

# Interpretation of Neural Networks Is Fragile

01/24/2020

Presenter: Zijie Pan

<https://qdata.github.io/deep2Read/>

# Motivation

Prediction made by learning algorithm is important. Interpretations are needed to gain trusts. However, the robustness of interpretations are considered. It is disconcerting that indistinguishable images having very different salient features are classified as the same.

# Related Work

Prediction side, instead of interpretation side:

(Szegedy et al. 2013): indistinguishable images, different predictions

Interpretation Side:

- Feature Importance Interpretation: Simple gradient method, Integrated gradients, DeepLIFT (Testing sample)
- Sample Importance Interpretation (Training Samples)

Quantify Similarity of Interpretation:

- Spearman's rank order correlation
- Top-k intersection

# Claim / Target Task

Introduce a notion of adversarial perturbation to neural network.

Define fragile:

For a given neural network: indistinguish images with same predictions, yet very different interpretation.

# An Intuitive Figure Showing WHY Claim

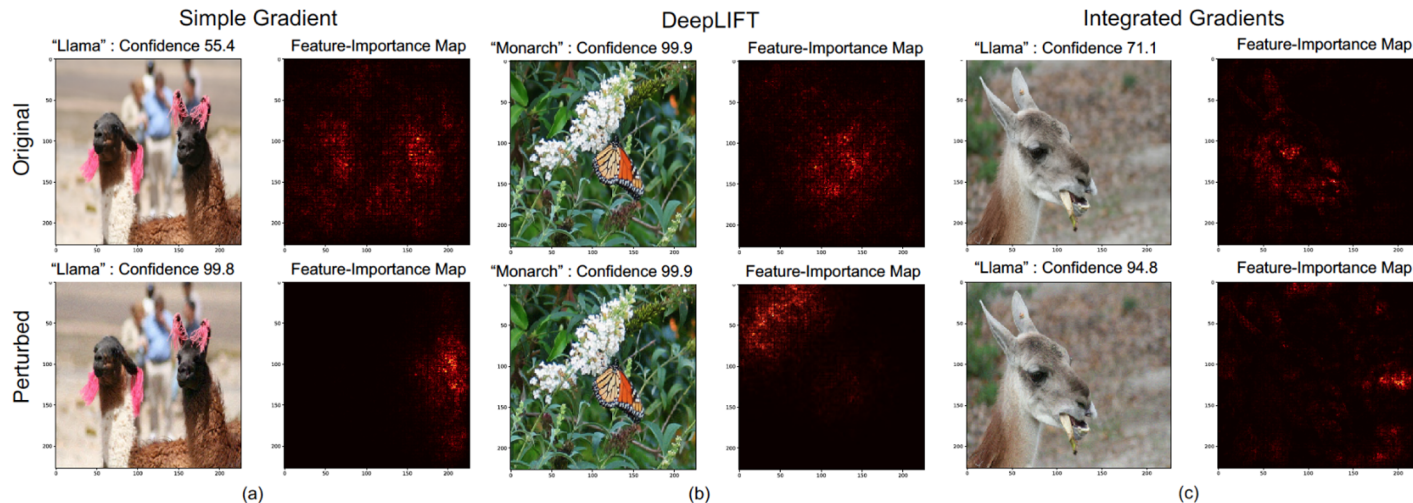
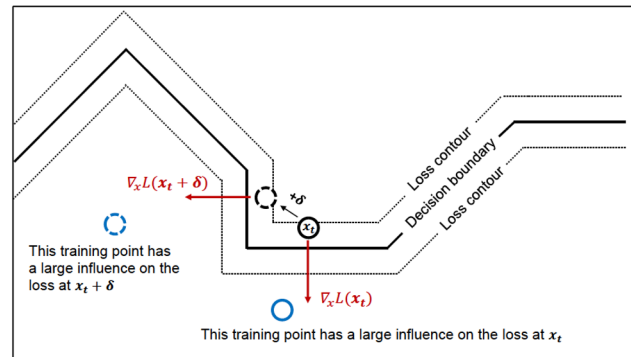


Figure 1: **Adversarial attack against feature-importance maps.** We generate feature-importance scores, also called saliency maps, using three popular interpretation methods: (a) simple gradients, (b) DeepLIFT, and (c) integrated gradients. The **top row** shows the original images and their saliency maps and the **bottom row** shows the perturbed images (using the center attack with  $\epsilon=8$ , as described in Section 2) and corresponding saliency maps. In all three images, the predicted label does not change from the perturbation; however, the saliency maps of the perturbed images shifts dramatically to features that would not be considered salient by human perception.

