Data-centric Prediction Explanation via Kernelized Stein Discrepancy

ICLR 2025

Mahtab Sarvmaili, Hassan Sajjad, Ga Wu

Faculty of Computer Science, Dalhousie University

April, 2025

Data-centric, Post hoc prediction explanation

- Explaining models prediction
  - focusing on **data quality, patterns, and examples** rather than internal model mechanics
  - without accessing the model training dynamic
- Shifts the focus from
  - From ”what are the internal processing of the model?"
  - To ”How the model learn the data and what is the correlation between samples?”
- Emphasizes **training data** and its impact on predictions.

Data-centric, Post hoc prediction explanation pt.2

- Generally, is done by constructing the influence chain between training and test data

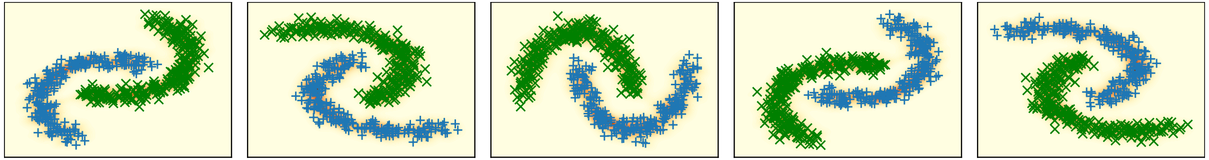
Method	Explanation of	Need optimization as sub-routine	Whole model explanation: Theoretical	Whole model explanation: Practical	Inference computation complexity bounded by	Memory/cache (**of each training sample**) bounded by
Influence Function	Original Model	Yes (Iterative HVP approximation)	Yes	No	1.$H_\theta^{-1}\nabla_\theta L(\mathbf{x}_t, \theta)$ approximation 2. $<\nabla_\theta L(\mathbf{x}, \theta), H_\theta^{-1}\nabla_\theta L(\mathbf{x}_t, \theta)>$	Size of model parameters
RPS	Fine-tuned Model	Yes (L2 regularized last layer retrain)	No	No	1.last layer representation $\mathbf{f}_t$ 2.$<\alpha_i \mathbf{f}_i, \mathbf{f}_t>$	Size of model parameters of the last layer
TracIn*	Original Model	No	Yes	No	1.$\nabla_\theta L(\mathbf{x}_t, \theta)$ approximation 2. $<\nabla_\theta L(\mathbf{x}, \theta), \nabla_\theta L(\mathbf{x}_t, \theta)>$	Size of model parameters
HD-Explain	Original Model	No	Yes	Yes	1.$\nabla_{\mathbf{x}_t} f(\mathbf{x}_t, \theta)_{y_t}$ 2. Closed-form $k_\theta(\mathbf{x}, \mathbf{x}_t)$ defined by KSD	Size of data dimension

Highly Precise and Data-Centric Explanation (HD-Explain)

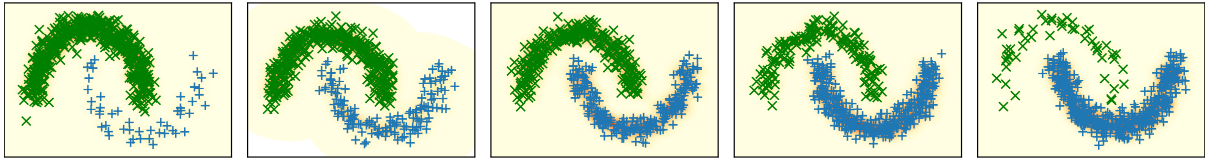
- Treating the model as the estimator of data distribution
- Understanding the correlation between samples conditioned on the model
- Retaining the influence chain using Kernelized Stein Discrepancy (KSD)
- Exploiting this probability distance measure to obtain sample correlation

KSD between model and data

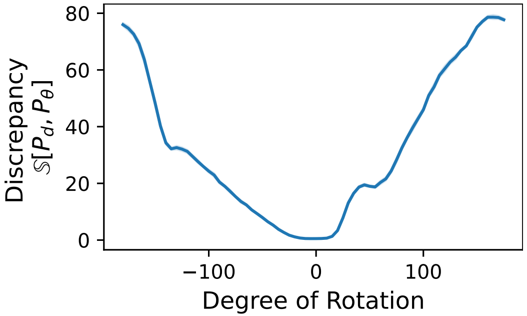
- KSD between training data and augmented data
  - Rotation
  - Density change

