ADA-SISE: ADAPTIVE SEMANTIC INPUT SAMPLING FOR EFFICIENT EXPLANATION OF CONVOLUTIONAL NEURAL NETWORKS

Mahesh Sudhakar[1], Sam Sattarzadeh[1], K. N. Plataniotis[1], Jongseong Jang[2] , Yeonjeong Jeong[2], Hyunwoo Kim[2]

1. The Edward S. Rogers Sr. Department of Electrical & Computer Engineering, University of Toronto
2. LG AI Research

The Edward S. Rogers Sr. Department of Electrical & Computer Engineering
UNIVERSITY OF TORONTO

FACULTY OF APPLIED SCIENCE & ENGINEERING

Overview of the presentation

- Explainable AI: Motivation, Applications
- Problem statement
  Semantic Input Sampling for Explanation (SISE)
- Our proposed method: Adaptive Semantic Input Sampling for Explanation (Ada-SISE)
- Empirical results
- Conclusion
- References

Motivation

Explainable AI (XAI):
provides human-satisfying interpretations of the behavior of "black-box" AI-based models, increasing users' trust on these cumbersome models[1].

Why did the model predict this?
When the model fails to predict correctly?
What features are important for the model?
...

Applications:

- **Medicine, Autonomous Driving:** remarkable demand for reasoning due to the catastrophic side effects of single false predictions.
- **Criminal Justice:** Regulations forcing computer-based models to provide rationale for their decisions.
- **Novelty detection:** detecting abnormally-shaped patterns in real-world industrial data-sets.

[1] Lipton, Z. C. 2018. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. Queue 16(3): 31–57. ISSN 1542-7730. doi:10.1145/3236386.3241340.

Background

The problem of visual explainability

- To visualize the behavior of models trained for image recognition tasks.
- Using a heatmap representing the evidence leading the model to decide.

Our problem: Visual explainable AI

- A branch of *post-hoc* and *local* XAI algorithms
- Specialized on *all* feed-forward CNNs (*model-specific*)

Terminology:

Post-hoc: models the behavior of the target model after training has concluded.

Local: Illustrates the relationship between the outcome of the target model with the input

Model-specific : Specialized for a certain type of AI-based models, using assumptions regarding their architecture and properties

Existing Works

Visual explanation algorithms:

- **Backpropagation-based methods:** Calculating the gradient of a model's output to the input features or the hidden neurons *(e.g., Vanilla Gradient, Integrated Gradient, Full Gradient)*.

- **CAM-based methods:** Visualizing the features extracted in a single layer of the CNNs *(e.g., Grad-CAM, Grad-CAM++, Score-CAM)*.

- **Perturbation-based methods**: Probing the model's behavior using perturbed copies of the input image *(e.g., RISE, Extremal Perturbation, SISE)*.

How Perturbation-based Methods Work

Randomized Input Sampling for Explanation[2] (RISE):

