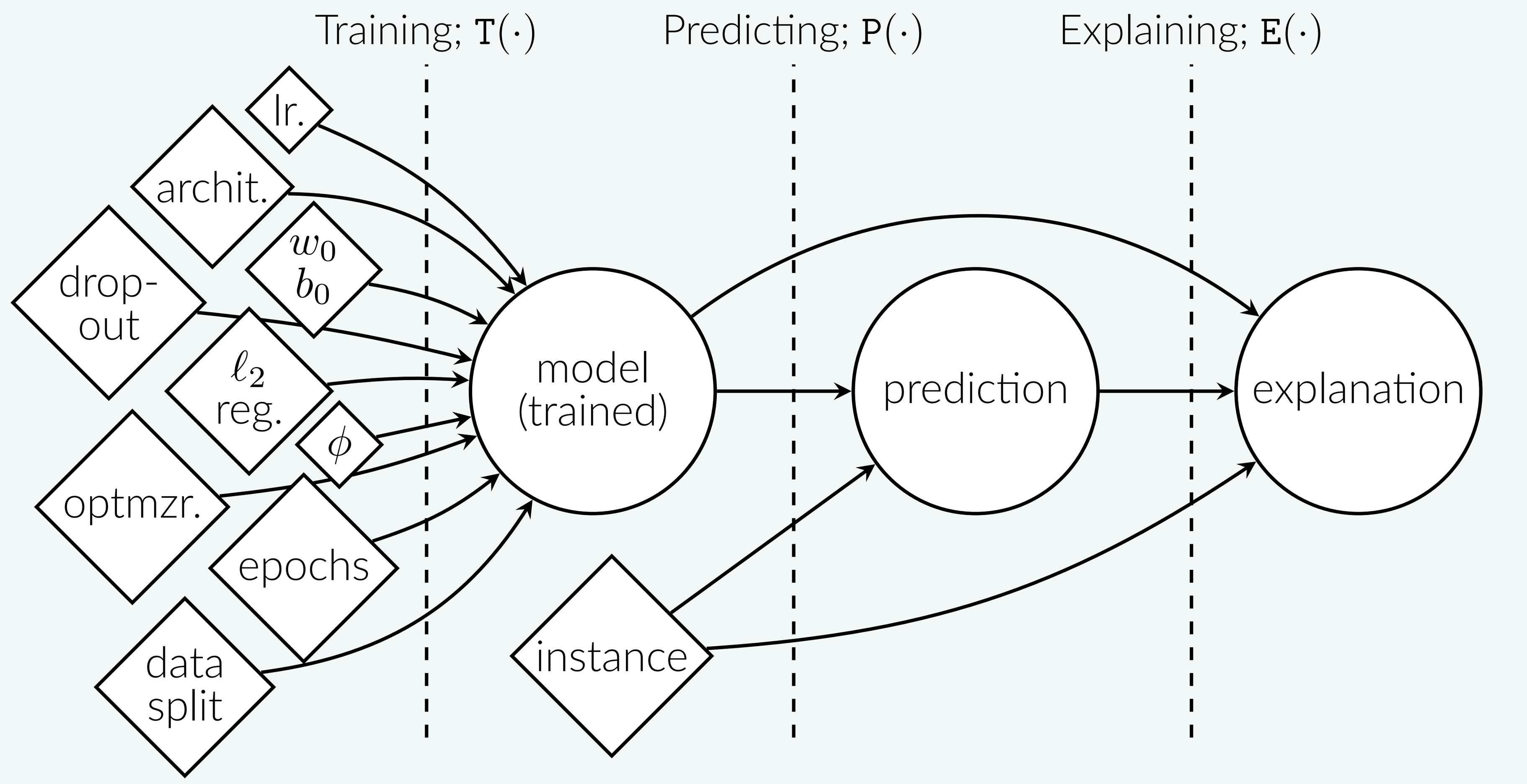
On the Relationship Between Explanation & Prediction: A Causal View

Amir-Hossein Karimi, Krikamol Muandet, Simon Kornblith, Bernhard Schölkopf, Been Kim

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS · ETH zürich · Google AI · ICML International Conference On Machine Learning

Summary



Explainability is a critical aspect in the lifecycle of machine learning (ML) models.

We lack understanding as to **what ML explanation methods can and cannot do**.

Various factors including data, random initialization, **model predictions** (Y), and **training hyperparameters** (H) have significant effects on **explanations** (E).

