Explaining Convolutional Neural Networks through Attribution-Based Input Sampling and Block-Wise Feature Aggregation

LG AI Research

Sam Sattarzadeh, Mahesh Sudhakar, Anthony Lem, Shervin Mehryar, K. N. Plataniotis,
Jongseong Jang, Hyunwoo Kim, Yeonjeong Jeong, Sangmin Lee, Kyunghoon Bae

University of Toronto, LG AI research

UNIVERSITY OF TORONTO

Introduction

- **Explainable AI (XAI):** Opening "black-box" AI-based models by providing human-understandable interpretations of their behavior.
- **Our aim: Visual Explainability**
  - Visualizing the behavior of models trained for image recognition tasks.
  - Generating a heatmap that represents the evidence leading the model to decide.
- **Our approach:** Proposing a visual explanation algorithm that is specialized to the family of Convolutional Neural Networks (CNNs).

Contributions

- **SISE (Semantic Input Sampling for Explanation):** A novel approach to provide interpretations for CNNs by aggregating the information extracted from multiple layers of the model.
- A strategy to select the minimum number of layers in each CNN to be visualized in order to provide a comprehensive view of the whole CNN.

Semantic Input Sampling for Explanation (SISE)

- Inspired by **RISE (Randomized Input Sampling for Explanation)**.
- A **CNN-specific** solution to address the limitations of RISE.
- **Perturbation-based:** Runs by feeding the model with masked copies of the input.

Major ideas:

- **Block-wise Feature Explanation:** Which layers of the CNN are required to be visualized?
- **Attribution-based Input Sampling:** How the input should be masked so that a RISE-based framework will be able to visualize each individual layer of the CNN?

Block-wise Feature Explanation

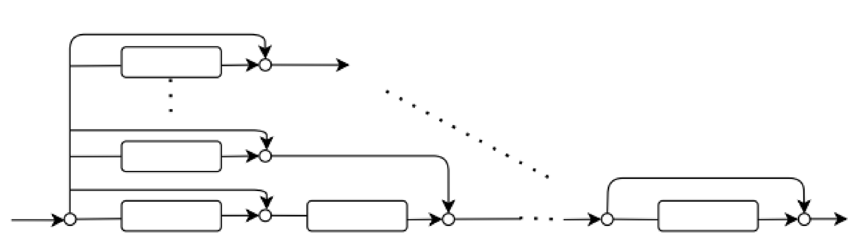

Figure: Unravelled architecture of residual CNNs. (Veit et al 2016.)

By decomposing the filters in non-residual CNNs, it can be shown that this architecture can be applied to these models as well.

Two implications from the unravelled architecture:

- During a forward/backward pass, the information may be processed by a convolutional layer or skip that layer.
- On the other hand, in pooling layers, all signals are downsampled. Thus, the implication above is NOT applied to the pooling layers of a CNN.

Conclusion: By visualizing the last convolutional layers in each convolutional block, representing the features captured through the CNN is achievable.

Attribution-based Input Sampling

Randomized Input Sampling for Explanation (Petsiuk et al. 2018):

- Creating a set of random masks M.
- Perturbing copies of the input (I) with the random masks ($I \odot M$).
- Passing the masked images to the model ($\Psi(.)$).
- Inferring the explanation map by combining the masks.

$$S_{RISE} = \mathbb{E}_M[m \times \Psi(I \odot m)] \tag{1}$$

The limitations of the RISE framework:

- Low visual quality of the explanation maps.
- Increase of failure chance while dealing with small object instances.
- Excessive computational overhead.

By replacing random masks with **attribution masks**, we infer the perspective of single layers of the target CNN.

Attribution masks:

- Getting the feature maps from a specific layer l, that are denoted as $A_k^{(l)}$.
- Selecting a class-distinctive set of features (using average gradient terms).

$$\alpha_k^{(l)} = \sum \frac{\partial \Psi(I)}{\partial A_k^{(l)}} \tag{2}$$

- Upscaling the features, using bilinear interpolation and normalization in the range [0,1]. This function is denoted as $\Omega(.)$

The set of attribution masks for each layer l are calculated as:

$$M_d^{(l)} = \{\Omega(A_k^{(l)}) | k \in \{1, ..., N\}, \alpha_k^{(l)} > \mu \times \max_k(\alpha_k^{(l)})\} \tag{3}$$

μ is a non-negative threshold parameter that is set to 0 by default.

