Understanding Black-box Predictions via Influence Functions

Pang Wei Koh & Percy Liang

Outline

Why do we need this method?

How does it work?

What can it be used for?

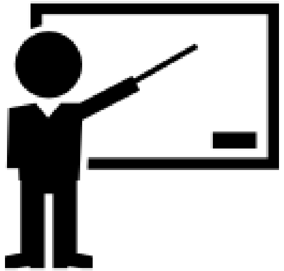
Summary

Why do we need this method?

- Authors views
- Existing approaches
- A new approach

Authors views

High performing



Provide explanation

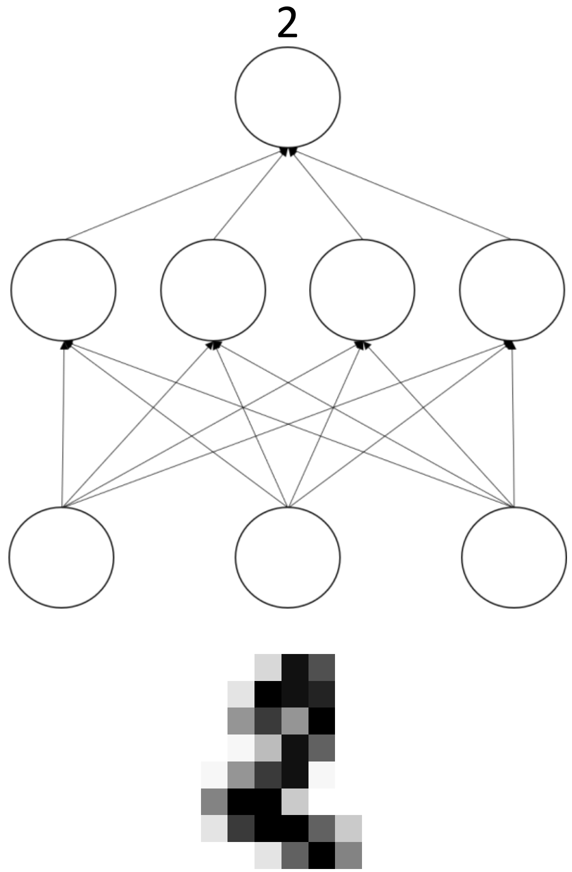
Discover new science

Improve models

Existing approaches

- Simplification
- Maximally activate neurons
- Find responsible parts

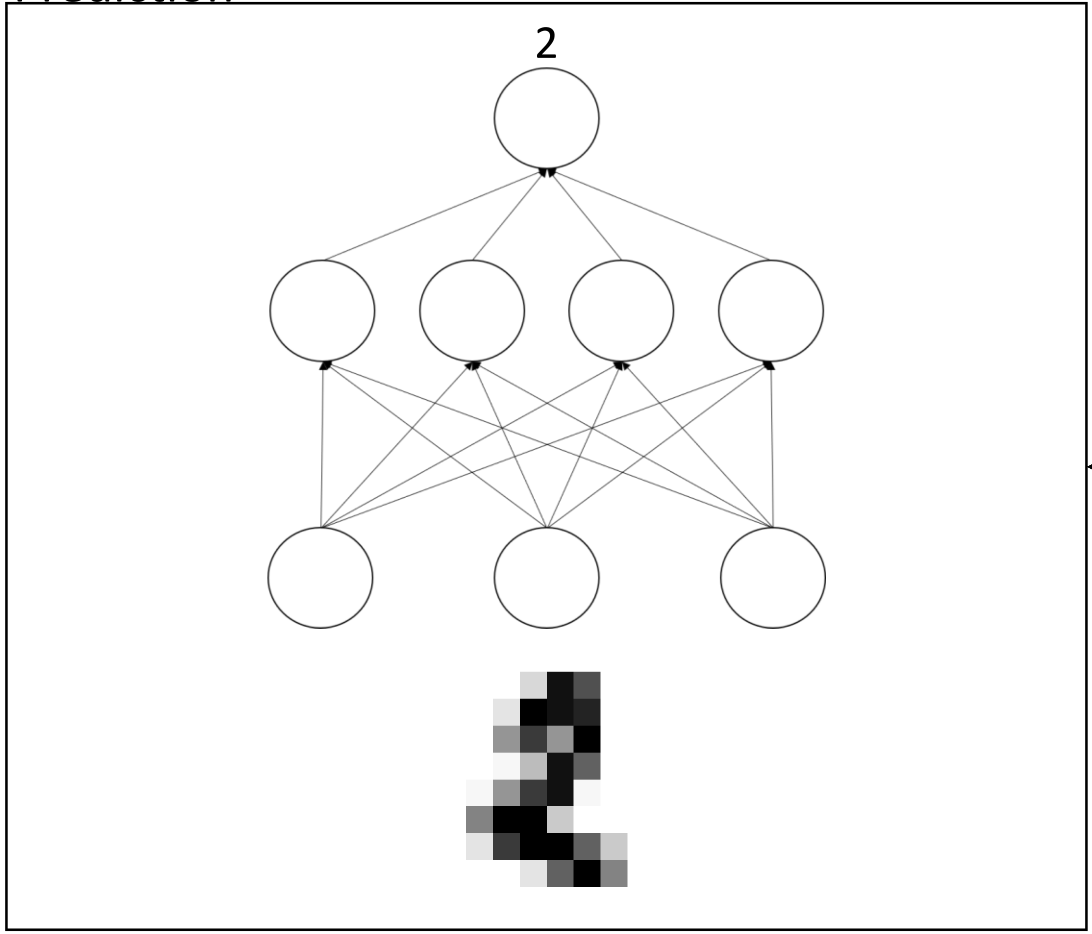
