# Towards Explanation of DNN-based Prediction with Guided Feature Inversion


Mengnan Du, Ninghao Liu, Qingquan Song, Xia Hu

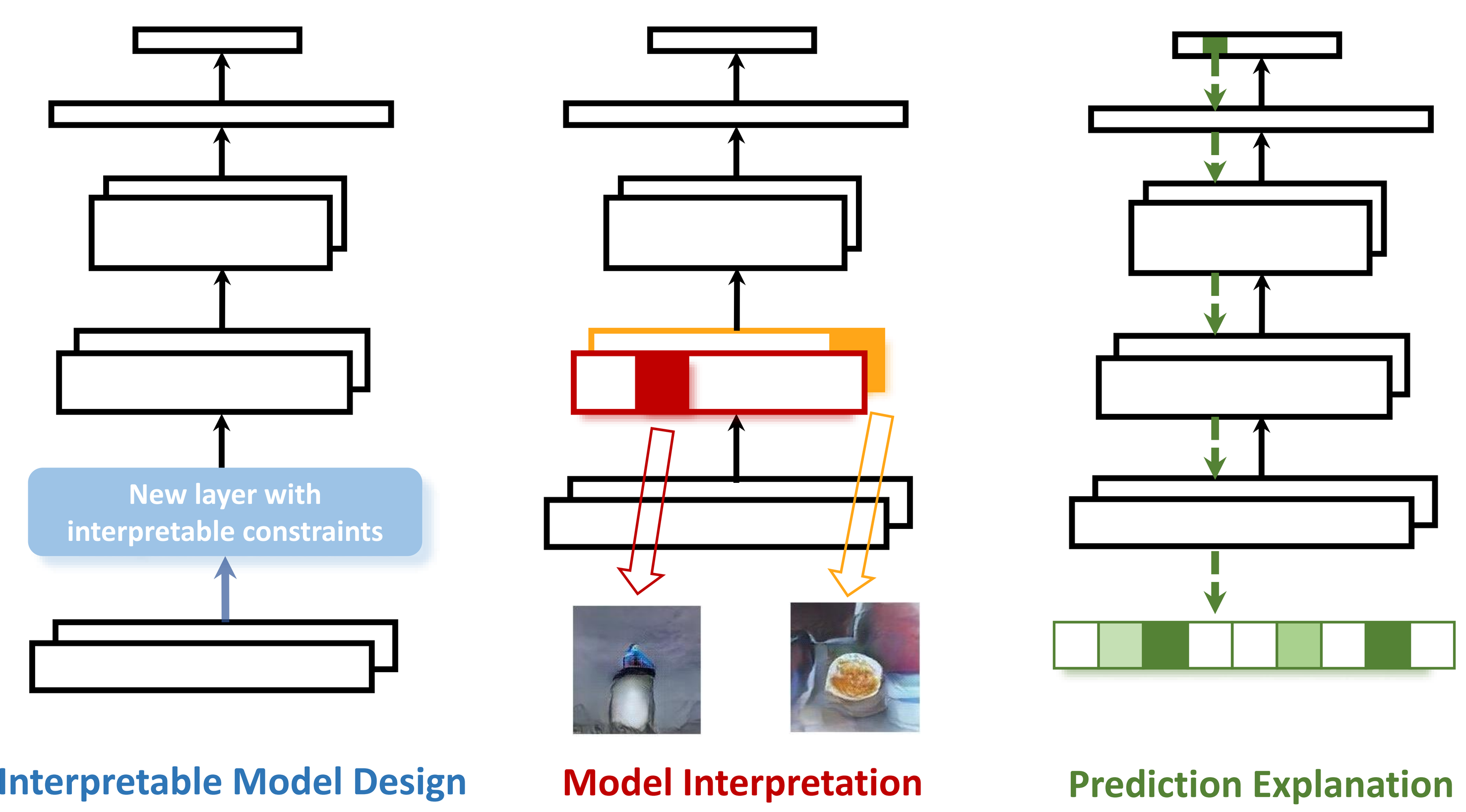
Department of Computer Science and Engineering, Texas A&M University

{dumengnan,nhliu43,song_3134,xiahu}@tamu.edu


## Introduction

DNN interpretation techniques can be grouped into three categories [1]:

