Ada-SISE: Adaptive Semantic Input Sampling for Efficient Explanation of Convolutional Neural Networks

4216

LG AI Research

UNIVERSITY OF TORONTO

Mahesh Sudhakar, Sam Sattarzadeh, Konstantinos N. Plataniotis, Jongseong Jang, Yeonjeong Jeong, Hyunwoo Kim

University of Toronto, LG AI research

Introduction

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

Background

- **Methods for visual explainability**.
  - **Backpropagation-based methods:**
    Computing the gradient of CNN's output to the input features or hidden neurons.
  - **CAM-based methods:**
    Visualizing the features extracted in a single layer of the CNNs.
  - **Perturbation-based methods:**
    Probing the model's behavior using perturbed copies of the input image.
- Despite the outperforming performance of perturbation-based methods, they have room for improvement in terms of **speed** and **visual clarity**.

Our approach: Ada-SISE

- Produces visual explanations by aggregating the information extracted from multiple layers of the Convolutional Neural Network.
- Build upon the 'perturbation-based' method *SISE*.
- Eliminates the need for tuning hyper-parameters.
- Considerable decrease in computational overhead.
- Removes outliers and background features in the generated explanations.

SISE: Semantic Input Sampling for Explanation

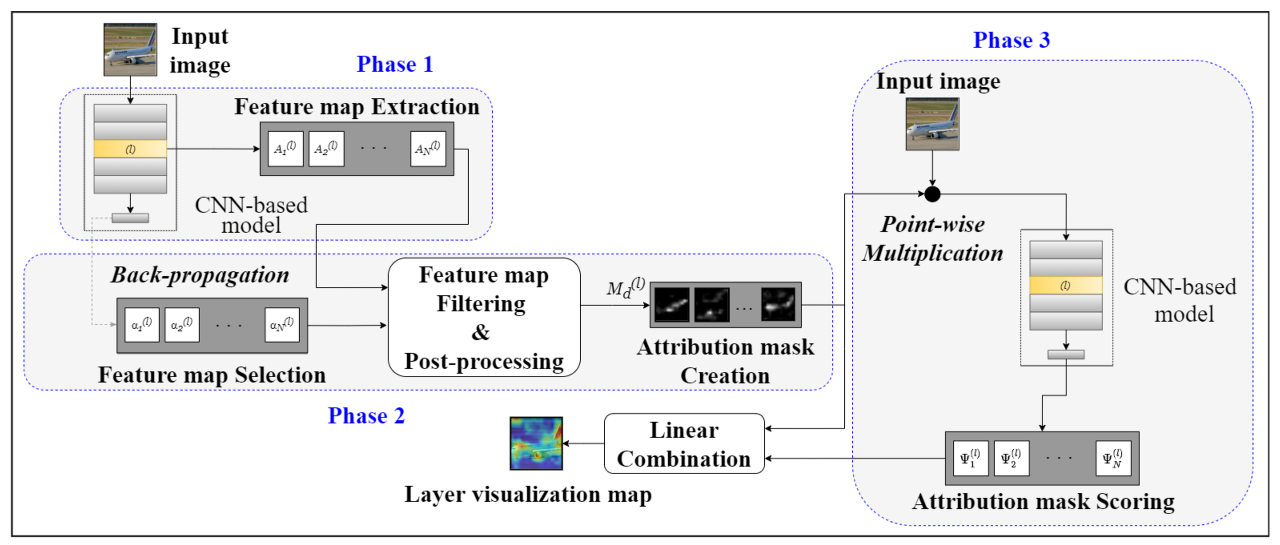


Figure: Credit: (Sattarzadeh et al. '20)

Novelty:

- A framework for visualizing layers of CNNs (See the figure above).
- A simple strategy for fusing information in various depths of CNNs.

Run-time Bottleneck: The 3rd phase, where the target model is fed with numerous *Attribution Masks*.

Limitation:

- Some attribution masks contain outliers and background information.
- Excessive number of attribution masks are utilized in the third phase.

Our Solution

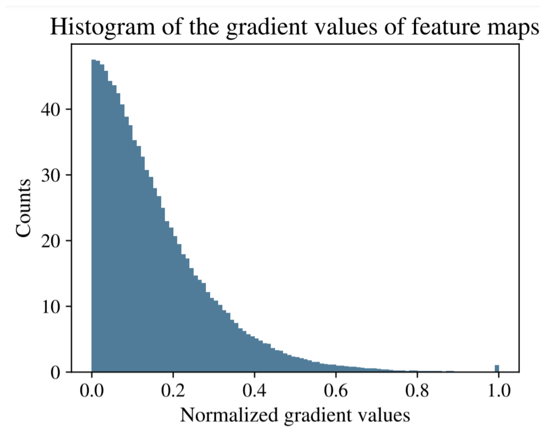


Figure: Histogram of the average-gradient values for the feature maps in the last convolutional layer of a ResNet-50 model.