- See model as fixed



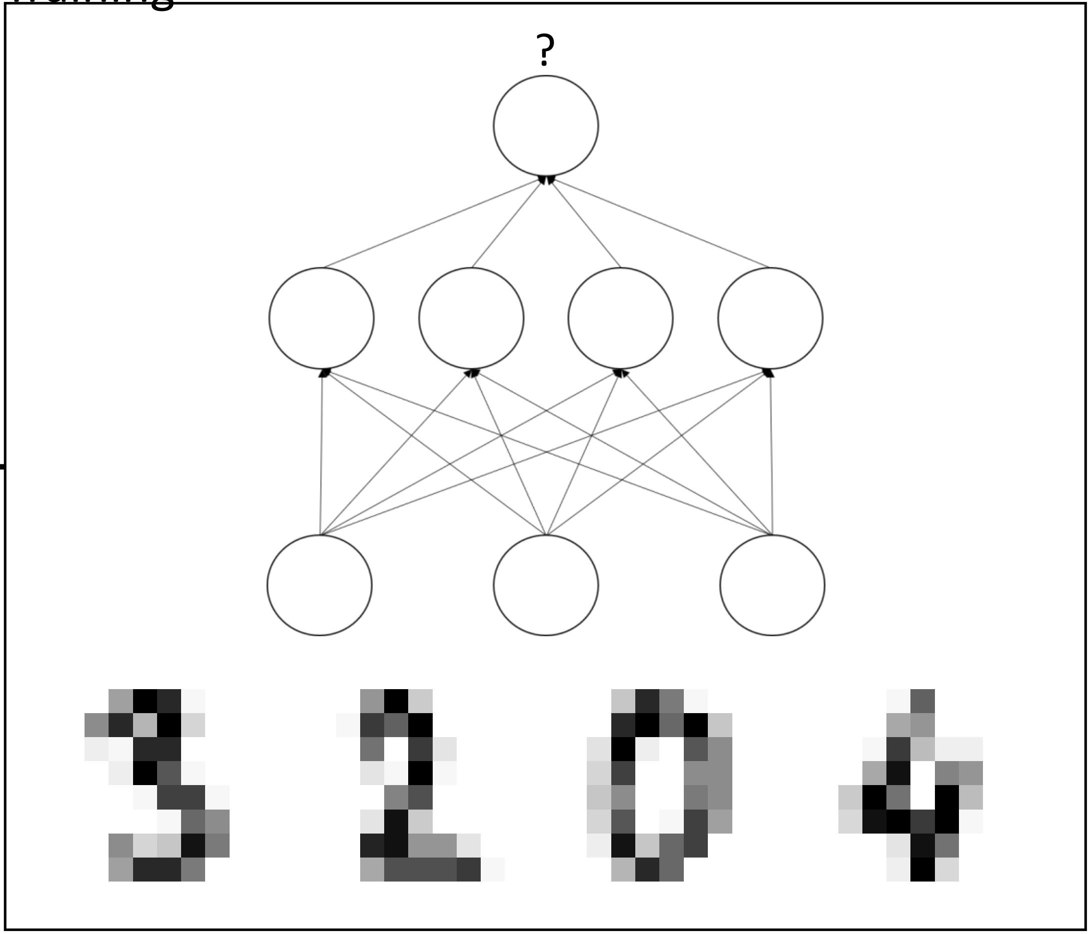
A new Approach

-Model is learned

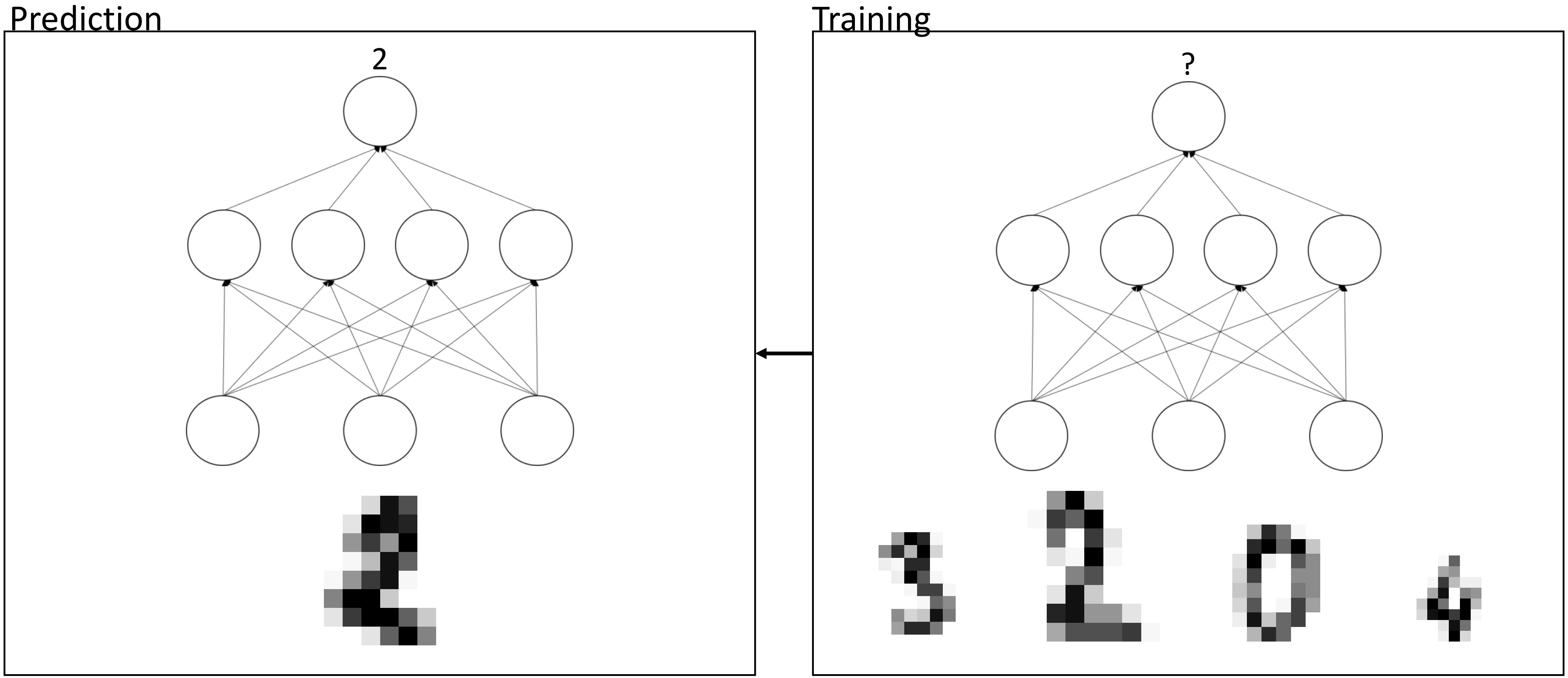
Prediction



Training



A new Approach
- Impact of training points



A new Approach
- Summary

How did the model come to its result?