Previous research [Adebayo et al., 2018] suggests a **weak correlation between** E **and** Y, calling for a definitive study to quantify this relationship.

Using the **Potential Outcomes framework** [Rubin, 2005], we systematically examine the relationship between Y and E.

By measuring the treatment effect when intervening on their causal predecessors (H), we introduce a causally-based quantitative metric for investigating the relationship between Y and E.

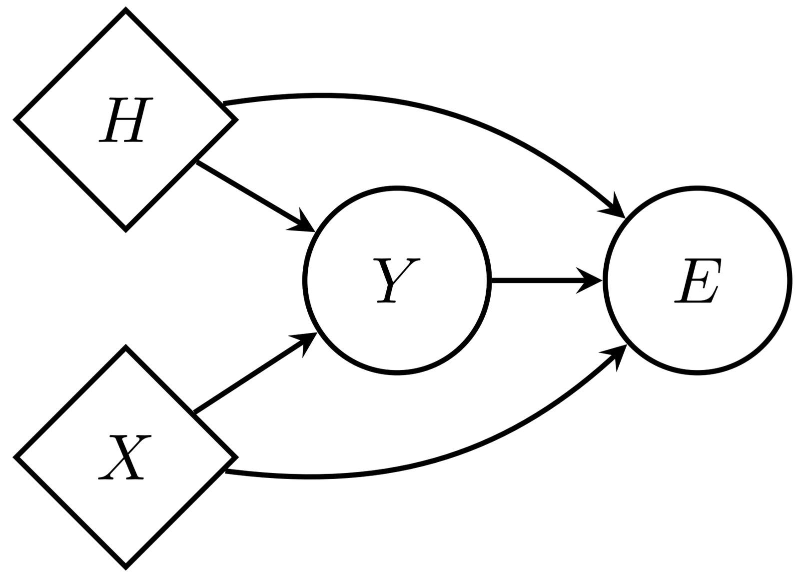
Conclusion: **explanations might be providing insights beyond just the model prediction.**

References

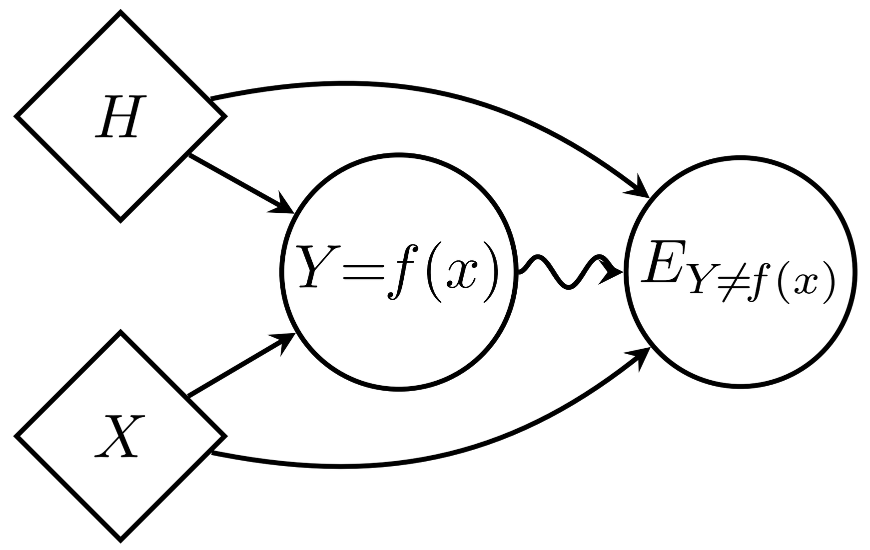
Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Methodology



Intervening on factors (H, X) allow for studying their treatment effect (i.e., causal influence) on down-stream targets (i.e., Y, E)



ITE_E measures the total effect (direct & indirect effect). How to tease them apart? We can sever the flow of dependence from H to E by randomizing Y.