A two-fold approach to filter the feature maps in the second phase. The feature maps are filtered based on their 'average-gradient' scores.

- Collecting the feature maps with positive average-gradient scores.
- Calculating an adaptive threshold to maximize the 'inter-class' variance between the selected and discarded feature maps.

Target CNN: $\Psi(.)$ and Input image: I,

Feature maps extracted from the layer l: $\{A_k^{(l)} | k \in \{1, ..., N\}\}$

Average-gradient score for $A_k^{(l)}$: $\alpha_k^{(l)} = \sum \frac{\partial \Psi(I)}{\partial A_k^{(l)}}$

The set of positive-gradient feature maps:

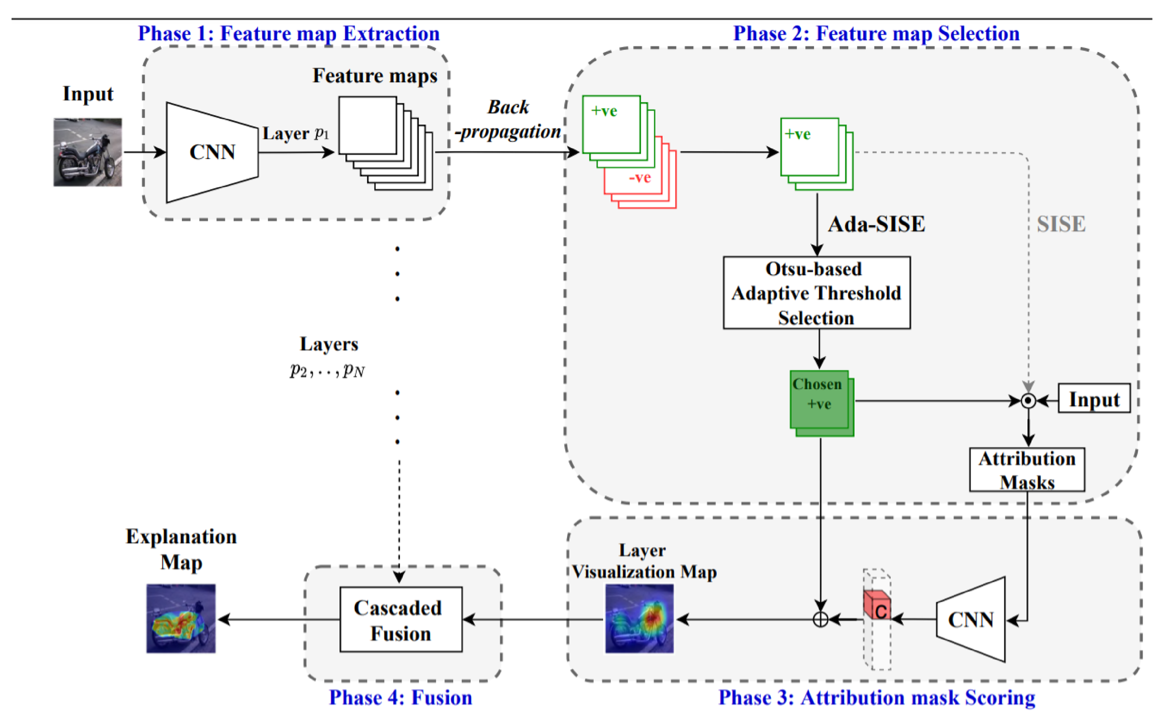
$$A^{+(l)} = \{A_k^{(l)} | k \in \{1, ..., N\}, \alpha_k^{(l)} > 0)\} \tag{1}$$

The set of normalized average-gradient values:

$$\Upsilon^{(l)} = \{\alpha_k^{(l)} | k \in \{1, ..., N\}, \alpha_k^{(l)} > 0)\} \tag{2}$$

Assumption: the values in the set $\Upsilon^{(l)}$ are sorted incrementally.

Methodology



The i-th minimum value in the set $\Upsilon^{(l)}$ :: $\Upsilon^{(l)}(i)$.

Lower/Upper mean function:

$$\omega_L^{(l)}(i) = \frac{\sum_{j=1}^{i} (\Upsilon^{(l)}(j))}{i} \times |\Upsilon^{(l)}|, \text{ and } \omega_H^{(l)}(i) = \frac{\sum_{j=i}^{|\Upsilon^{(l)}|} (\Upsilon^{(l)}(j))}{|\Upsilon^{(l)}| - i} \times |\Upsilon^{(l)}|$$

Inter-class variance:

$$\tau^{(l)}(i) = \omega_L^{(l)}(i) \times \omega_H^{(l)}(i) \times \left[\frac{|\Upsilon^{(l)}| - i}{|\Upsilon^{(l)}|} - \frac{i}{|\Upsilon^{(l)}|}\right]^2 = \omega_L^{(l)}(i) \times \omega_H^{(l)}(i) \times \left[\frac{|\Upsilon^{(l)}| - 2i}{|\Upsilon^{(l)}|}\right]^2$$

