f(x)$ from $f'(x)$,

Zhang et al. *i-Algebra: Towards Interactive Interpretability of Deep Neural Networks*. AAAI. 2021.

SQL !

```
select * from f(x)
```
– the identity operator $\phi(x)$.

```
select * from f(x) where w
```
– the projection operator $\Pi_w(x)$.

```
select l from f(x)
```
– the selection operator $\sigma_l(x)$.

```
select * from f(x) join (select * from f(x'))
```
– the join operator $x \bowtie x'$.

```
select * from f(x) left join (select * from f(x'))
```
– the anti-join operator $x \diamond x'$.

```
select l from f(x) where w
```
– the composition of selection and projection $\Pi_w \sigma_l(x)$.

```
select l from f(x) join (select l from f(x'))
```
– the composition of join and selection $\sigma_l(x) \bowtie \sigma_l(x')$.

Zhang et al. *i-Algebra: Towards Interactive Interpretability of Deep Neural Networks*. AAAI. 2021.

Interactive Analysis

Drill-Down Analysis – Here the user applies a sequence of projection and/or selection to investigate how the DNN model f classifies a given input x at different granularities of x and at different stages of f. This analysis helps answer important questions such as: (i) how does the importance of x's features evolve through different stages of f? (ii) which parts of x are likely to be the cause of its misclassification? (iii) which stages of f do not function as expected?

What-If Analysis

Comparative Analysis – In a comparative analysis, the user applies a combination of join and/or anti-join operators on the target input x and a set of reference inputs $\mathcal{X}$ to compare how the DNN f processes x and $x' \in \mathcal{X}$. This analysis helps answer important questions, including: (i) from f's view, why are x and $x' \in \mathcal{X}$ similar or different? (ii) does f indeed find the discriminative features of x and $x' \in \mathcal{X}$? (iii) if x is misclassified into the class of x', which parts of x are likely to be the cause?

Zhang et al. *i-Algebra: Towards Interactive Interpretability of Deep Neural Networks*. AAAI. 2021.

Empirical Evaluation

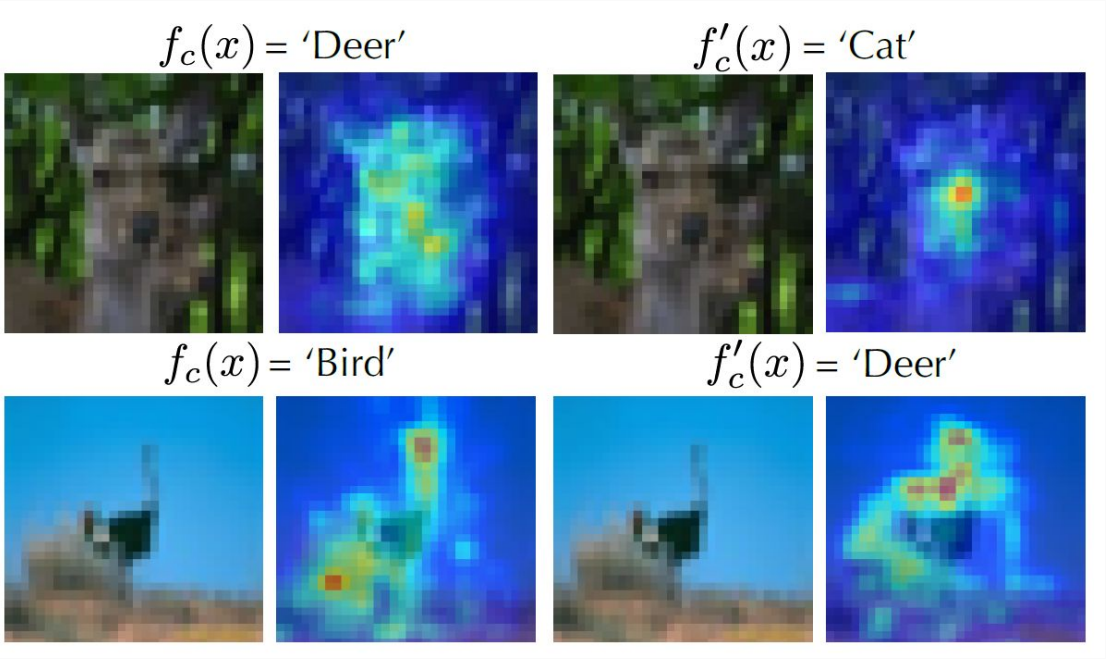
- RQ1: Versatility – Does i-Algebra effectively support a range of analysis tasks?
- RQ2: Effectiveness – Does it significantly improve the analysis efficacy in such tasks?
- RQ3: Usability – Does it provide intuitive, user-friendly interfaces for analysts?

