Interpretation of Neural Networks is Fragile

Amirata Ghorbani, Abubakar Abid, James Zou

2018

Presented by Eli Lifland, 10/18/2019

Interpretation of Neural Networks

- Explanations for why an algorithm makes a decision
- Needed for trust between user and algorithm
- Motivating examples:
  - Doctors understanding diagnoses
  - Lender and borrower understanding credit risk

Interpretation Methods: Feature Importance

- Explains predictions in terms of importance of features
- Simple gradient method: detects sensitivity of score to perturbing each dimension
- Integrated gradients: gradients calculated with respect to several scaled versions of input
- DeepLIFT: Decomposes score backwards through network, layer-wise propagation (LRP) method

Interpretation Methods: Sample Importance

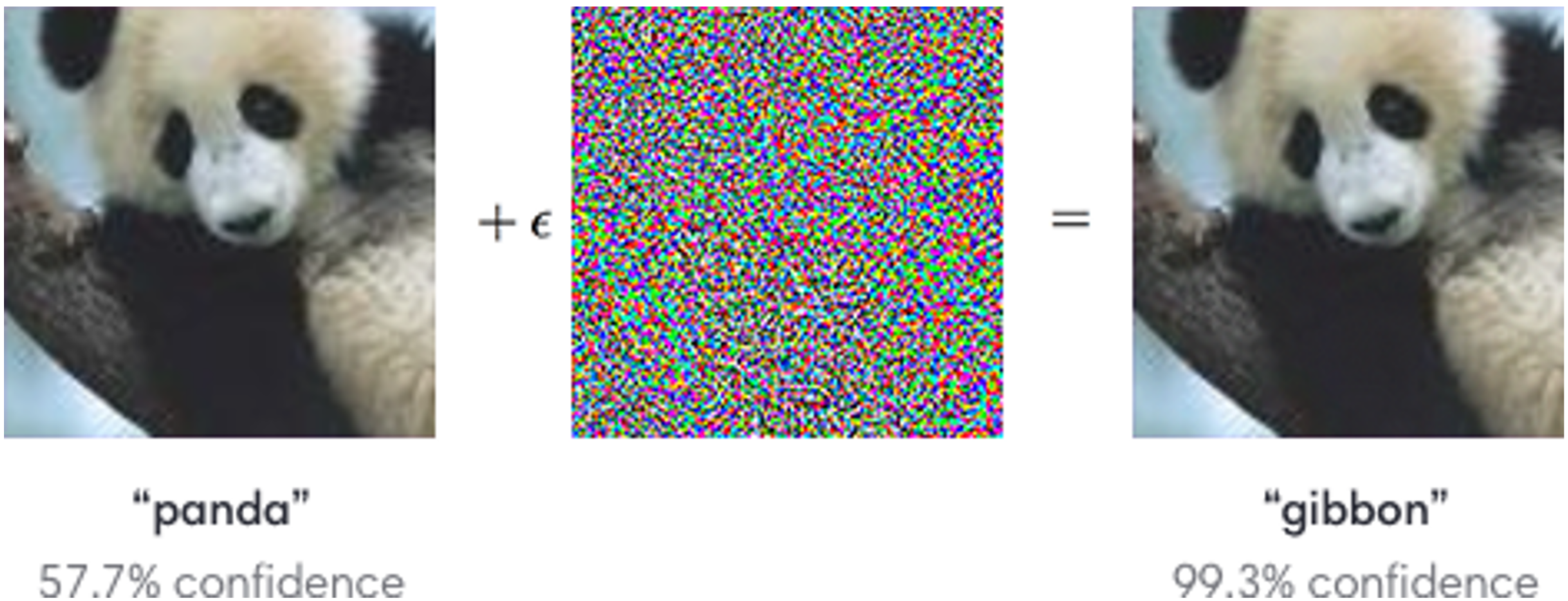
- Explains predictions in terms of importance of training examples
- Influence equation used to calculate influence of each example, derived by Koh and Liang (2017)

$$I(z_i, z_t) = -\nabla_\theta L(z_t, \hat{\theta})^\top H_{\hat{\theta}}^{-1} \nabla_\theta L(z_i, \hat{\theta}),$$