↓

Which training points where most influential?

↓

What would happen if we change the weights?

How does it work?

- Approach
- Issues

Approach
- Fundamentals

Training points:

$$z_1, \dots, z_n \text{ with } z_i = (x_i, y_i) \in X \times Y$$

empirical risk minimizer:

$$\hat{\theta} = \underset{\theta \epsilon \Theta}{\operatorname{argmin}} \frac{1}{N} \Sigma_i^N L(z_i, \theta)$$

Approach
- Formalizing the problem

$$\hat{\theta}_{\epsilon,z} = \underset{\theta\epsilon\Theta}{\operatorname{argmin}} \frac{1}{N} \Sigma_i^N L(z_i, \theta) + \epsilon L(\mathrm{z}, \theta)$$

$$\Rightarrow \hat{\theta}_{\epsilon,z} - \hat{\theta}$$

- Problem: retraining expensive

Approach

- Influence Functions

- concept in robust statistics (Hampel, 1974)
- *effect of a change in one observation on an estimator* (Kahn, 2015)
- Based on Gâteaux derivative

Approach
- Influence of weight changes

$$I_{up,params}(z)$$
$$= -H_{\hat{\theta}}{}^{-1} \nabla_{\theta} L(z, \hat{\theta})$$

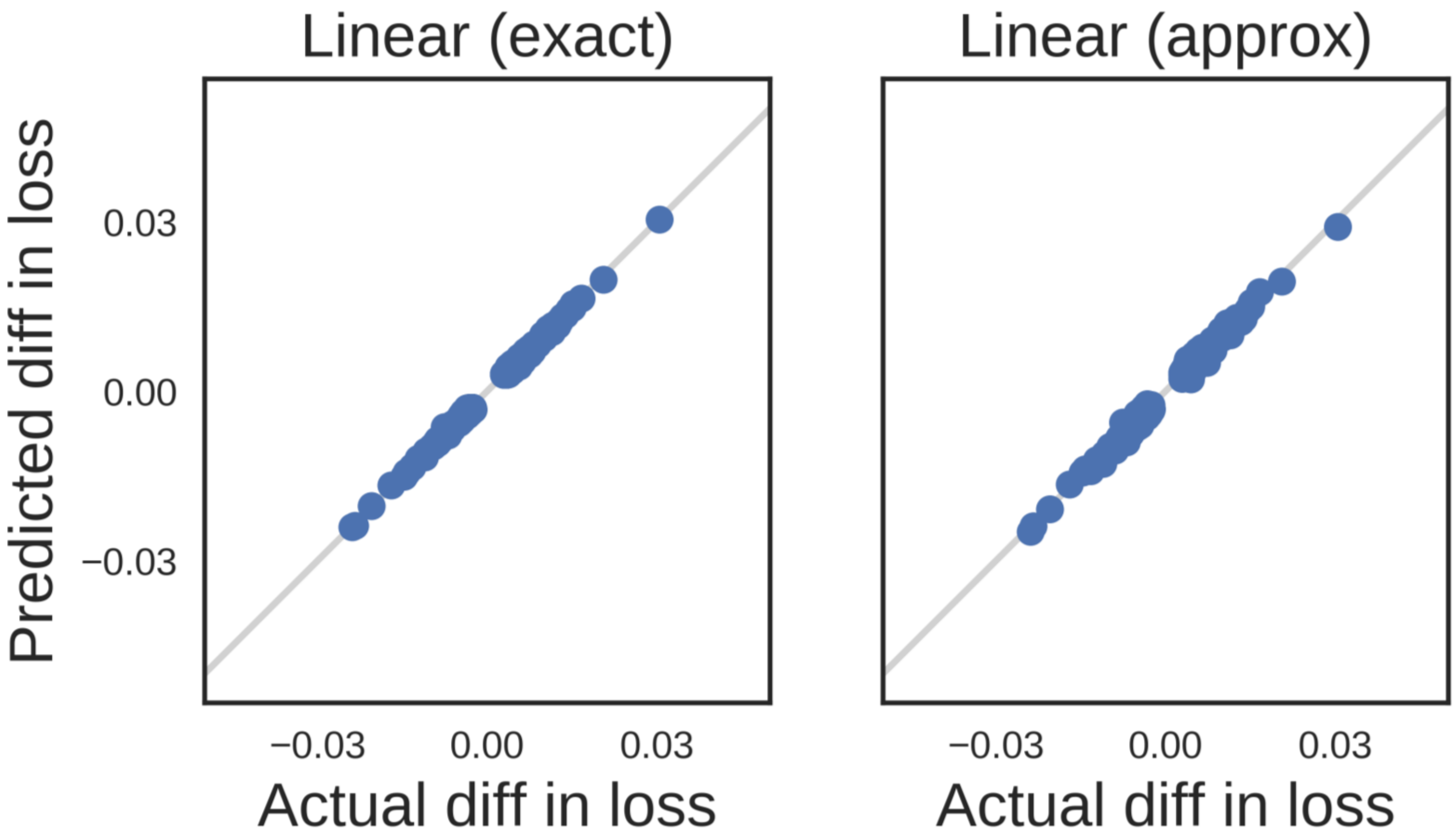
- Calculations for:
  - $I_{up,\text{loss}}$ using chain rule
  - $I_{\text{pert,loss}}$ analogous