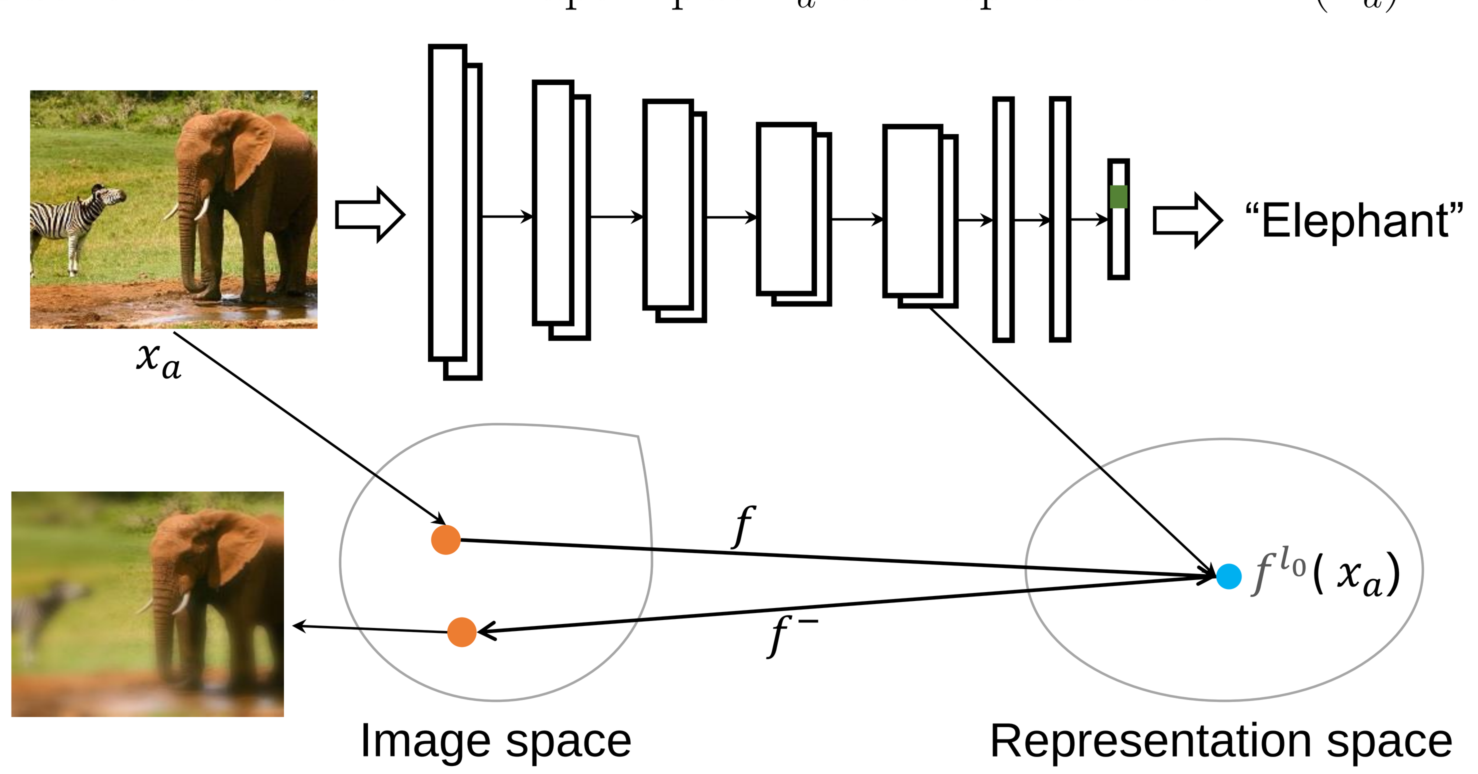
- Design interpretable network architectures
- Post-hoc interpret a pre-trained model
- Post-hoc explain a prediction of a pre-trained model




In this paper, we provide **post-hoc explanation** for predictions made by DNNs in order to promote the interpretability of DNNs.