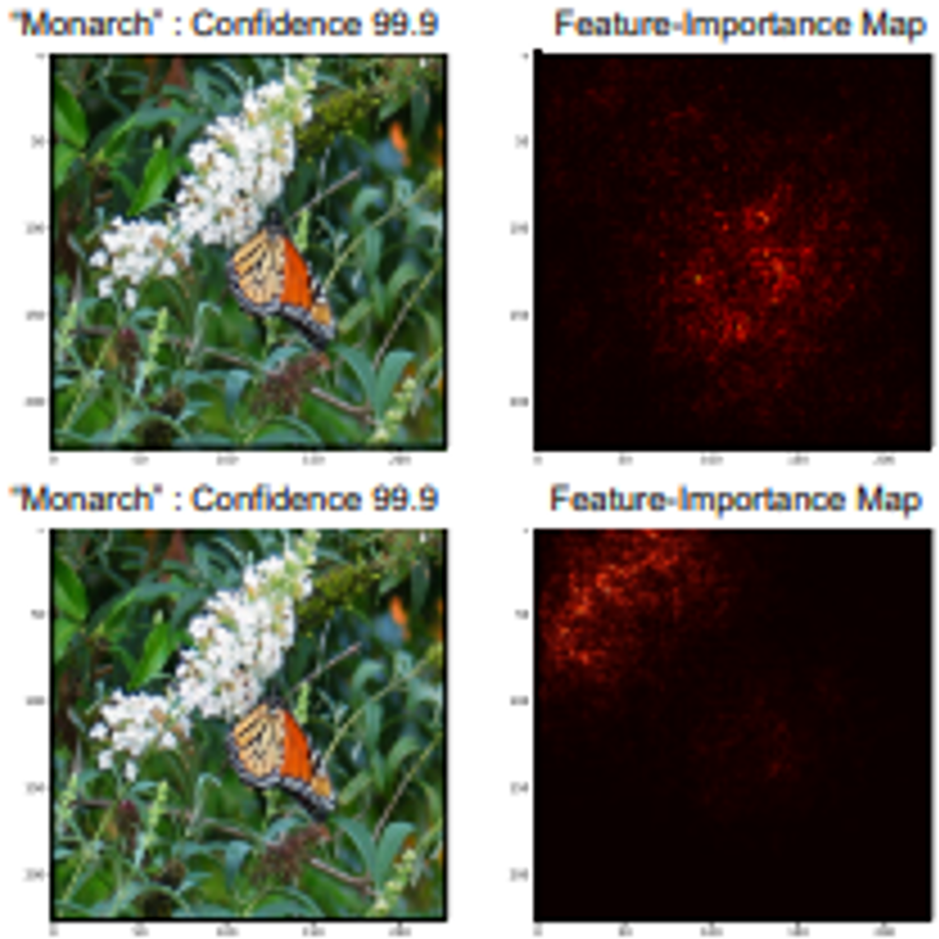
Importance of Robustness

- Interpretations not robust to indistinguishable perturbations may be security concern
  - Doctor selecting wrong intervention, e.g. location of biopsy
  - Incorrect causal conclusions

