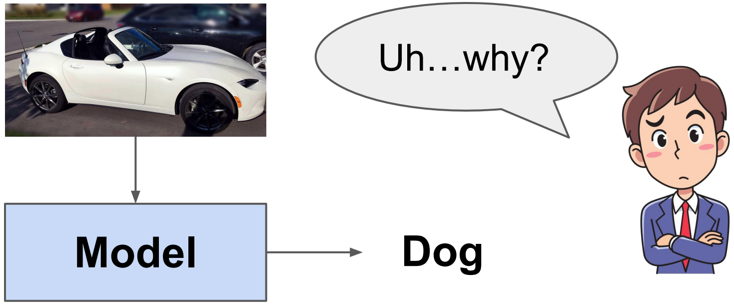
Representer Point Selection via Local Jacobian Expansion for Classifier Explanation of Deep Neural Networks and Ensemble Models

Yi Sui, Ga Wu, Scott Sanner
Dept. of Mechanical and Industrial Engineering, University of Toronto
NeurIPS 2021

UNIVERSITY OF TORONTO

Model Explanation with Training Samples

Model explanation:
Why does a model make a certain prediction?

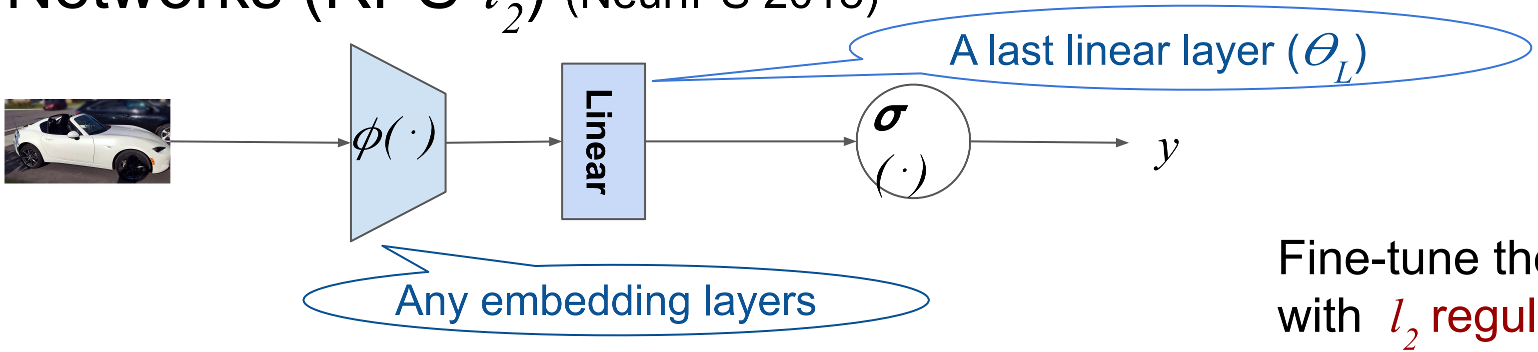


Explain with training data:
Identify the most influential training data samples on the prediction.

Dog

Dog

Representer Point Selection for Explaining Deep Neural Networks (RPS-l_2) (NeurIPS 2018)



Fine-tune the last layer with l_2 regularization

Train sample

Test sample

Loss

Last layer parameter

$$\hat{\mathbf{y}}_t = \sum_i^n \boldsymbol{\alpha}_i \mathcal{K}(\mathbf{x}_i, \mathbf{x}_t) = \sum_i^n \boxed{-\frac{1}{2\lambda n} \frac{\partial \mathcal{L}(\mathbf{x}_i, y_i, \boldsymbol{\Theta}^*)}{\partial \Theta_L^* \phi(\mathbf{x}_i)} \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_t)}$$