We conduct user studies on the Amazon MTurk platform, in which each task involves 1,250 assignments conducted by 50 qualified workers. We apply the following quality control: (i) the users are instructed about the task goals and declarative queries, and (ii) the tasks are set as small batches to reduce bias and exhaustion.

Zhang et al. *i-Algebra: Towards Interactive Interpretability of Deep Neural Networks*. AAAI. 2021.

Case A: Resolving Model Inconsistency

Two models with divergent predictions



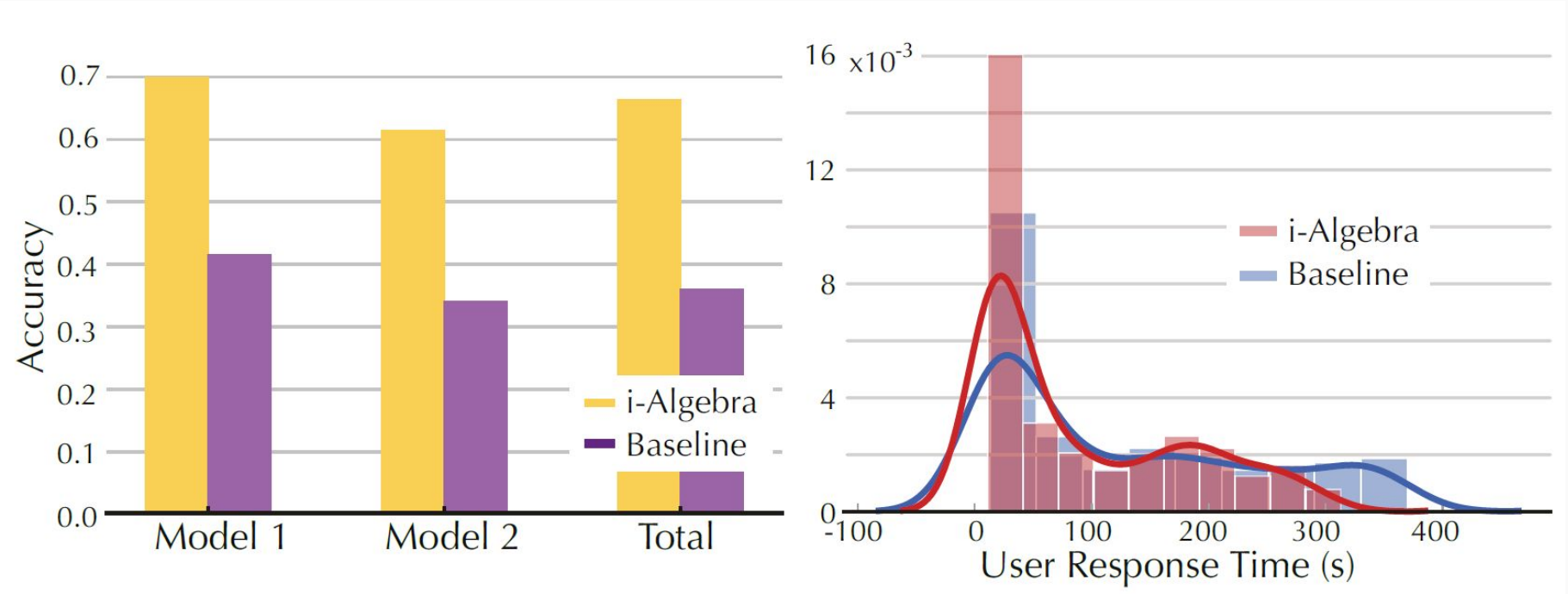

```
select * from f(x) left join (select * from f'(x))
```

Zhang et al. *i-Algebra: Towards Interactive Interpretability of Deep Neural Networks*. AAAI. 2021.