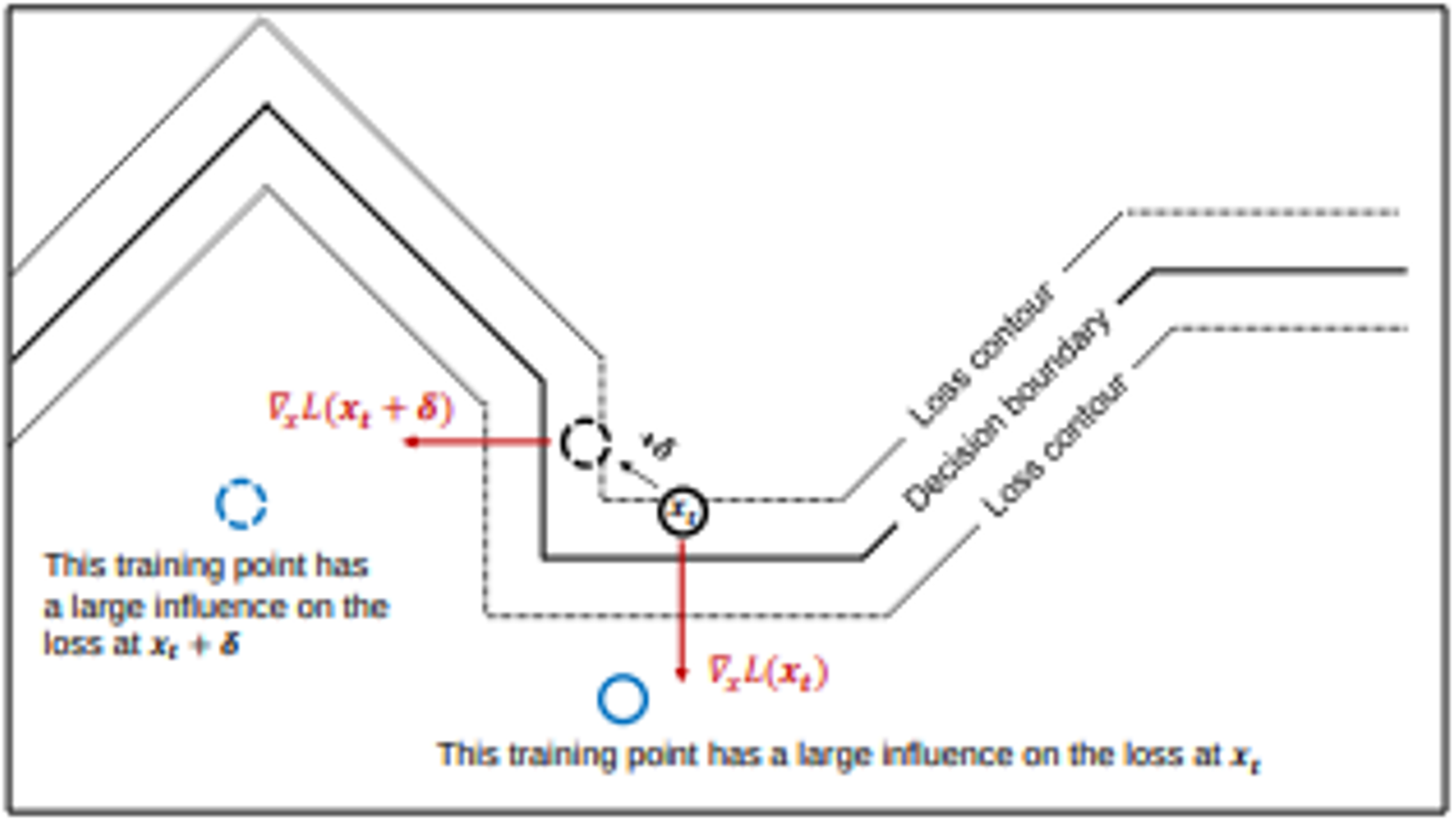
Adversarial Perturbations for Prediction



Adversarial Perturbations for Interpretation



Intuition for Fragility of Interpretation



Problem Statement

$$\arg\max_{\boldsymbol{\delta}} \mathcal{D}\left(\boldsymbol{I}(\boldsymbol{x}_t; \mathcal{N}), \boldsymbol{I}(\boldsymbol{x}_t + \boldsymbol{\delta}; \mathcal{N})\right)$$

$$\text{subject to: } ||\boldsymbol{\delta}||_\infty \leq \epsilon,$$

$$\text{Prediction}(\boldsymbol{x}_t + \boldsymbol{\delta}; \mathcal{N}) = \text{Prediction}(\boldsymbol{x}_t; \mathcal{N})$$