Issues
- Efficiency

- We require ${H_{\hat{\theta}}}^{-1}$
- Training points: n, $\hat{\theta} \in \mathbb{R}^p \rightarrow O(np^2 + p^3)$

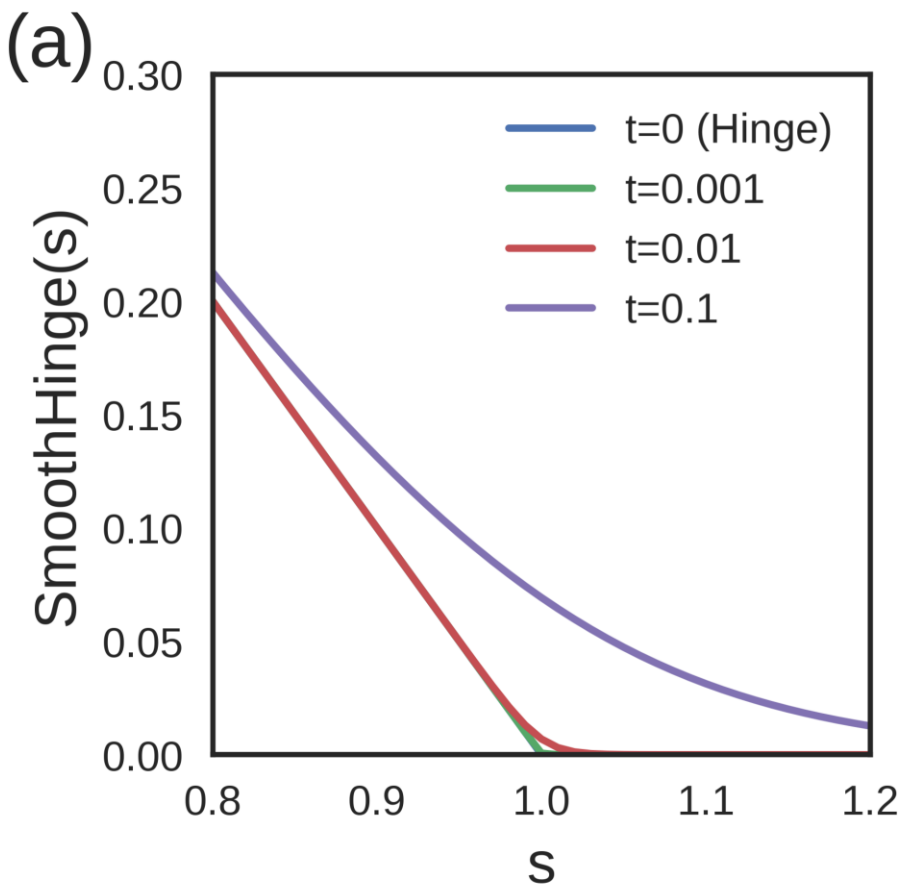
$$v \xrightarrow[\text{(Pearlmutter, 1994)}]{} H_{\hat{\theta}} v \xrightarrow[\substack{\text{CG (Martens, 2010)} \\ \text{SE (Agarwal et al., 2016)}}]{} {H_{\hat{\theta}}}^{-1} v$$

