Interpreting Convolutional Neural Networks By Explaining Their Predictions

Toon Meynen, Hamed Behzadi-Khormouji, José Oramas
University of Antwerp, imec-IDLab

Abstract

- Exploit capabilities of well-studied model explanation and frequent-itemset mining methods.
- Identify relevant units by mining units that contribute to visual explanations.
- Pin-point class-specific and class-shared relevant units.

Background

Model Explanation: Justify the predictions made by a model

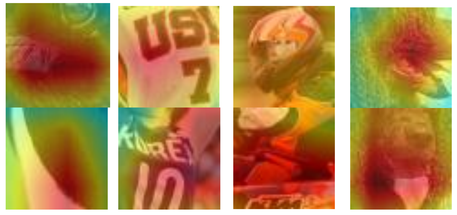
Model Interpretation: What has a model actually learned?

Proposed Method

1- Patch Extraction

- Visual Explanation are generated per input via Grad-CAM [1]

- Extracting highlighted patches from visual explanation.



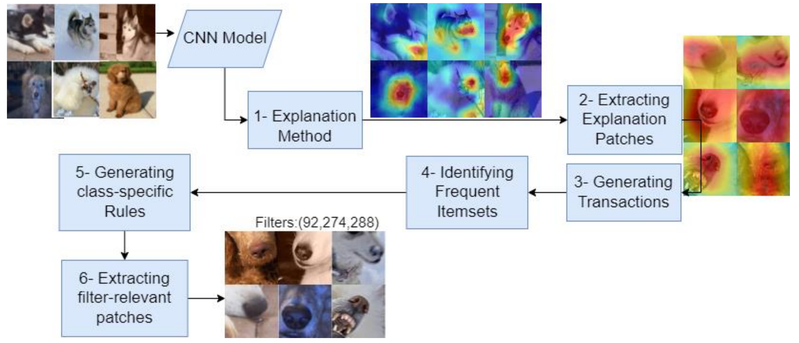


Fig 1. Proposed Deep Model Interpretation Method.

2- Transaction Database Generation

- Converting each visual patch into a *N* discreet values. The first *N-1* element shows indices of internal units, and the last element shows the predicted class label.

Unit 0	Unit 1	Unit 2	...	Unit N-1	Unit N
123	405	34	...	193	Bird
43	34	481	...	193	Bird
24	193	34	...	403	Dog
512	43	193	...	405	Car

3 – Frequent Itemset & Association Rule Mining

- Applying eclat algorithm [2]:
- The transactions of each class → Identifying relevant class-specific units.
- The whole transaction dataset → Identifying relevant class-shared units.

34,193 —> Bird, Dog
193,512 —> Car

Evaluation

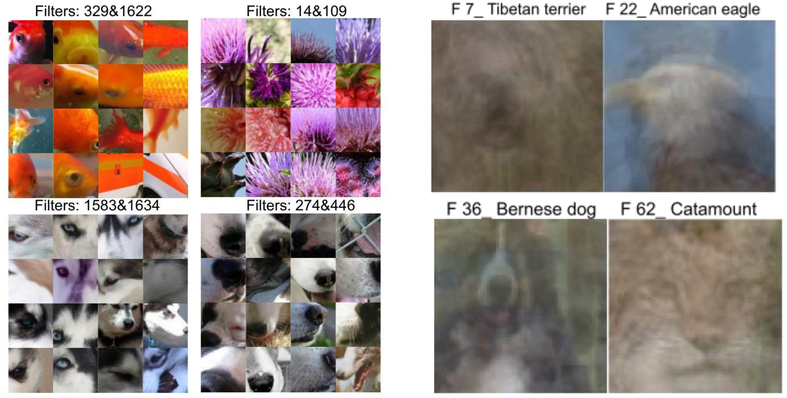
Quantitative Results

Table 1. Model accuracy comparison when internal units identified by each of the methods are perturbed.

Models/Methods	Ours-5 Acc. / avg. perturb.	Ours-10 Acc. / avg. perturb.	[6] Acc. / avg. perturb.	Random Acc. / avg. perturb.	Baseline Acc. / avg. perturb.
VGG16	55.17%±0.20 / 6.79	**45.67%±0.20 / 10.20**	50.53%±0.20 / 5.40	71.39%±0.17 / 10.00	71.59%±0.17 / 0.00
ResNet50	71.11% ±0.17 / 6.79	**68.63% ±0.18 / 10.20**	69.74% ±0.18 / 5.44	76.08% ±0.16 / 10.00	76.13%±0.16 / 0.00

Qualitative Results

Visual patterns from class-shared (L) & class-specific units (R)



Conclusion

We propose an interpretation method to identify class-specific and class-shared relevant units by mining explanations of its predictions.

Acknowledgments

This Project is partially supported by the UAntwerp BOF DOCPRO04 Project 41612(id) and FWO project G0A4720N.

References

[1] Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization", IJCV 2019.

[2] Zaki, Mohammed Javeed, et al. "New algorithms for fast discovery of association rules" *KDD* 1997.