Attacking Feature Importance Methods

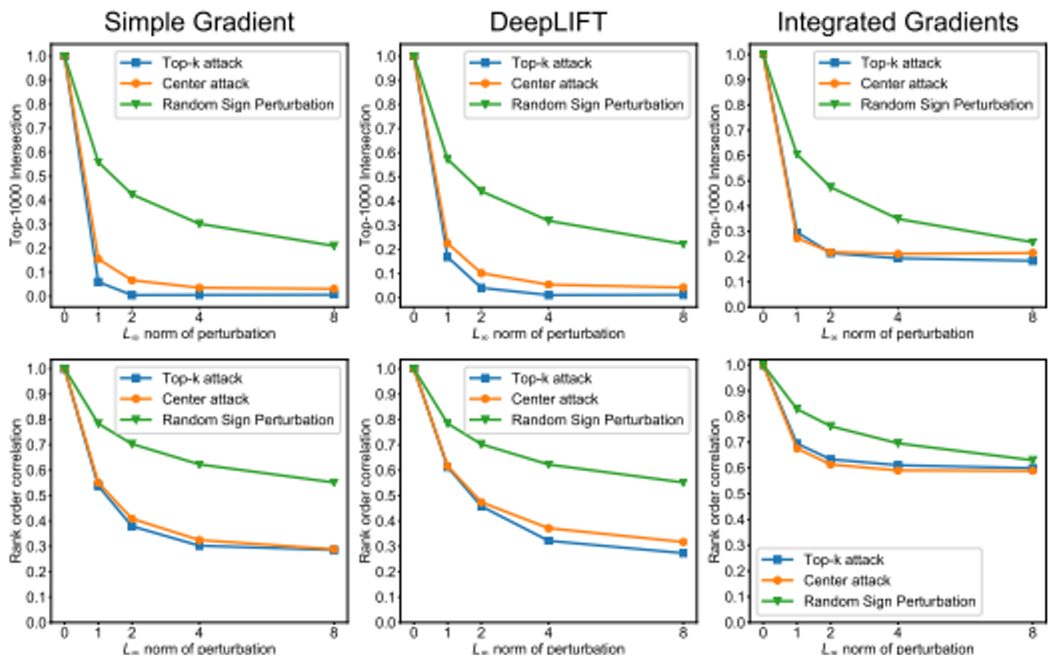
- Series of steps in direction which maximizes differentiable dissimilarity function between original, perturbed interpretation
    - Top-k attack: Decreases relative importance of k most important features
    - Mass-center attack for image data: maximizes spatial displacement of center of mass of feature importance map
    - Targeted attack for image data: Increases concentration of feature importance scores in pre-defined region of image

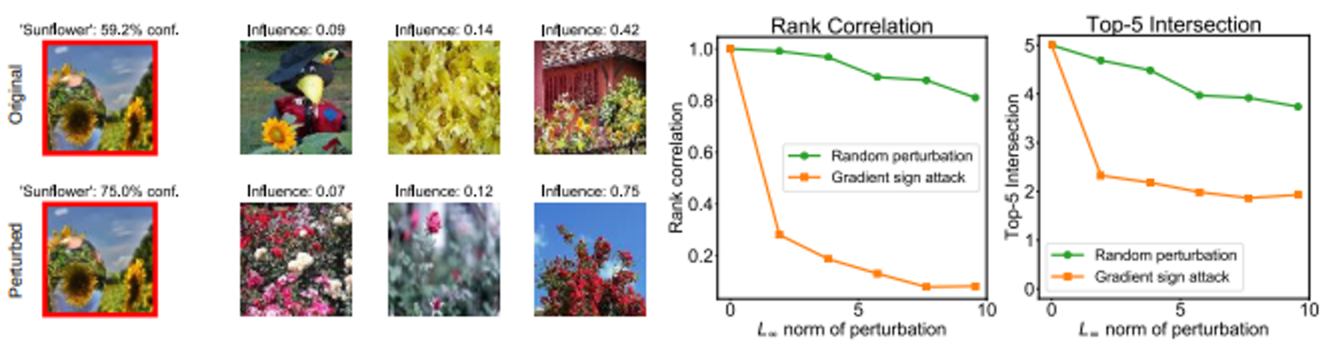
Attacking Influence Function

- Optimal single-step perturbation to decrease influence of 3 most influential training examples

$$\boldsymbol{\delta} = \epsilon \text{sign}(\nabla_{\boldsymbol{x}_t} I(z_i, z_t)) =$$
$$- \epsilon \text{sign}(\nabla_{\boldsymbol{x}_t} \nabla_\theta L(z_t, \hat{\theta})^\top \underbrace{H_{\hat{\theta}}^{-1} \nabla_\theta L(z_i, \hat{\theta})}_{\text{independent of } \boldsymbol{x}_t})$$