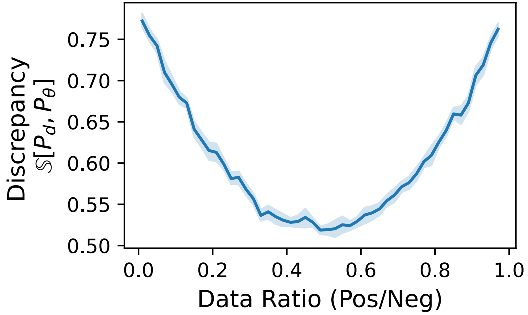


Rotated: -120° | Rotated: -60° | Training Data | Rotated: 60° | Rotated: 120°



Class Ratio: 1:9 | Class Ratio: 3:7 | Training Data | Class Ratio: 7:3 | Class Ratio: 9:1





Kernelized Stein Discrepancy (KSD)

- Stein Identity for a smooth distribution $p(x)$ and function $\phi(x)$

$$\mathbb{E}_{x\sim p}[\phi(x)\nabla_x \log p(x) \ + \nabla_x \phi(x)] = 0$$

- Stein Identity characterizes the $p(x)$ by stein operator A_p

$$A_p \phi(x) = \phi(x)\nabla_x \log p(x) \ + \nabla_x \phi(x)$$

- Stein Discrepancy to measure the distance between two distributions

$$\sqrt{\mathbb{S}(p,q)} = \max_{\phi \in F} \mathbb{E}_{x\sim q}\left[A_p \phi(x)\right]$$

- Limiting the F to be the unit ball of Reproducing Hilbert Kernel Space

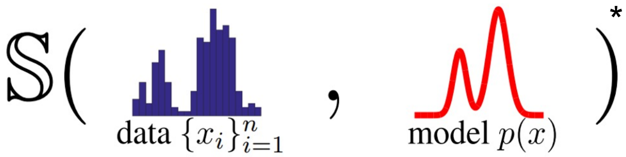
$$\text{KSD* } = \mathbb{S}(p,q) = \mathbb{E}_{x,x'\sim q}\left[\kappa_p \phi(x)\right] = [A_p^x A_p^{x'} k(x,x')]$$

*(Liu, Q., et al (2016). A kernelized Stein discrepancy for goodness-of-fit tests and model evaluation. *NeurIPS 2016*.)

DALHOUSIE UNIVERSITY

KSD between model and data

- A trained classifier to maximize the log likelihood of data
  - minimizes its KL-distance to the data distribution
  - Learns the data distribution



- KSD between model and training data

$$\mathbb{S}(p,q) = \mathbb{E}_{x,x'\sim q}\left[\kappa_p \phi(x)\right] = \left[A_p^x A_p^{x'} k(x,x')\right]$$

- Considering the uniform data distribution and relaxing the score function calculation

$$\nabla_{x,y} \log P_\theta(x,y) = [\nabla_x \log f_\theta(x)_y \,||\log f_\theta(x)]$$

* Liu, Q. (n.d.). (A short introduction to kernelized Stein discrepancy).

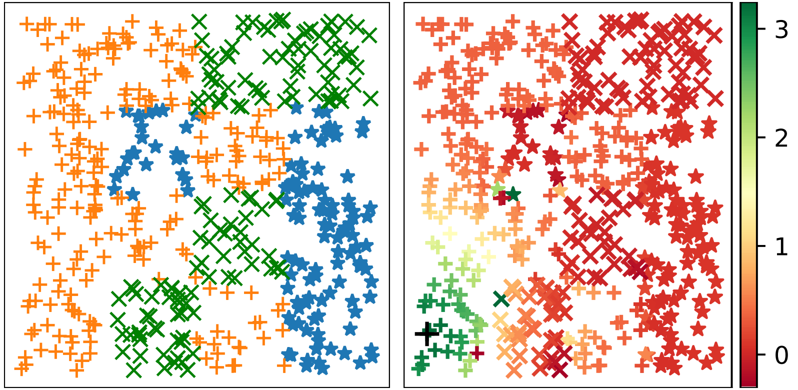
KSD between model and data

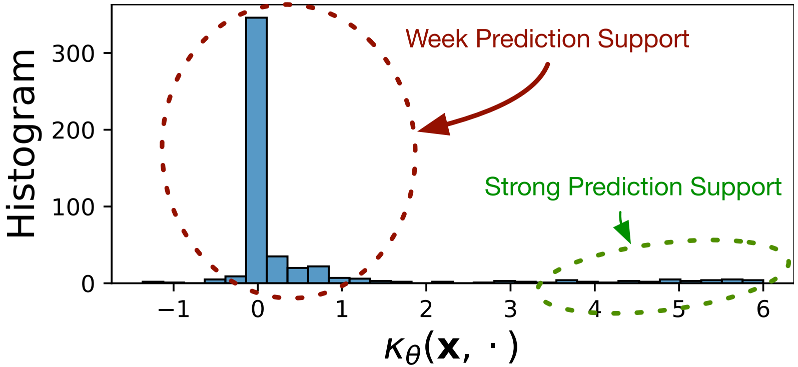
- A trained classifier $f_\theta \sim \operatorname{argmax} \mathbb{E}_{(x,y)\sim P_D} [\log P_\theta (y|x)]$
- KSD only models the discrepancy between joint distributions rather than conditional distributions
- Relaxing the discrepancy measure by considering the uniform data distribution