Issues
- Efficiency



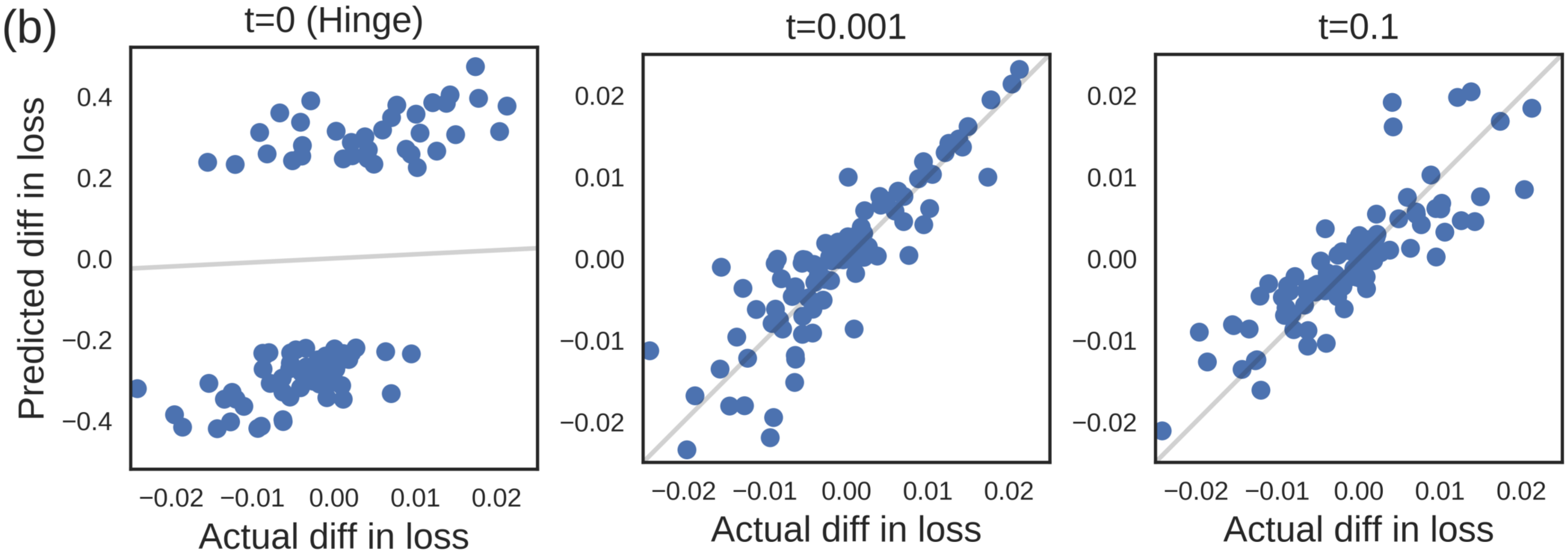
Issues

- Non-differentiable loss



Issues

- Non-differentiable loss



What can it be used for?