## Motivation

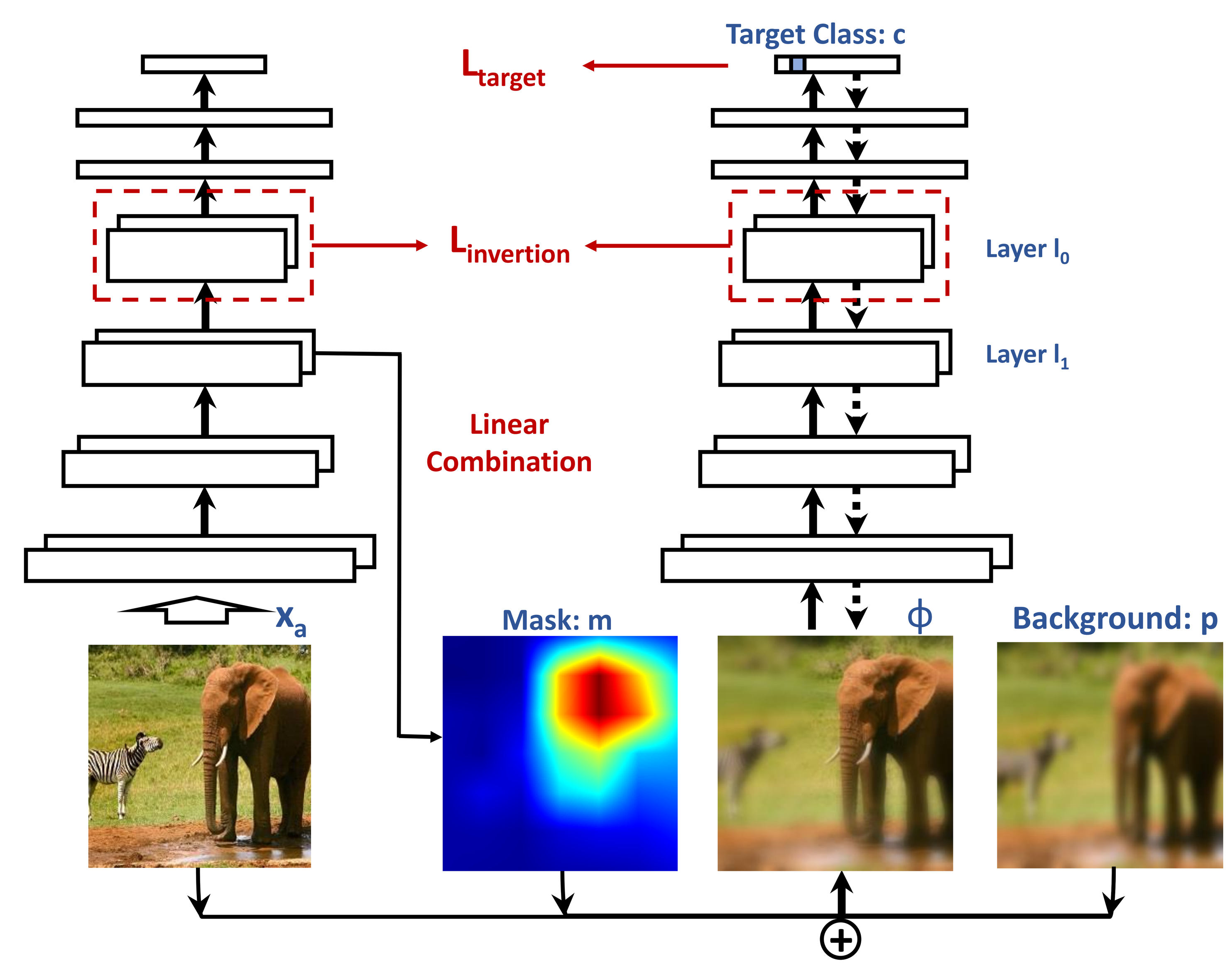
Subnetwork of a DNN maps input $\mathbf{x}_a$ to a representation $\mathbf{f}^{l_0}(\mathbf{x}_a)$




### Motivations based on feature inversion [2]

- DNN compresses the input information as the layer goes deeper
- Inversion reveals the amount of information contained in a layer
- Leverage the inversion of deep representations to derive more accurate interpretations