- Correlation of any pairs of test data with training data
  - Predicting the label
  - Constructing the data point $(x_t, \widehat{y_t})$
  - Apply KSD function
  - Select the top k-th samples

HD-Explain pt.2

- Correlation of any pairs of training data condition on data
  - Calculating the Stein Kernels for every pair
  - Sorting the values from the highest to lowest
  - Selecting the top samples



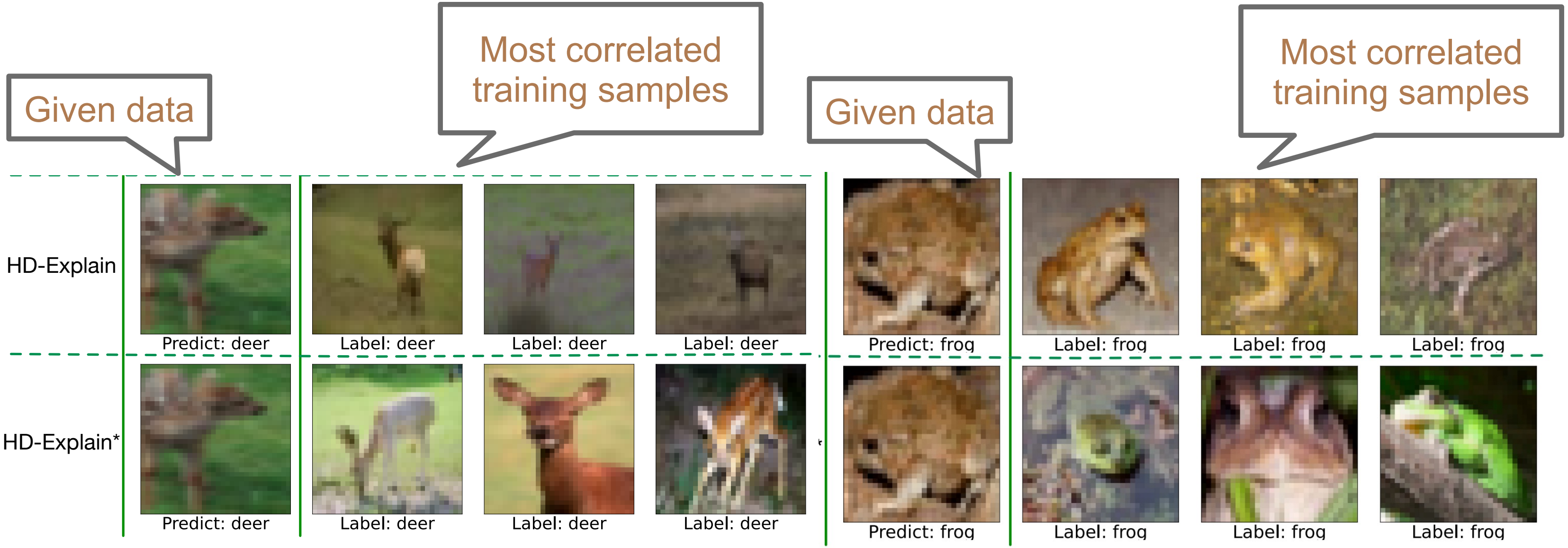


Experimental Results

Evaluation and analysis

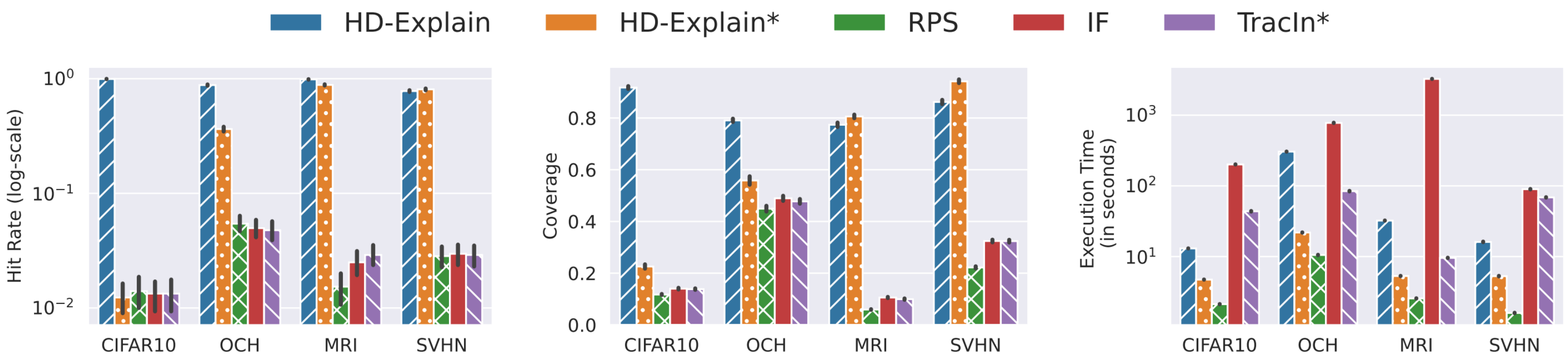
- Metrics
  - Hit Rate
  - Coverage
  - Run time
- Quantitative Evaluation
  - Noise Injection $x_t = x_i + \varepsilon$ s.t. $\varepsilon \sim \mathcal{N}(0,\ 0.01\sigma)$
  - Horizontal Flip $x_t = \text{flip}(x_i)$