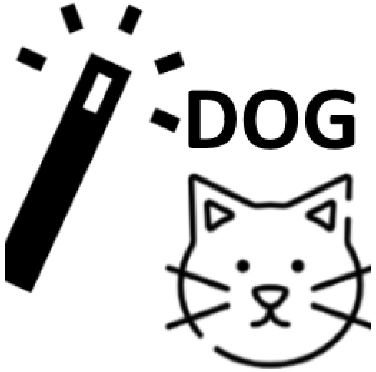
- Applications

Applications

Understanding

Debugging

Identifying mislabeled training data

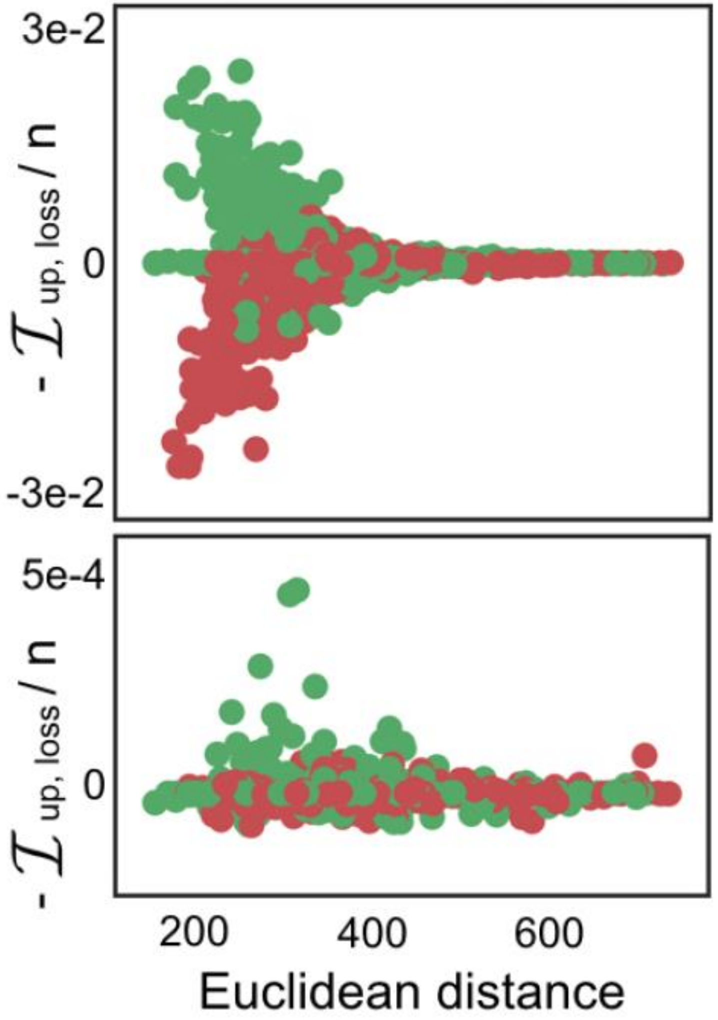


Generating adversarial training examples

Applications
- Understanding model prediction

RBF SVM

ANN



Applications

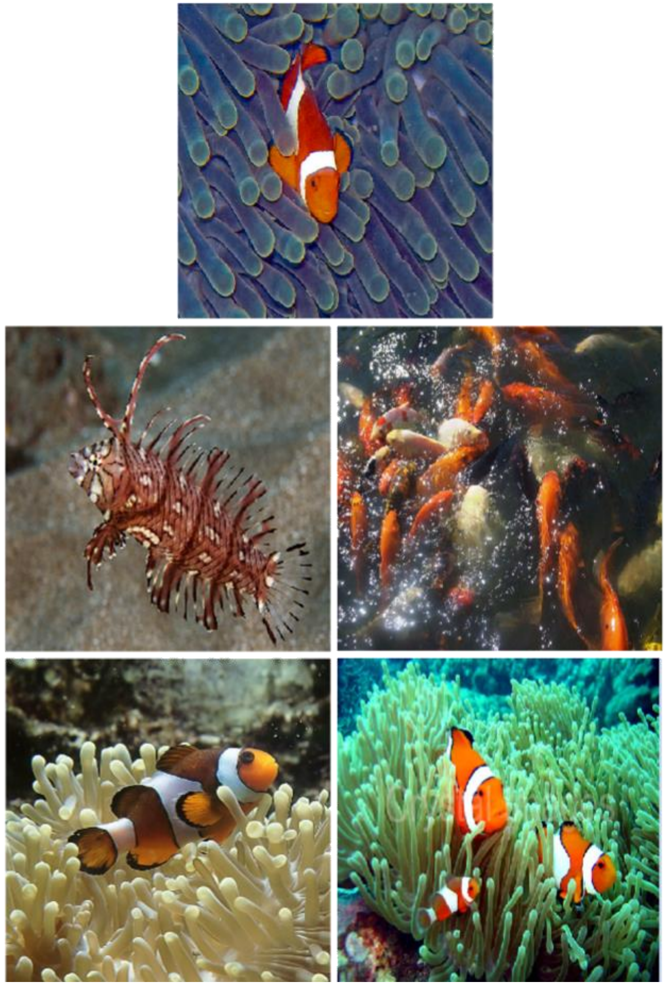
- Understanding model prediction

Most helpful training examples

Test example

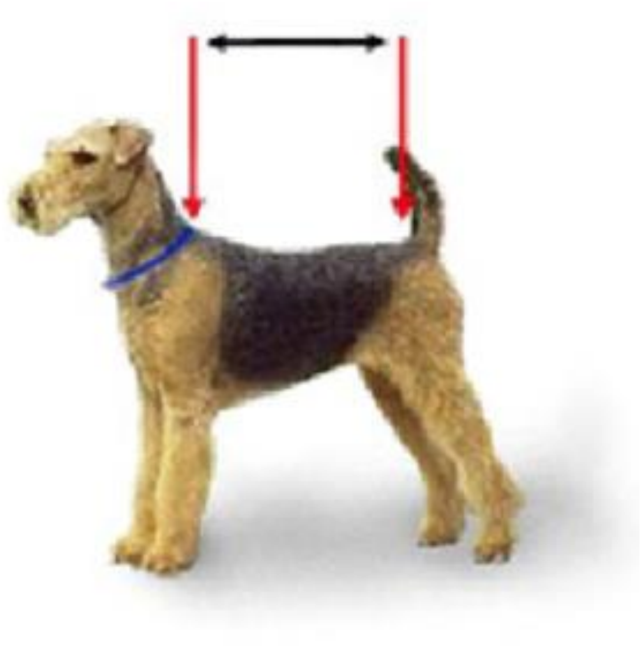
RBF SVM

ANN



Applications
- Understanding model prediction

Further helpful example for ANN



Summary

Summary

Based on training

Efficient calculation

Global trends undetectable

Measure upweighting

Many applications

Presentation by:
Philipp de Sombre
philipp.de.sombre@gmail.com
14 June 2018
Explainable Machine Learning
Ruprecht-Karls-Universität Heidelberg

Sources - Literature

- Agarwal, N., Bullins, B., and Hazan, E. *Second order stochastic optimization in linear time*. arXiv preprint arXiv:1602.03943, 2016.
- R. Hampel, Frank. (1974*). The Influence Curve and Its Role in Robust Estimation*. In *Journal of The American Statistical Association* - J AMER STATIST ASSN. 69. 383-393. 10.1080/01621459.1974.10482962.