## Proposed Approach




- **Interpretation via Guided feature inversion**
  The expected inversion input is reformulated as the weighted sum of the original image $\mathbf{x}_a$ and another noise background image $\mathbf{p}$:

$$\Phi(\mathbf{x}_a, \mathbf{m}) = \mathbf{x}_a \odot \mathbf{m} + \mathbf{p} \odot (1 - \mathbf{m}). \quad (1)$$

  We use perceptual loss to minimize the representation difference between the original input $\mathbf{x}_a$ and the inverted input $\Phi(\mathbf{x}_a, \mathbf{m})$:

$$L_{\text{inversion}}(\mathbf{x}_a, \omega) = \|\mathbf{f}^{l_0}(\Phi(\mathbf{x}_a, \omega)) - \mathbf{f}^{l_0}(\mathbf{x}_a)\|^2 + \gamma \cdot \|\omega\|_1, \quad (2)$$

- **Class-Discriminative Interpretation**
  We further use target neuron in the output layer to make the final interpretation results class-discriminative:

$$L_{\text{target}}(\mathbf{x}_a, \omega) = -\mathbf{f}^L_c(\Phi(\mathbf{x}_a, \omega)) + \lambda \mathbf{f}^L_c(\Phi_{bg}(\mathbf{x}_a, \omega)) + \delta \cdot \|\omega\|_1, \quad (3)$$

- **Regularization by Utilizing Intermediate Layers**
  We build the weight mask $\mathbf{m}$ as the weighted sum of the channels at a specific layer $l_1$:

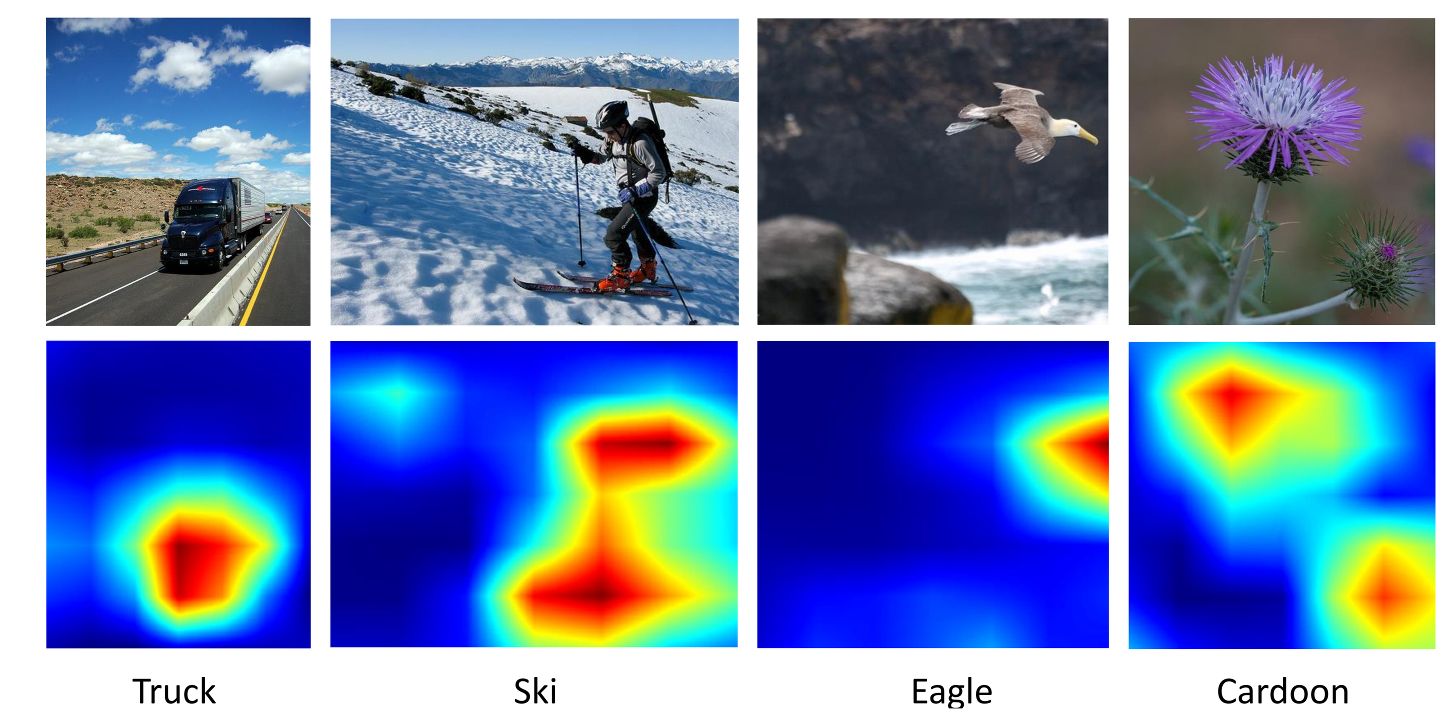
$$\mathbf{m} = \sum_i \omega_i \mathbf{f}^{l_1}_i(\mathbf{x}_a). \quad (4)$$