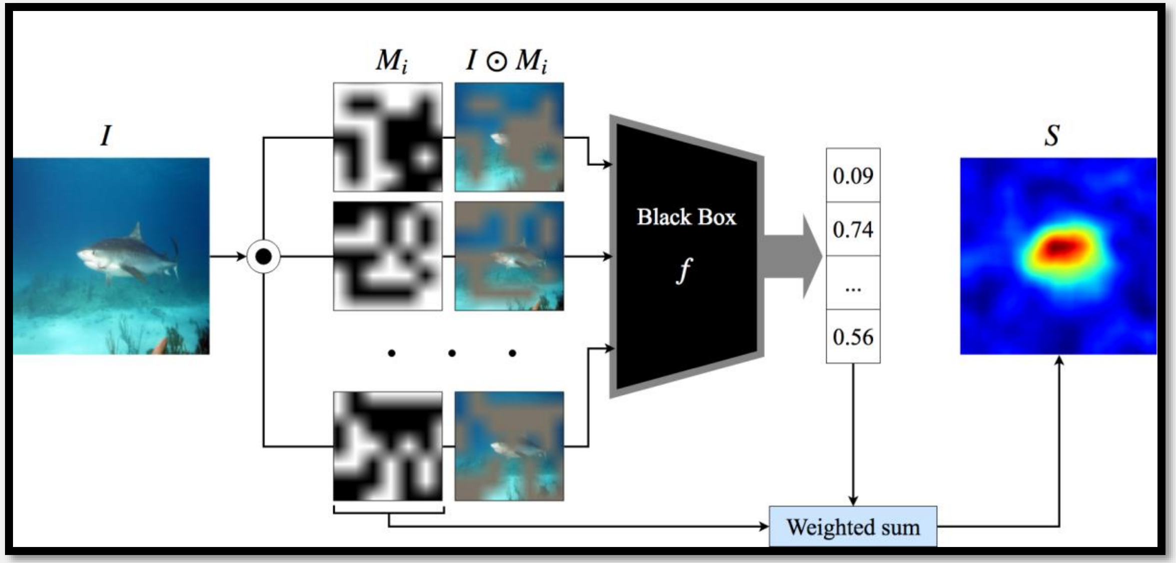


Image credit: [2]

Attribution masks: $m \in M$
The set of locations in the input domain: Λ
Explanation map: $S = \mathbb{E}_M[\Psi(I \odot m).M]$

Novelty:

- Investigating for the model's explanation by feeding the model with copies of the input image perturbed with *random masks.*

[2] Petsiuk, Vitali, Abir Das, and Kate Saenko. "Rise: Randomized input sampling for explanation of black-box models." arXiv preprint arXiv:1806.07421 (2018).

How Perturbation-based Methods Work

Semantic Input Sampling for Explanation[3] (SISE):

Research Gap Filled:

- Addressing the "Gradient Saturation" problem (Grad-CAM).
- Enhanced spatial resolution and clarity in the produced explanations (Grad-CAM, RISE).
- Improved consistency in the explanations (RISE).
- Considerable Decrease in the runtime (RISE, Score-CAM).

Novelty:

- Visualizing the perspective of individual layers via *attribution-based input sampling.*
- Replacing the *random masks* in RISE method with *attribution masks*.

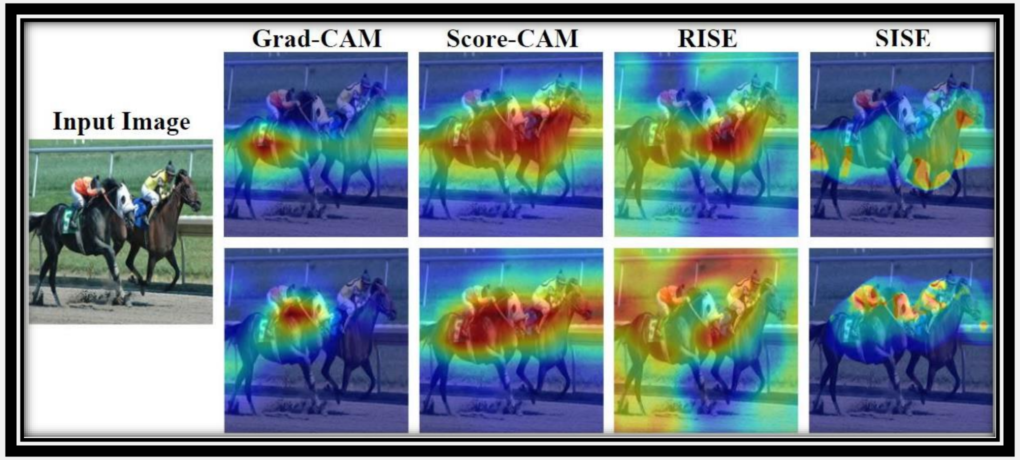
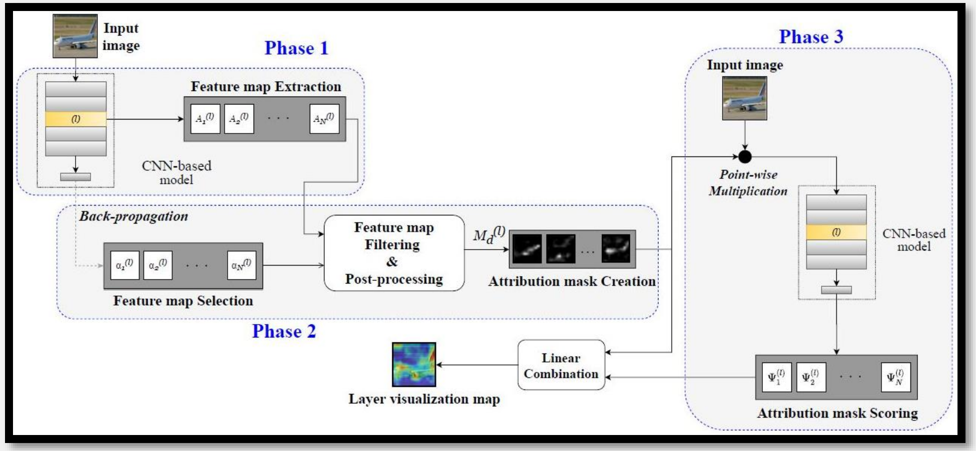