Methodology

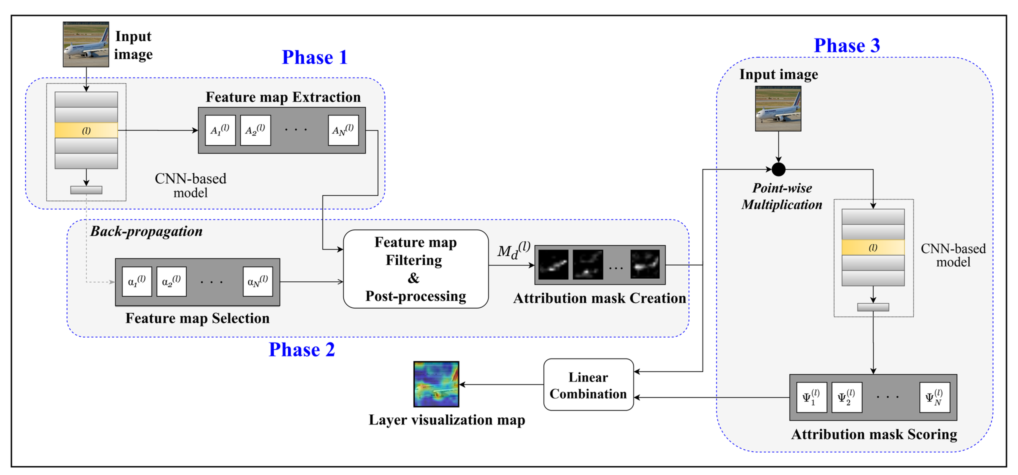

Figure: Layer visualization module (The first 3 steps).

SISE consists of 4 phases:

1. Feature Map Extraction
2. Feature Map Selection
3. Attribution Mask Scoring
4. Visualization Map Fusion

The first 3 steps are applied to the last layer in all convolutional blocks of the CNN.

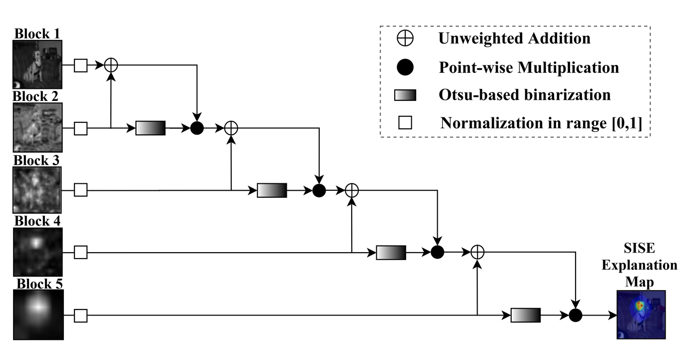

Figure: Fusion module (4th step)

The output of the third phase for each layer, is a visualization map that is computed as ($\lambda \in \Lambda$:: the set of locations in the input image domain):

$$V_{I,\Psi}^{l}(\lambda) = \mathbb{E}_{M_d^l}[\Psi(I \odot m) \times \frac{m(\lambda)}{\sum_{\lambda \in \Lambda} m(\lambda)}] \tag{4}$$

The visualization maps are fused into the explanation map by the fusion module.

Experimental Setup

Dataset: PASCAL VOC 2007:

- **Purpose:** Multi-label image classification, Object Detection.
- Containing 4963 test images in 20 classes, Bounding boxes provided.
- A VGG-16 model and a ResNet-50 model trained on this dataset are utilized.