- Jay Kahn, *Influence Functions for Fun and Profit,* Ross School of Business, University of Michigan, July 10, 2015.
- Koh, P.W. & Liang, P.. (2017). *Understanding Black-box Predictions via Influence Functions*. Proceedings of the 34th International Conference on Machine Learning, in PMLR 70:1885-1894
- Martens, J. Deep learning via hessian-free optimization. In *International Conference on Machine Learning (ICML)*, pp. 735–742, 2010.
- Pearlmutter, B. A. Fast exact multiplication by the hessian. *Neural Computation*, 6(1):147–160, 1994.

Sources - Assets

- Tachometer: by Freepik from www.flaticon.com
- Graph: by Gregor Cresnar from www.flaticon.com
- Magnifying glass: by Smashicons from www.flaticon.com
- Teacher: by Freepik from www.flaticon.com
- Lightbulb: by Freepik from www.flaticon.com
- Bug with target: by Freepik from www.flaticon.com
- Cats head: by Freepik from www.flaticon.com
- Magic wand: by Freepik from www.flaticon.com
- Clock: by Good Ware from www.flaticon.com
- Money: by Pause08 from www.flaticon.com
- Pushups: by Freepik from www.flaticon.com
- Gears: by Freepik from www.flaticon.com
- Globe with arrow: by Freepik from www.flaticon.com
- Scale: by Freepik from www.flaticon.com