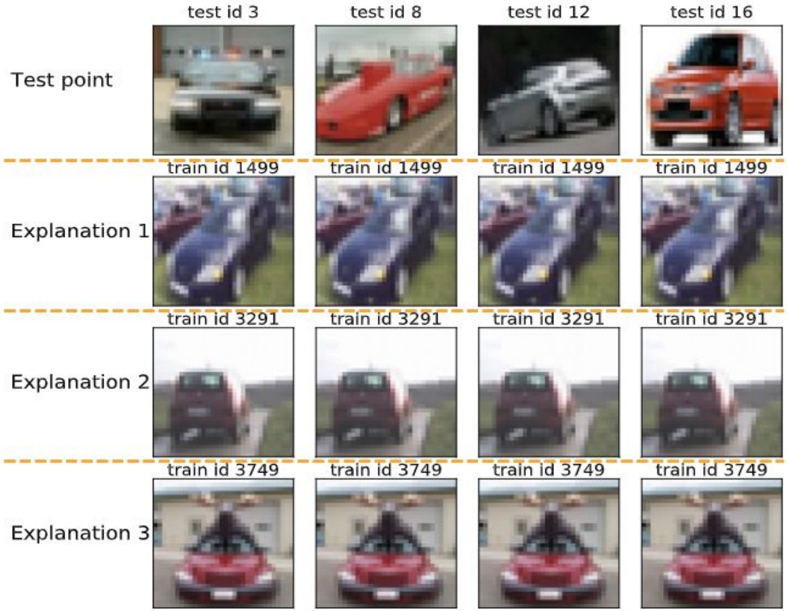
Training data influence

Linear combination

l_2 regularization weight

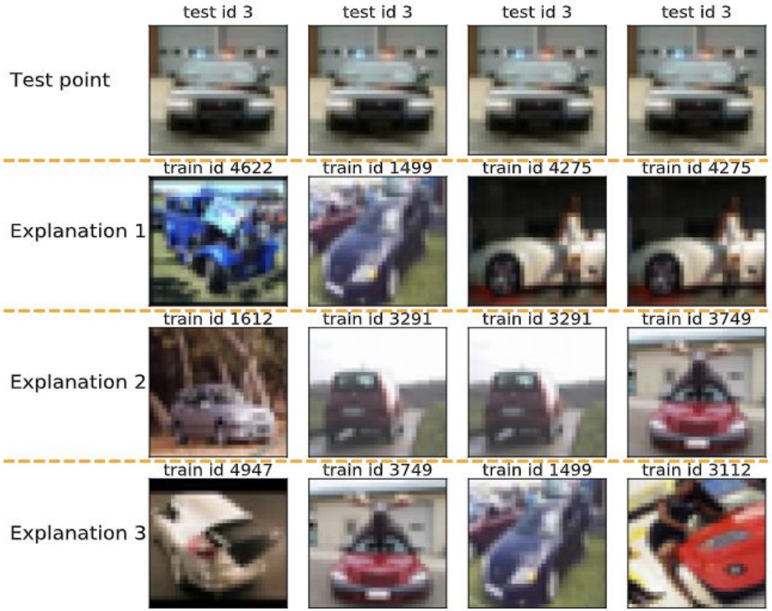
RPS-l_2 Caveats

1. Class-level explanation



Different images in the same class

2. Unfaithful explanation



Different l_2 coefficients

Our Method: Representer Point Selection via Local Jacobian Expansion (RPS-LJE)

- Motivation: avoid altering the model by imposing the l_2 regularization
- Solution: an alternative derivation with *Taylor expansion* on the first order gradient (Jacobian)

⇒ derives RPS-like result without the l_2 regularizer

RPS-LJE Data Influence Estimation

$$\underbrace{\left[\frac{1}{\phi(\mathbf{x}_i)n}\Theta_L^* - \frac{1}{n}\mathcal{H}_{\Theta_L^*}^{-1}\frac{\partial\mathcal{L}(\mathbf{x}_i, y_i, \mathbf{\Theta}^*)}{\partial\Theta_L^*\phi(\mathbf{x}_i)}\right]}_{\boldsymbol{\alpha}_i}\underbrace{\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_t)}_{\mathcal{K}(\mathbf{x}_i,\mathbf{x}_t)}$$

$\Theta_L^\dagger$: last layer of the given model

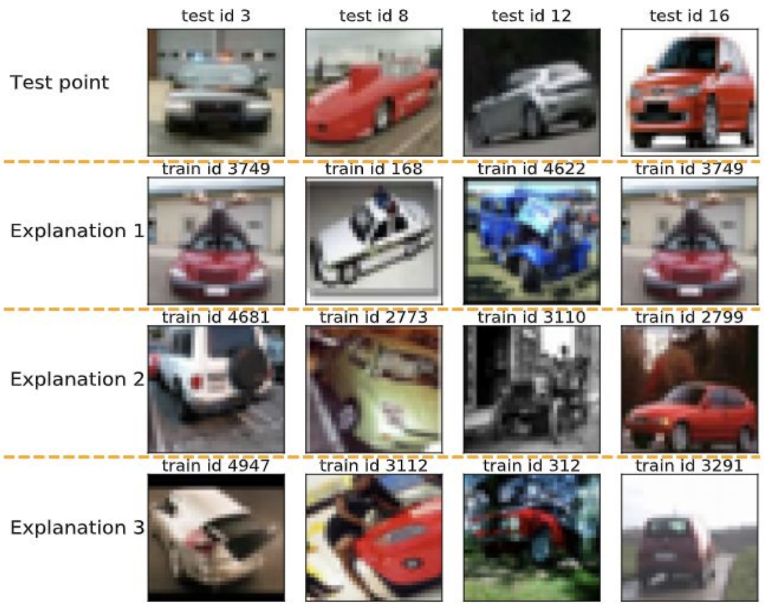
Θ_L^* : a *nearby* anchor point for Taylor expansion

Our result

- has a similar form with RPS-l_2
- does not contain dominant term (Hessian matrix w.r.t *all* data points)
- does not fine-tune the original model (one-step gradient ascent Θ_L^*)