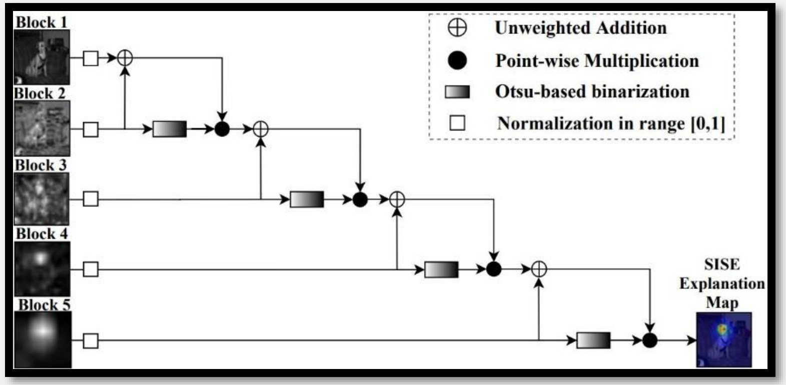
Image credit: [3]

[3] Sattarzadeh, Sam, et al. "Explaining Convolutional Neural Networks through Attribution-Based Input Sampling and Block-Wise Feature Aggregation." *arXiv e-prints* (2020): arXiv-2010.

How SISE Works

- Consists four consecutive phases:
  1. Feature map extraction
  2. Feature map selection
  3. Attribution mask scoring
  4. Feature aggregation





Phase 4:
Fusion block

- The first phases are applied on multiple layers. Corresponding to each layer, the third phase outputs a 2-dimensional map called ***visualization map***.
- The visualization maps are **aggregated** in the last phase to form the desires explanation map.