Case A: Resolving Model Inconsistency

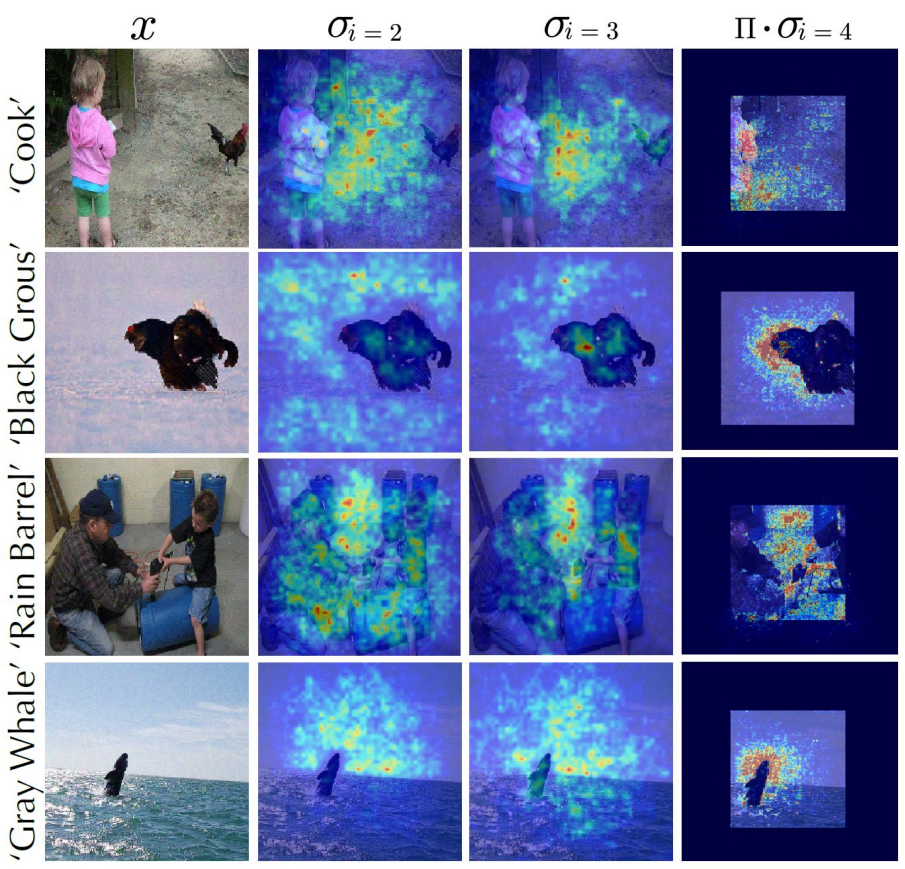
SOTA in DL -IF- ACC high -AND- TIME low

baseline - explanation
i-Algebra - Anti-Join



Zhang et al. *i-Algebra: Towards Interactive Interpretability of Deep Neural Networks*. AAAI. 2021.

