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.

Proposed Approach

- **Interpretation via Guided feature inversion**
  The expected inversion input is reformulated as the weighted sum of the original image \( x_a \) and another noise background image \( p \):
  \[
  \Phi(x_a, m) = x_a \odot m + p \odot (1 - m),
  \]
  We use perceptual loss to minimize the representation difference between the original input \( x_a \) and the inverted input \( \Phi(x_a, m) \):
  \[
  L_{\text{inversion}}(x_a, \omega) = \| \Phi(x_a, \omega) - \Phi(x_a) \|^2 + \gamma \cdot \| \omega \|_1,
  \]

- **Class-Discriminative Interpretation**
  We further use target neuron in the output layer to make the final interpretation results class-discriminative:
  \[
  L_{\text{target}}(x_a, \omega) = -\alpha L^2(\Phi(x_a, \omega)) + \lambda L^2(\Phi_{bg}(x_a, \omega)) + \delta \cdot \| \omega \|_1,
  \]

- **Regularization by Utilizing Intermediate Layers**
  We build the weight mask \( m \) as the weighted sum of the channels at a specific layer \( l_i \):
  \[
  m = \sum_i \omega_i f_i^l(x_a).\]

Experimental Results

1. **Visualization results on ImageNet dataset**
   - Interpretation results for four illustrative instances

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

Acknowledgements

The authors are thankful to the helps from collaborators at DATA Lab and are grateful for generous support by DARPA and NSF.

References

Techniques for interpretable machine learning.
Communications of the ACM (CACM), 2019.

Understanding deep image representations by inverting them.
In CVPR, 2015.