Global Explanations for Image Classifiers (Student Abstract)

Bhavan K. Vasu, Prasad Tadepalli

Oregon State University, Corvallis, Oregon 97331
vasub@oregonstate.edu, prasad.tadepalli@oregonstate.edu

Abstract

We hypothesize that deep network classifications of complex scenes can be explained using sets of relevant objects. We employ beam search and singular value decomposition to generate local and global explanations that summarize the deep model's interpretation of a class.

Introduction

Deep networks have achieved tremendous success in recent years, yet they are generally treated as black boxes that accept data and are optimized toward the desired objective. We attempt to demystify deep network decisions with the help of symbolic representations. Explanations can be either local or global; where a local explanation explains decisions at the image level, the global explanation explains decisions at a class or dataset level. A global explanation of a dataset with fine-grained classes could answer questions like - "What makes a luxurious bedroom different from a normal bedroom". The insight obtained for such decisions could lead to building applications geared towards customized interior design to improve the aesthetic look of a real-world scene. Alternatively, it could be used adversarially to affect other deep models or obtain human-interpretable rule-based models. In addition to the above applications, the proposed method could also be used for identifying anomalous training examples by looking for the disparity between local and global explanations.

Saliency maps are often used to explain single images. However, their interpretation is often subjective and they are further maligned by cognitive biases such as overgeneralization, correlation fallacy, and confirmation bias (Kliegr, Bahník, and Fürnkranz 2021). While structured explanations presented in (Shitole et al. 2021) provide a local quantitative measure and graphical representations, they are not expressed in terms of human-understandable language.

Approach

Notation. The dataset we are attempting to explain is $\mathcal{D} = \{(x_i, y_i, s_i)|i = 1, \ldots, N\}$. It contains N input images $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}, y_i \in \{1, 2, \ldots, c\}$ categorical labels indicating the image scene and a segmentation map s_i of the same spatial resolution as x_i. Let $s_i = \{s_i^1, s_i^2, \cdots s_i^p\}$ be a set of binary segmentation maps with respect to p object categories present in x_i.

For a given image x_i, $f(x_i, y_i; \theta)$ is a deep convolutional network that predicts y_i with probability p_i. Following (Shitole et al. 2021), we utilize beam search to search over the subsets of s_i to find minimal subsets that result in a score of at least $0.95p_i$. We call such subsets minimal sufficient explanations (MSXs). We restrict the beam search to a subset of objects by blurring its complement objects in the input to the neural network model. In previous work (Shitole et al. 2021), it has been found that a single image might contain multiple MSXs.

Symbolic Representation. To represent the MSXs consistently across all images of a class, we propose to represent them as a sparse matrix $\mathbf{G}_{y_c} \in R^{k \times m}$ containing information regarding the presence or absence of each object in the MSX. Let k represent the number of objects and m be the number of local MSXs for class y_c across the whole dataset. A value of 1 at position $\mathbf{G}_{y_c}[i, j]$ indicates the presence of object i in the MSX j for class y_c while 0 represents its absence.

Global Explanations. We formulate the problem of identifying global explanations as one of finding a low-rank representation $\mathbf{G}^r_{y_c}$ of $\mathbf{G}_{y_c}$, where the rank r of $\mathbf{G}^r_{y_c}$ is much lower than m, i.e., $r \leq min\{k, m\}$. The low-rank representation problem can be reformulated as the following optimization problem.

$$\min_{\mathbf{G}_{y_c} \in R^{k \times m}, rank(\mathbf{G}^r_{y_c}) \leq r} \|\mathbf{G}_{y_c} - \mathbf{G}^r_{y_c}\|_F^2 \qquad (1)$$

where $\mathbf{G}^r_{y_c}$ is the low rank representation of $\mathbf{G}_{y_c}$. Although the constraint $rank(\mathbf{G}^r_{y_c}) \leq r$ is non-convex from (Eckart and Young 1936), we know that it can be optimally solved for a low-rank matrix of rank r using methods such as truncated Singular Value Decomposition (SVD).

Preliminary Results

To investigate the ability of the proposed approach to explain deep CNN decisions, we experiment with a VGG19 network trained on the AED20K dataset (Zhou et al. 2019). The VGG19 network was trained for 40 epochs to classify between the top 10 scene categories with the most images