The adaptive threshold value is achieved by maximizing the inter-class variance:

$$\mu^{(l)} = \Upsilon^{(l)}\left(\underset{j \in \{1, ..., |\Upsilon^{(l)}|\}}{\text{argmax}} \left(\tau^{(l)}(j)\right)\right) \tag{3}$$

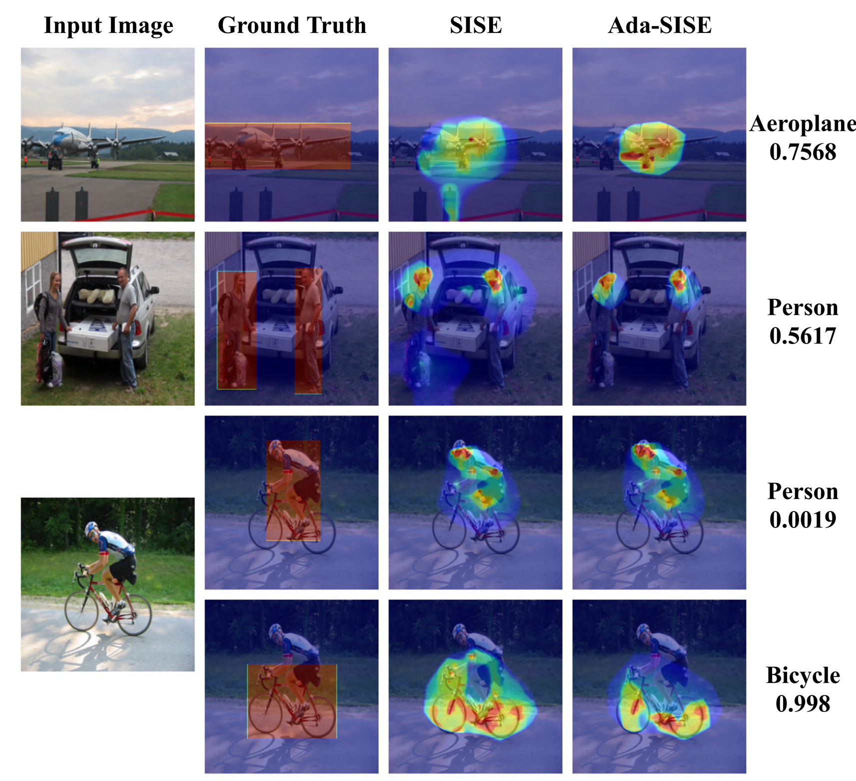
The set of feature maps utilized by Ada-SISE:

$$A_{Ada-SISE}^{(l)} = \{\alpha_k^{(l)} | k \in \{1, ..., N\}, \alpha_k^{(l)} > \mu^{(l)})\} \tag{4}$$

Experiments

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.



Quantitative Evaluation

Evaluation metrics:

- *Ground truth-based:* Energy-based Pointing Game (**EBPG**) and Bounding Box (**Bbox**) are used to verify the meaningfulness of explanation methods, and their ability in feature visualization.
- *Model truth-based:* **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	Ada-SISE
VGG16	EBPG	55.44	46.29	**61.19**	33.44	46.42	36.87	60.54	60.79
	Bbox	51.7	55.59	51.2	54.59	54.98	33.97	55.68	**55.73**
	Drop	49.47	60.63	43.90	39.62	39.79	64.74	**38.40**	38.87
	Increase	31.08	23.89	32.65	37.76	36.42	26.17	37.96	**38.25**
ResNet-50	EBPG	60.08	47.78	63.24	32.86	35.56	40.62	66.08	**66.4**
	Bbox	60.25	58.66	52.34	55.55	60.02	34.79	61.59	**61.77**
	Drop	35.80	41.77	39.38	39.77	35.36	66.12	**30.92**	**30.92**
	Increase	36.58	32.15	34.27	37.08	37.08	24.24	40.22	**40.75**

Table: Quantitative results on PASCAL VOC 2007 test set.

Conclusion

Ada-SISE Takeaways:

- Reducing the computational overhead in the bottleneck of SISE by approximately 33%.
- Decreasing attention on the outliers and regions ineffective in the CNN's prediction.

References

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

Sattarzadeh, Sam, Mahesh Sudhakar, Anthony Lem, Shervin Mehryar, K. N. Plataniotis, Jongseong Jang, Hyunwoo Kim, Yeonjeong Jeong, Sangmin Lee, and Kyunghoon Bae. "Explaining Convolutional Neural Networks through Attribution-Based Input Sampling and Block-Wise Feature Aggregation." (2020).

Fong, Ruth, Mandela Patrick, and Andrea Vedaldi. "Understanding deep networks via extremal perturbations and smooth masks" (2019).