- HD-Explain* is variant of HD-Explain that reduces the computation

Qualitative analysis



Quantitative analysis

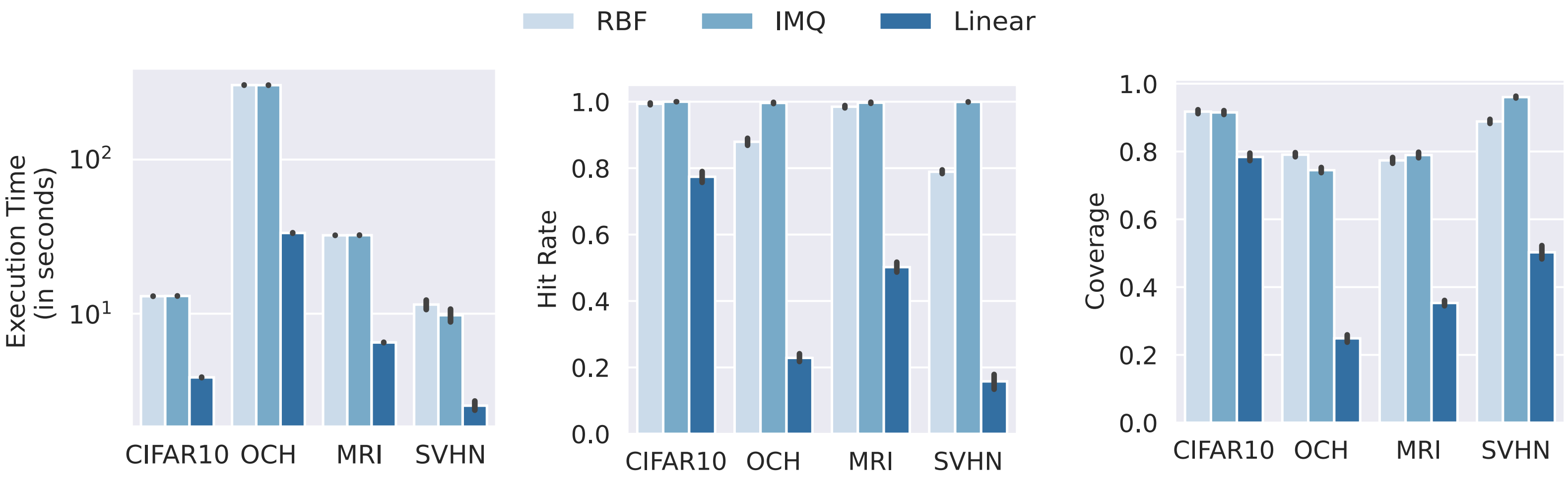
- Noise injection



(a) Hit Rate (b) Coverage (c) Execution Time

Quantitative analysis pt.2

- Choice of Kernel



Conclusion

- A novel **Kernel Stein Discrepancy-driven** example-based prediction explanation method.
- Outperforms three baseline methods across **three datasets** in both qualitative and quantitative assessments.
- Provides **accurate and effective explanations** at granular levels.
    1. **Flexibility**: Applicable to any layer of interest in a model.
    2. **Analysis Across Layers**: Enables tracking the evolution of predictions across layers.
- Retrieve the diverse and high coverage explanations for test data

Thanks for your attention!