Single binary treatment effect — $Y^*_{h=1}(x) - Y^*_{h=0}(x)$
effect of $h = 1$ w.r.t $h = 0$ on $x \in X$

Single non-binary treatment effect — $\mathbb{E}_{m\neq n}\left[Y^*_{h=n}(x) - Y^*_{h=m}(x)\right]$
effect of $h = n$ w.r.t $h \neq n$ on $x \in X$

Multiple non-binary treatment effect — $\mathbb{E}_{h_{\setminus i}}\left[\mathbb{E}_{m\neq n}\left[Y^*_{[h_i=n,h_{\setminus i}]}(x) - Y^*_{[h_i=m,h_{\setminus i}]}(x)\right]\right]$
effect of $h_i = n$ w.r.t $h_i \neq n$ on $x \in X$

Kernelized treatment effect —
$$\|\phi(Y^*_h(x)) - \phi(Y^*_{h'}(x))\|^2_{\mathcal{G}} = k(Y^*_h(x), Y^*_h(x)) - 2k(Y^*_h(x), Y^*_{h'}(x)) + k(Y^*_{h'}(x), Y^*_{h'}(x))$$

(In)Direct treatment effect of H on E —
total effect: $ITE_E, y = f(x)$
direct effect: $ITE_E, y \neq f(x)$
indirect effect: Δ above

Observational Study

4 datasets	MNIST, FASHION, SVHN, CIFAR10
8 hyparparameters	drawn “indep. at random” from pre-specified ranges Fixed architecture. Fixed random seed.
$30{,}000$ pre-trained models	3-layer CNNs (4,970 parameters) trained to convergence (max 86 epochs)
$4+1$ saliency-based explanations	Gradient, SmoothGrad, Integrated Gradients, Grad-CAM Reference E: “identity”, i.e., $E = Y \implies \text{ITE}_E = \text{ITE}_Y$

Experiments

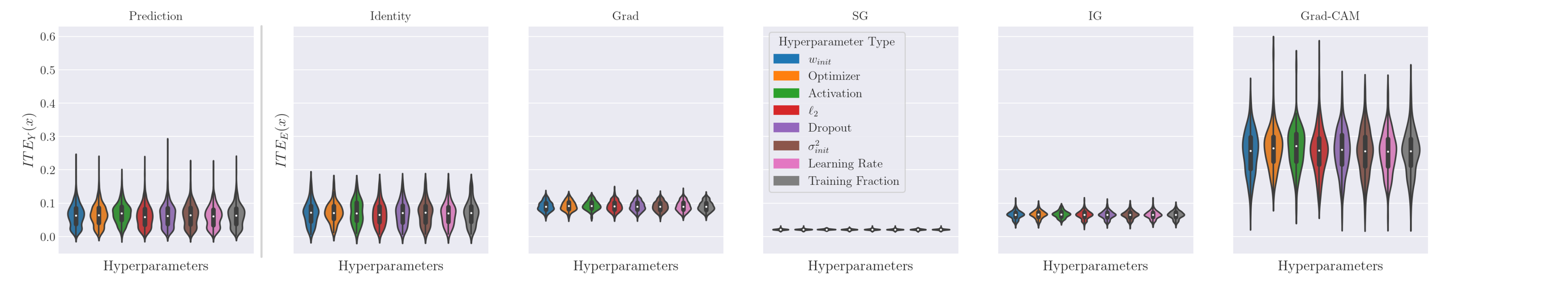


Figure 3. Comparison of ITE_Y and ITE_E for CIFAR10 shows that different types of H influence E and Y in a similar way.

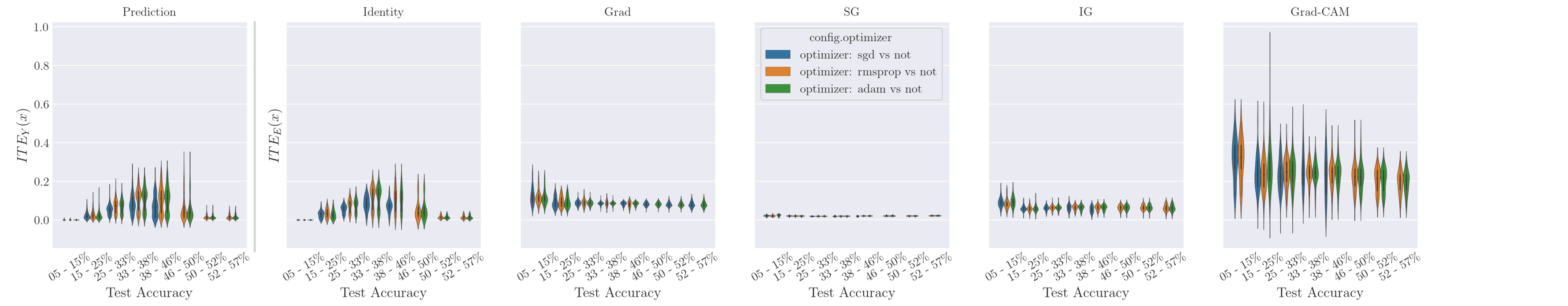


Figure 4. Comparison of ITE values of $h_{\text{optimizer}}$ on Y (left) and E (right) for models across different performance buckets. Interestingly, ITE_E differs across accuracy buckets. More importantly, none of the explainability methods resemble ITE_Y.