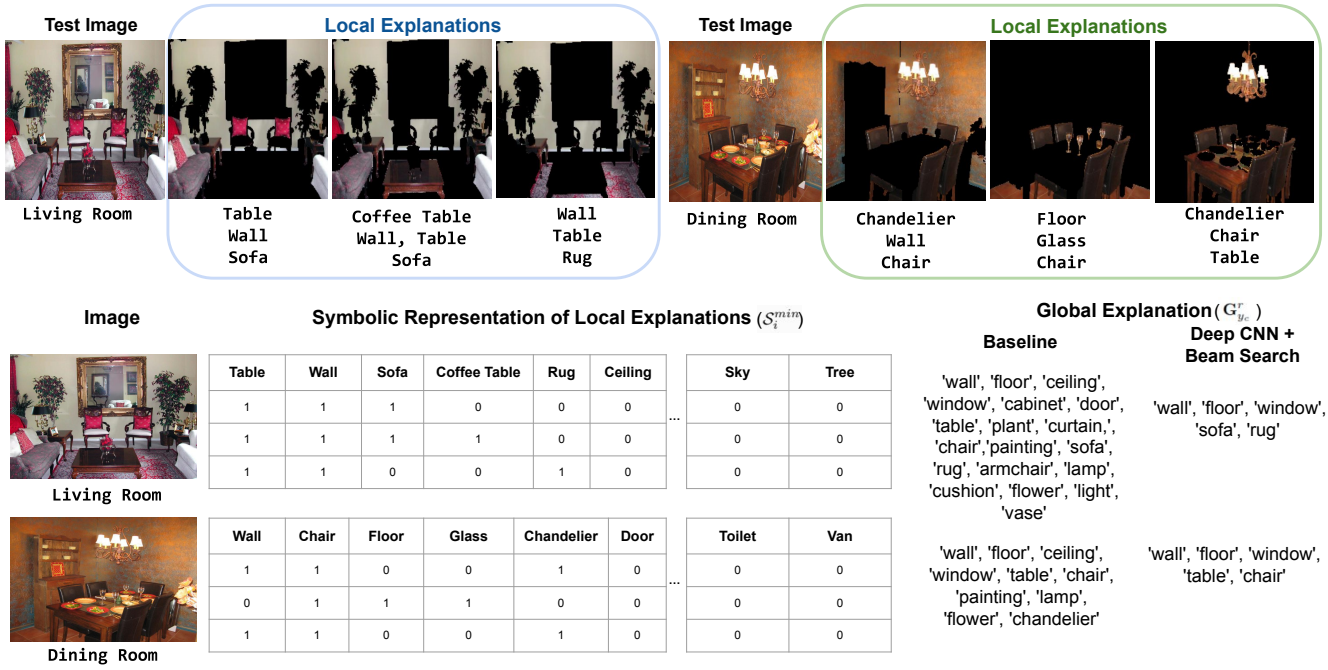

Figure 1: Figure shows the visualization of multiple local explanations along with their symbolic representations $\mathcal{S}_i^{min}$ for images from the classes *Living Room* and *Dining Room*. Local explanations across multiple images are used to generate global explanations ($\mathbf{G}^r_{y_c}$) obtained via low-rank matrix approximation of the symbolic representation encoding: 1) All objects in an image (Baseline); 2) Objects in MSXs found via beam search.

and achieved a test accuracy of 98%. To negate the effect of objects during beam search, we utilize gaussian blur (kernel size=51, $\sigma = 50$) to blur out the objects not being considered by the beam search for k = 150 objects across the whole dataset. Figure 1 presents the image, its MSXs, and symbolic local and global explanations. The fewer objects in the global explanation obtained using our method indicate that the CNN learned to ignore common or redundant objects during training. Secondly, global explanations can be considered as sets of objects and using concepts from set theory, the extracted information could help us answer questions such as "What is common between a Living Room and Dining Room?". In our case, the intersection of objects in the global explanations, i.e. ($\mathbf{G}^r_{LivingRoom} \cap \mathbf{G}^r_{DiningRoom}$) results in "wall", "floor", "window". Conversely, one could ask "What sets a Living Room apart from a Dining Room?" which is given by the difference between the two sets, i.e. ($\mathbf{G}^r_{LivingRoom} - \mathbf{G}^r_{DiningRoom}$), i.e. "sofa" and "rug" and ($\mathbf{G}^r_{DiningRoom} - \mathbf{G}^r_{LivingRoom}$) = "table", "chair". In the same spirit, one could ask, "How can I make CNN classify a living room as a dining room?" (i.e., adversarial setting). The answer would be "Replace the sofa and rug with a table and chair."

Future Work

The proposed work introduced a novel methodology for explaining Deep CNNs globally while adhering to a standard taxonomy of objects. For future work, we are exploring alternative symbolic representations and methods to quantitatively evaluate global explanations complemented by a user study to understand the impact of enforcing a common taxonomy on human trust and understanding.

Acknowledgements

This research was partially funded through the NSF grant CNS-1941892 and the Industry-University Cooperative Research Center on Pervasive Personalized Intelligence. We thank Kiri L. Wagstaff, Rahul Khanna, Raffa Giuseppe, Kai Ishikawa, and Kunihiko Sadamasa for their valuable feedback.

References

Eckart, C.; and Young, G. 1936. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3): 211–218.

Kliegr, T.; Bahník, Š.; and Fürnkranz, J. 2021. A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence*, 295: 103458.

Shitole, V.; Li, F.; Kahng, M.; Tadepalli, P.; and Fern, A. 2021. One Explanation is Not Enough: Structured Attention Graphs for Image Classification. *Advances in Neural Information Processing Systems*, 34.

Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3): 302–321.