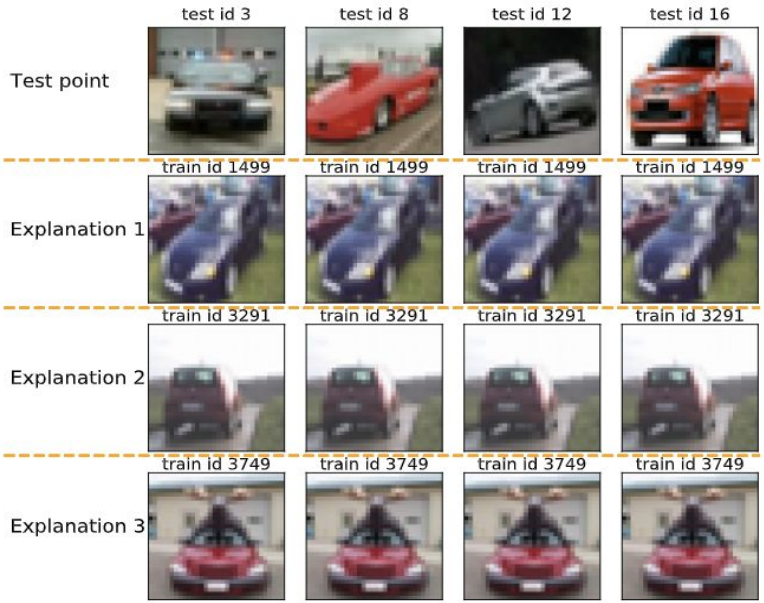
RPS-LJE with Instance-level Explanations

RPS-LJE



Instance-level explanations

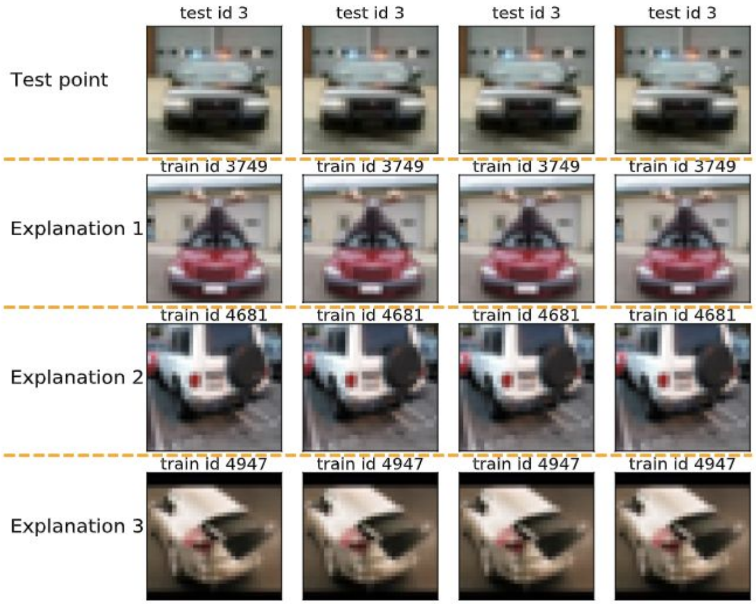
RPS-l_2



Class-level explanations

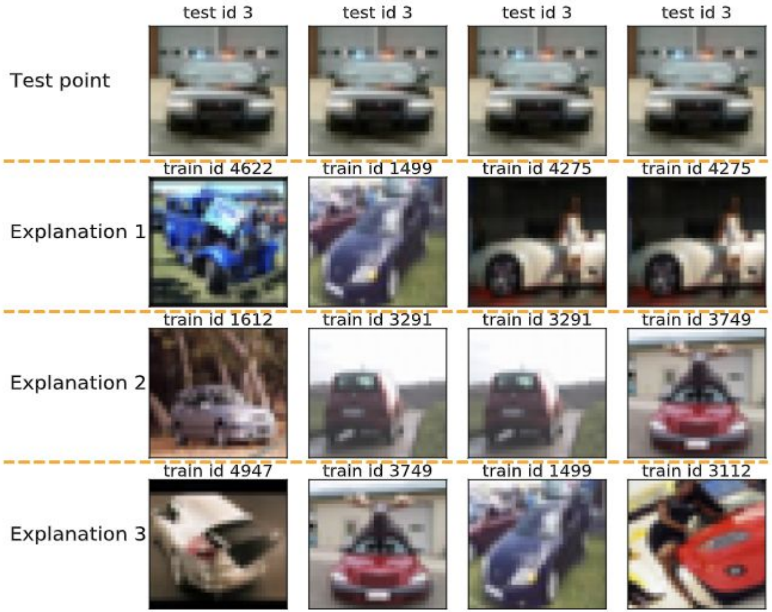
RPS-LJE with Faithful Explanations

RPS-LJE (different learning rate)



Faithful to the *given* model

RPS-l_2 (different l_2 weights)



Faithful to the *fine-tuned* model

Summary

- Identify two key drawbacks of RPS-l_2
  - Class-level explanation
  - Unfaithful to the given model (faithful to the fine-tuned model)
- Proposed an alternative sample-based explanation method with *Taylor Expansion* on Jacobian and derived a *RPS-like* data influence estimation
  - Instance-level explanation
  - Faithful to the given model
- Ability to explain common deep neural networks (e.g. ResNet, LSTM) as well as ensemble models like XGBoost classifiers by removing the l_2 requirement