Case B: Detecting Adversarial Inputs

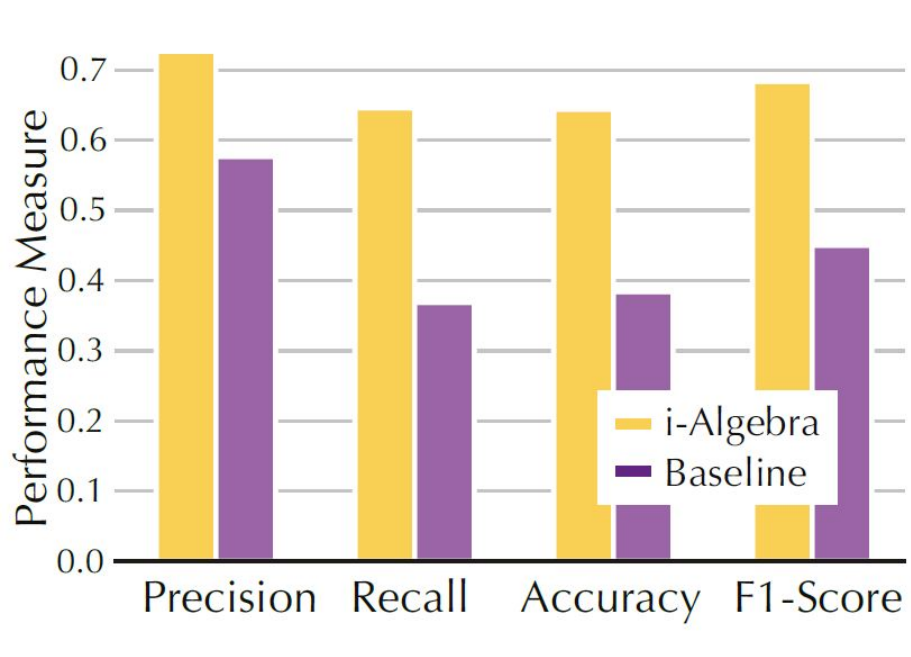


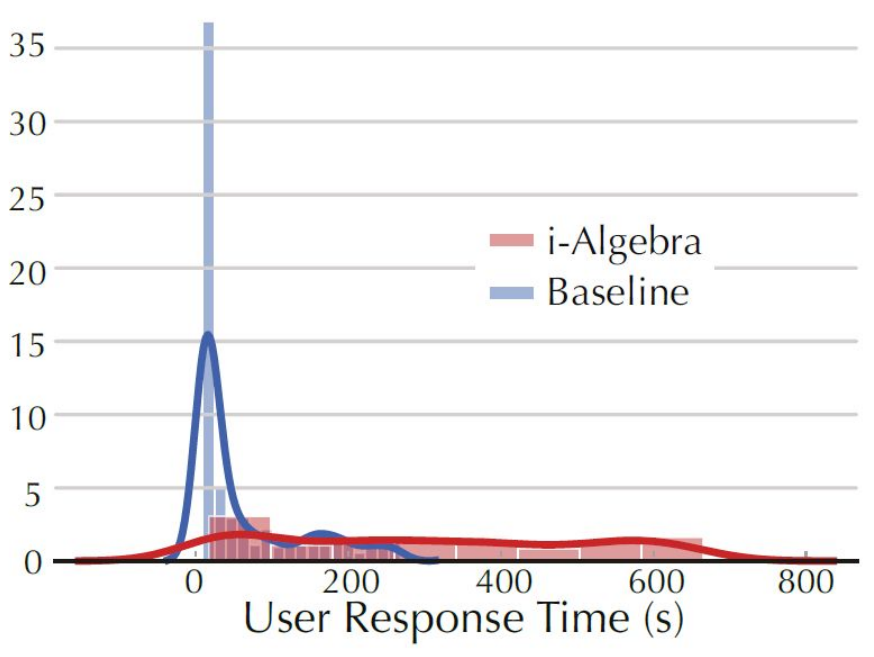

```
select l from f(x) where w
```

Zhang et al. *i-Algebra: Towards Interactive Interpretability of Deep Neural Networks*. AAAI. 2021.

Case B: Detecting Adversarial Inputs

baseline - explanation
i-Algebra - Selection + Partition





Zhang et al. *i-Algebra: Towards Interactive Interpretability of Deep Neural Networks*. AAAI. 2021.

Case C: Cleansing Poisoning Data

Figure 11: Sample trigger inputs misclassified as “deer”.

Prediction	Ground-truth	
	+	-
+	48	654
-	399	3574

Table 1: Statistics of samples in Case C.

```
select l from f(x)
```

Zhang et al. *i-Algebra: Towards Interactive Interpretability of Deep Neural Networks*. AAAI. 2021.

Case C: Cleansing Poisoning Data

Precision	Recall	Accuracy	F1-Score
0.609	0.6343	0.586	0.622

Table 2: Users' performance under i-Algebra in Case C.

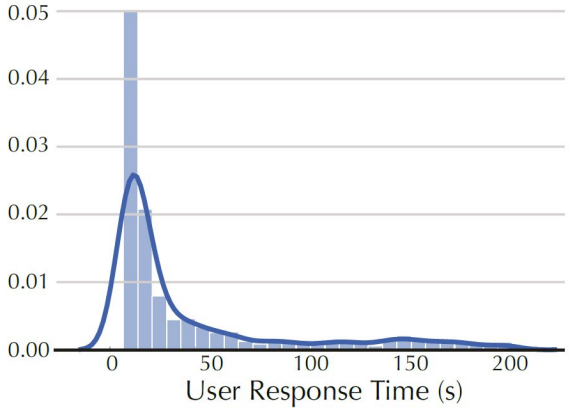


Figure 12: URT distribution in Case C.

baseline - ??????
i-Algebra - Selection

Conclusion

This work promotes a paradigm shift from *static* interpretation to *interactive* interpretation of DNNs, which significantly improves the usability of existing interpretation models in practice. We design and prototype i-Algebra, a first-of-its-kind interactive framework for DNN interpretation. The extensive studies in three representative analysis tasks all demonstrate the promising usability of i-Algebra.

Zhang et al. *i-Algebra: Towards Interactive Interpretability of Deep Neural Networks*. AAAI. 2021.