Results

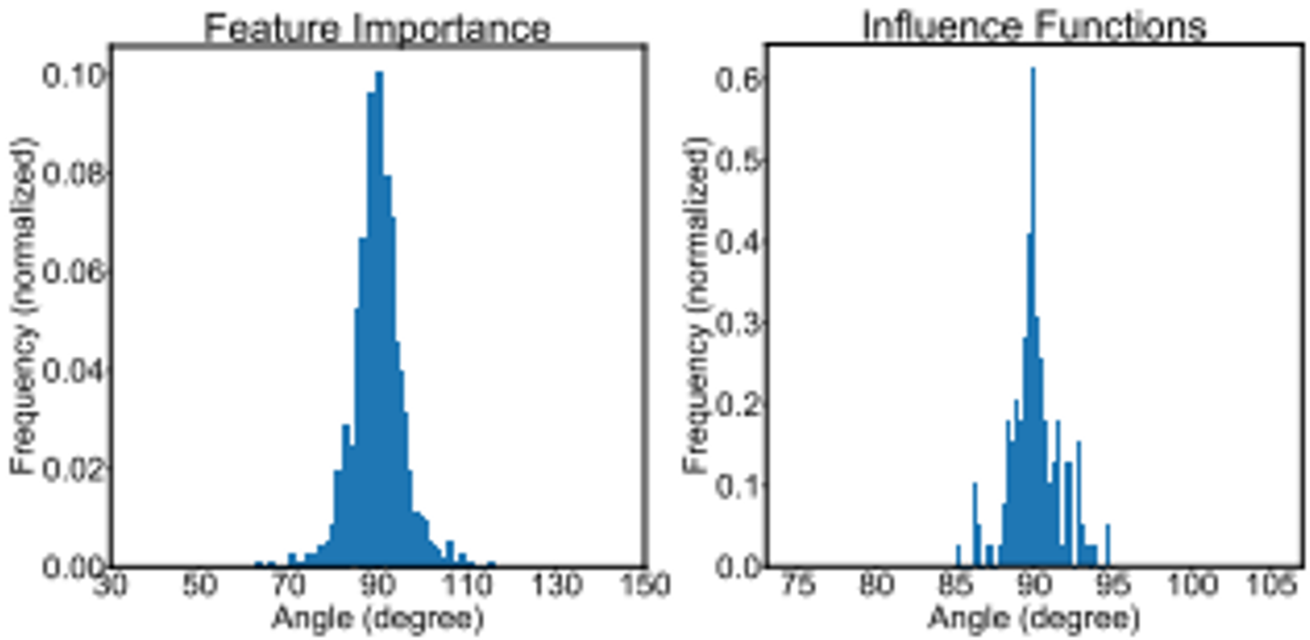

Simple Gradient
DeepLIFT
Integrated Gradients
Top-k attack
Center attack
Random Sign Perturbation
Top-1000 Intersection
Rank order correlation
L∞ norm of perturbation


'Sunflower': 59.2% conf.
Original
'Sunflower': 75.0% conf.
Perturbed
Influence: 0.09
Influence: 0.14
Influence: 0.42
Influence: 0.07
Influence: 0.12
Influence: 0.75
Rank Correlation
Top-5 Intersection
Rank correlation
Top-5 Intersection
Random perturbation
Gradient sign attack
L∞ norm of perturbation

Hessian Analysis

- Approximation of sensitivity of gradient-based interpretations to perturbation $\boldsymbol{\delta}$ is:
  - $\nabla_{\mathbf{x}}S(\mathbf{x}+\boldsymbol{\delta}) - \nabla_{\mathbf{x}}S(\mathbf{x}) \cong H\boldsymbol{\delta}$, where H is the Hessian
- Consider a linear model $\mathbf{w}^T\mathbf{x}$: $\nabla_{\mathbf{x}}S(\mathbf{x}) = \mathbf{w} \; \forall \; \mathbf{x}$. Thus the feature importance vector is robust
- Consider the same model followed by a non-linearity (e.g. softmax) $g(\mathbf{w}^T\mathbf{x})$. The change in feature importance map is now $H \cdot \boldsymbol{\delta} = \nabla^2_{\mathbf{x}}S \cdot \boldsymbol{\delta}$. $\nabla^2_{\mathbf{x}}S$ is no longer 0. Authors show that change in feature importance map grow with dimension of $\mathbf{w}$.
- Thus, non-linearity and high dimensionality are causes of lack of robustness of interpretations

Orthogonality of Fragile Directions



Conclusion

- Robustness of interpretation of a prediction is important and challenging
- Importance scores can be susceptible even just to random perturbations, but doubly so to targeted ones
- Potential defense techniques:
  - Discretizing inputs
  - Constraining non-linearity of networks