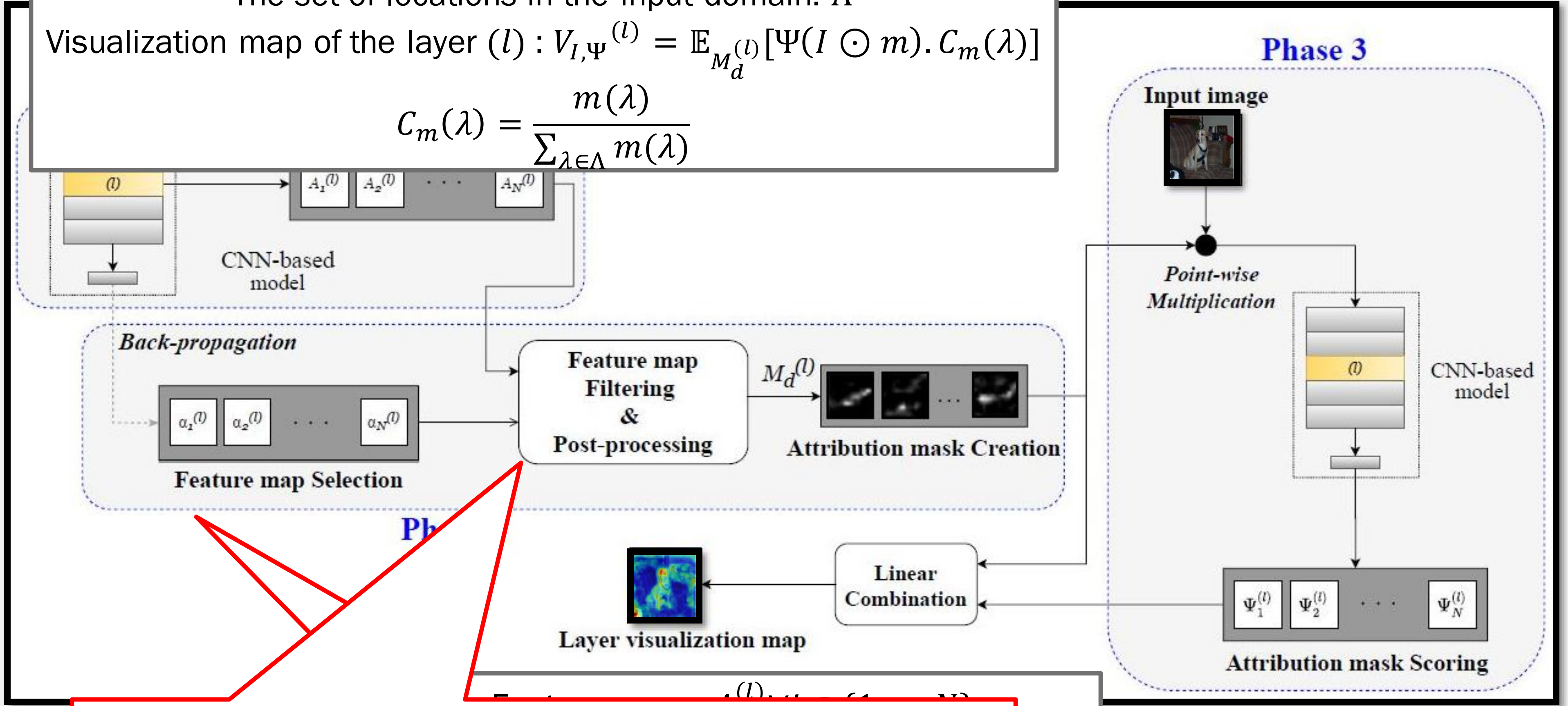
How SISE Works

Attribution masks: $m \in M_d^{(l)}$

The set of locations in the input domain: Λ

Visualization map of the layer (l) : ${V_{I,\Psi}}^{(l)} = \mathbb{E}_{M_d^{(l)}}[\Psi(I \odot m).C_m(\lambda)]$

$$C_m(\lambda) = \frac{m(\lambda)}{\sum_{\lambda \in \Lambda} m(\lambda)}$$



The feature maps satisfying $\frac{\alpha_k^{(l)}}{\max\limits_{k \in \{1,\dots,N\}} \alpha_k^{(l)}} > \mu$ are selected.

("μ" is a threshold parameter which is set to zero by default.)

Problem Statement

Let's take a close look into the second phase of the SISE method!

- **Are all attribution masks effective in the prediction procedure?**
- **Are "*all*" the feature maps with 'positive' average gradient scores (positive-gradient feature maps) free of outliers and background information?**
- **Is it yet possible to remove more unnecessary computational overhead from the SISE method?**

Feature maps: $A_i^{(l)} \forall i \in \{1, \ldots, N\}$
The set of locations in the feature maps: $\Lambda^{(l)}$
Average gradient scores: $\alpha_i^{(l)} = \sum_{\lambda^{(l)} \in \Lambda^{(l)}} \frac{\partial \Psi(I)}{\partial A_i^{(l)}(\lambda^{(l)})}$

The feature maps satisfying $\frac{\alpha_k^{(l)}}{\max_{k \in \{1,\ldots,N\}} \alpha_k^{(l)}} > \mu$ are selected.

("μ" is a threshold parameter which is set to zero by default.)