Leveraging the integration of the intermediate activation values as masks, we further lower the possibility to produce artifacts and increase the optimization efficiency.

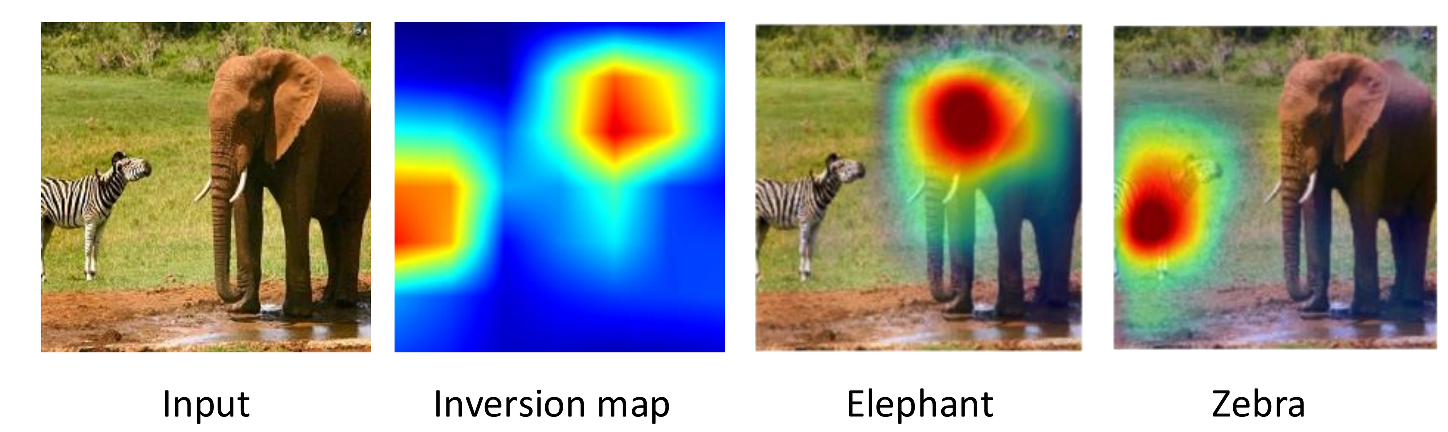
## Experimental Results

### 1. Visualization results on ImageNet dataset

- Interpretation results for four illustrative instances




- Class discriminability of our algorithm




### 2. Quantitative Evaluation

- Test the localization performance by applying the generated saliency maps to weakly supervised object localization tasks.

| | Grad | GuidedBP | LRP | CAM | Mask | Real | Ours |
|---|---|---|---|---|---|---|---|
| $\alpha$ | 5.0 | 4.5 | 1.0 | 1.0 | 0.5 | - | 1.1 |
| Error (%) | 41.7 | 42.0 | 57.8 | 48.1 | 43.2 | 36.9 | 38.2 |

## Acknowledgements


The authors are thankful to the helps from collaborators at DATA Lab.


## References


[1] Mengnan Du, Ninghao Liu, and Xia Hu.
Techniques for interpretable machine learning.
*Communications of the ACM (CACM)*, 2019.

[2] Aravindh Mahendran and Andrea Vedaldi.
Understanding deep image representations by inverting them.
In *CVPR*, 2015.