# Proposed Solution

- Random Sign Perturbation
- Iterative attacks against feature importance methods
  - Top-k : decreasing the relative importance of the k initially most important input features
  - Maximum spatial displacement of mass-center images
  - Semantically meaningful:targeted attacks: increase the concentration of feature importance scores in the predefined region of input image
- Gradient sign attack against influence functions

# Implementatio

## Algorithm 1 Iterative feature importance Attacks

**Input:** test image  $\mathbf{x}_t$ , maximum norm of perturbation  $\epsilon$ , normalized feature importance function  $I(\cdot)$ , number of iterations  $P$ , step size  $\alpha$

Define a dissimilarity function  $D$  to measure the change between interpretations of two images:

$$D(\mathbf{x}_t, \mathbf{x}) = \begin{cases} -\sum_{i \in B} I(\mathbf{x})_i & \text{for } \mathbf{top-k} \text{ attack} \\ \sum_{i \in \mathcal{A}} I(\mathbf{x})_i & \text{for } \mathbf{targeted} \text{ attack} \\ \|\mathbf{C}(\mathbf{x}) - \mathbf{C}(\mathbf{x}_t)\|_2 & \text{for } \mathbf{mass-center} \text{ attack,} \end{cases}$$

where  $B$  is the set of the  $k$  largest dimensions of  $I(\mathbf{x}_t)$ ,  $\mathcal{A}$  is the target region of the input image in targeted attack, and  $\mathbf{C}(\cdot)$  is the center of feature importance mass<sup>a</sup>.

Initialize  $\mathbf{x}^0 = \mathbf{x}_t$

**for**  $p \in \{1, \dots, P\}$  **do**

Perturb the test image in the direction of signed gradient<sup>b</sup> of the dissimilarity function:

$$\mathbf{x}^p = \mathbf{x}^{p-1} + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}} D(\mathbf{x}_t, \mathbf{x}^{p-1}))$$

If needed, clip the perturbed input to satisfy the norm constraint:  $\|\mathbf{x}^p - \mathbf{x}_t\|_{\infty} \leq \epsilon$

**end for**

Among  $\{\mathbf{x}^1, \dots, \mathbf{x}^P\}$ , return the element with the largest value for the dissimilarity function and the same prediction as the original test image.

# Data Summary

Image net

CIFAR-10



# Experimental Results

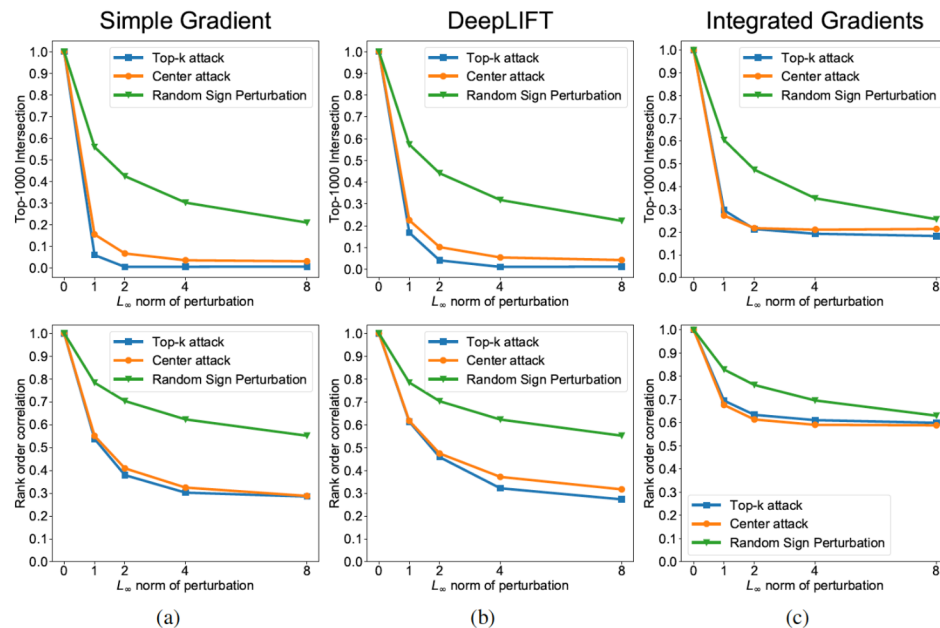


Figure 3: **Comparison of adversarial attack algorithms on feature-importance methods.** Across 512 correctly-classified ImageNet images, we find that the top- $k$  and center attacks perform similarly in top-1000 intersection and rank correlation measures, and are far more effective than the random sign perturbation at demonstrating the fragility of interpretability, as characterized through top-1000 intersection (**top**) as well as rank order correlation (**bottom**).

# Conclusion and Future Work

Interpretations are vulnerable to perturbation

Interpretation arises as a consequence of high dimensionality and non-linearity

Apply not only on image data

# References