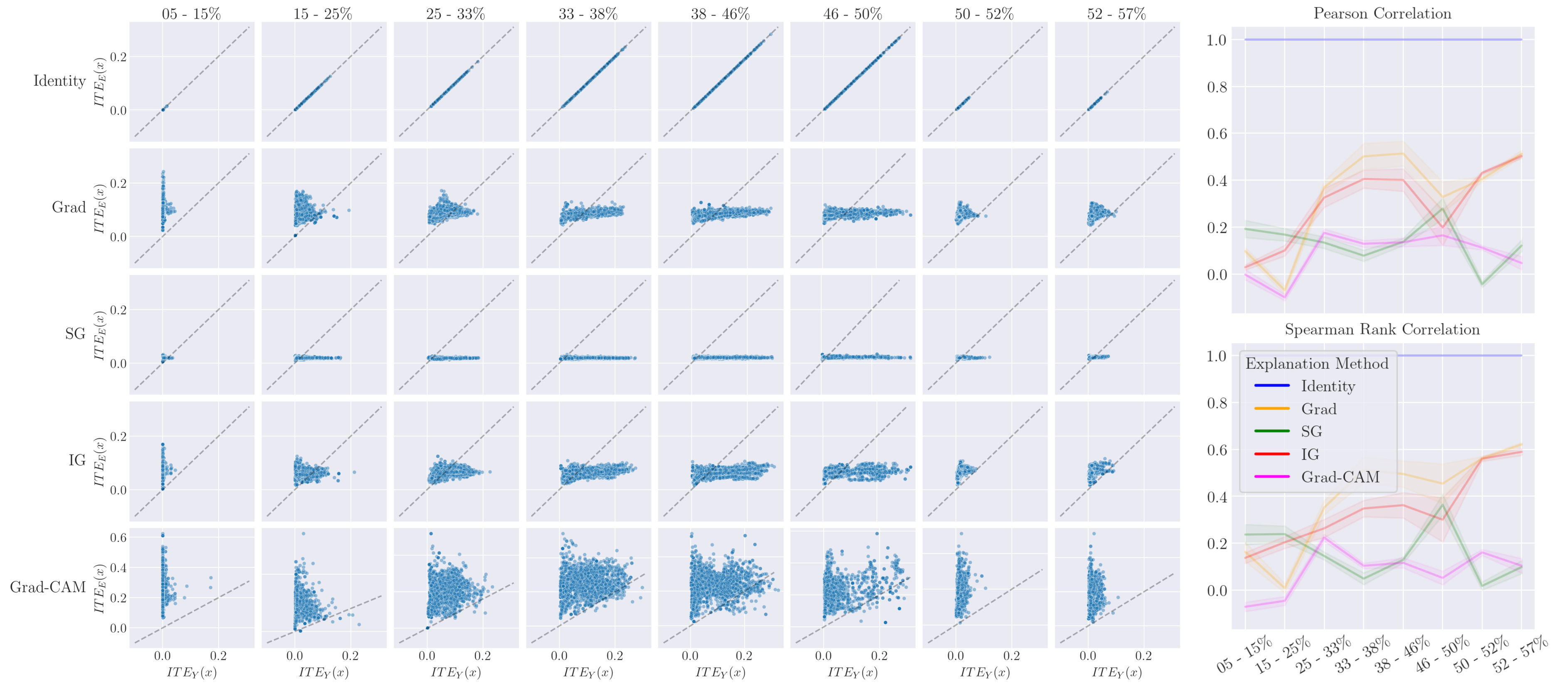


Figure 5. (left) Each column is a subset of models at each accuracy bucket, each row is a different explanation method. Whereas low-performing CIFAR10 models (first column) show little change in predictions as their explanations differ, top-performing models show the reverse of this trend. (right) Correlation measures of the scatter plots on the left.

Figure 6. Pearson correlation between ITE_Y and ITE_E in total and direct effect (first column). The second column is the difference between total and direct effect, where higher values mean that the influence of H on E flows more through Y (ideal). The third column plots the difference in delta correlations between the ideal case (Identity) and each method. In other words, it indicates how far each method moves away from the ideal case, as a model performs better.