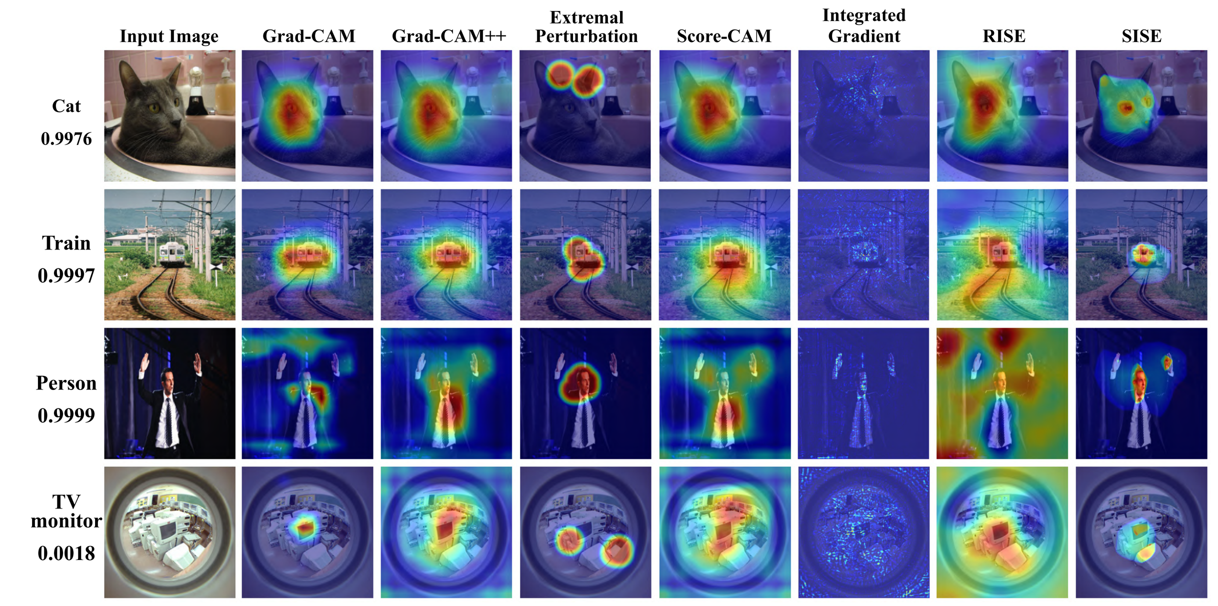

Figure: Qualitative evaluation of SISE on a VGG-16 trained on the PASCAL 2007 dataset.

Quantitative Evaluation

Evaluation metrics:

- *Ground truth-based* like Energy-based Pointing Game (**EBPG**), Mean Intersection-over-Union (**mIoU**) and Bounding Box (**Bbox**) are used to verify the meaningfulness of explanation methods, and their ability in feature visualization.
- *Model truth-based* like **Drop** and **Increase rate** are employed to justify the faithfulness and validity of the generated explanations from the model's perspective.

Model	Metric	Grad-CAM	Grad-CAM++	Extremal Perturbation	RISE	Score-CAM	Integrated Gradient	SISE
VGG16	EBPG	55.44	46.29	**61.19**	33.44	46.42	36.87	60.54
	mIoU	26.52	**28.1**	25.44	27.11	27.71	14.11	27.79
	Bbox	51.7	55.59	51.2	54.59	54.98	33.97	**55.68**
	Drop	49.47	60.63	43.90	39.62	39.79	64.74	**38.40**
	Increase	31.08	23.89	32.65	37.76	36.42	26.17	**37.96**
ResNet-50	EBPG	60.08	47.78	63.24	32.86	35.56	40.62	**66.08**
	mIoU	**32.16**	30.16	26.29	27.4	31.0	15.41	31.37
	Bbox	60.25	58.66	52.34	55.55	60.02	34.79	**61.59**
	Drop	35.80	41.77	39.38	39.77	35.36	66.12	**30.92**
	Increase	36.58	32.15	34.27	37.08	37.08	24.24	**40.22**

Table: Quantitative results on PASCAL VOC 2007 test set.

Conclusion

Multi-layer approach to CNN interpretation:

- Integrates both semantic and spatial information discovered by the CNN, in the explanation map.
- Represents features in multiple semantic levels, while discarding class-indistinctive attributions.

Attribution-based layer visualization:

- Highlights the class-distinctive features leading the model to make its prediction.
- Takes account for small-size instances extracted by the CNN.

References

Petsiuk, Vitali, Abir Das, and Kate Saenko. "RISE: Randomized Input Sampling for Explanation of Black-box Models." (2018).

Veit, Andreas, Michael J. Wilber, and Serge Belongie. "Residual networks behave like ensembles of relatively shallow networks." (2016)