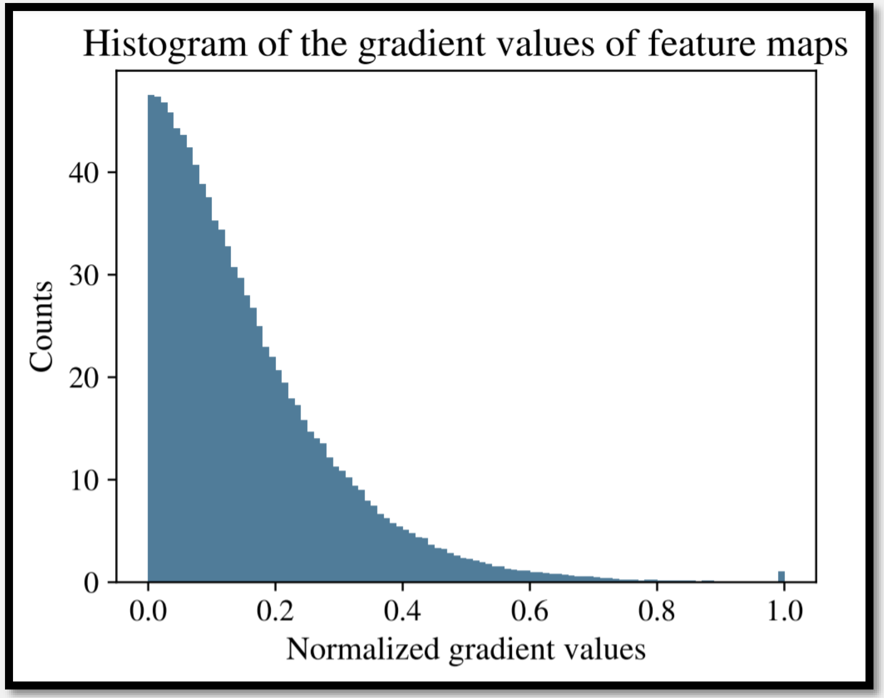
Problem Statement

Limitations of SISE

- The computational bottleneck of SISE is **on its third phase**, when a large set of attribution masks are passed through the target model.
- Most of the 'positive-gradient' attribution masks are **not effective** in the model's prediction procedure.
- The performance of SISE is dependent to the hyper-parameter "μ".

Goal

- Propose a strategy to tune the threshold parameter "μ" in a *positive value* in an *adaptive manner*.
- Reach an acceptable trade-off between the performance and runtime of the explanation method.



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

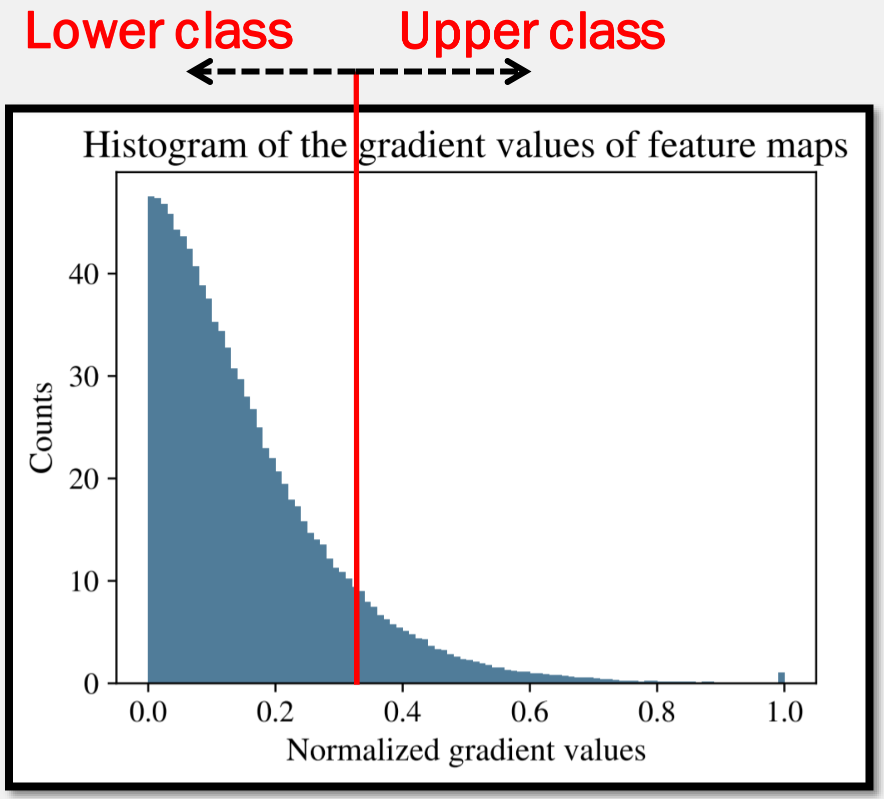
Problem Statement

Goal

- Propose a strategy to tune the threshold parameter "μ" in a *positive value* in an *adaptive manner*.
- Reach an acceptable trade-off between the performance and runtime of the explanation method.

Idea

- Maximizing the 'inter-class' variance between the feature maps in the 'lower class' and 'upper class'.
- Discarding the maximum number of ineffective feature maps, while retaining the explanation information.
- Ada-SISE only uses the positive-gradient feature maps in the upper class to infer the explanation.



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

Adaptive Semantic Input Sampling for Explanation (Ada-SISE)

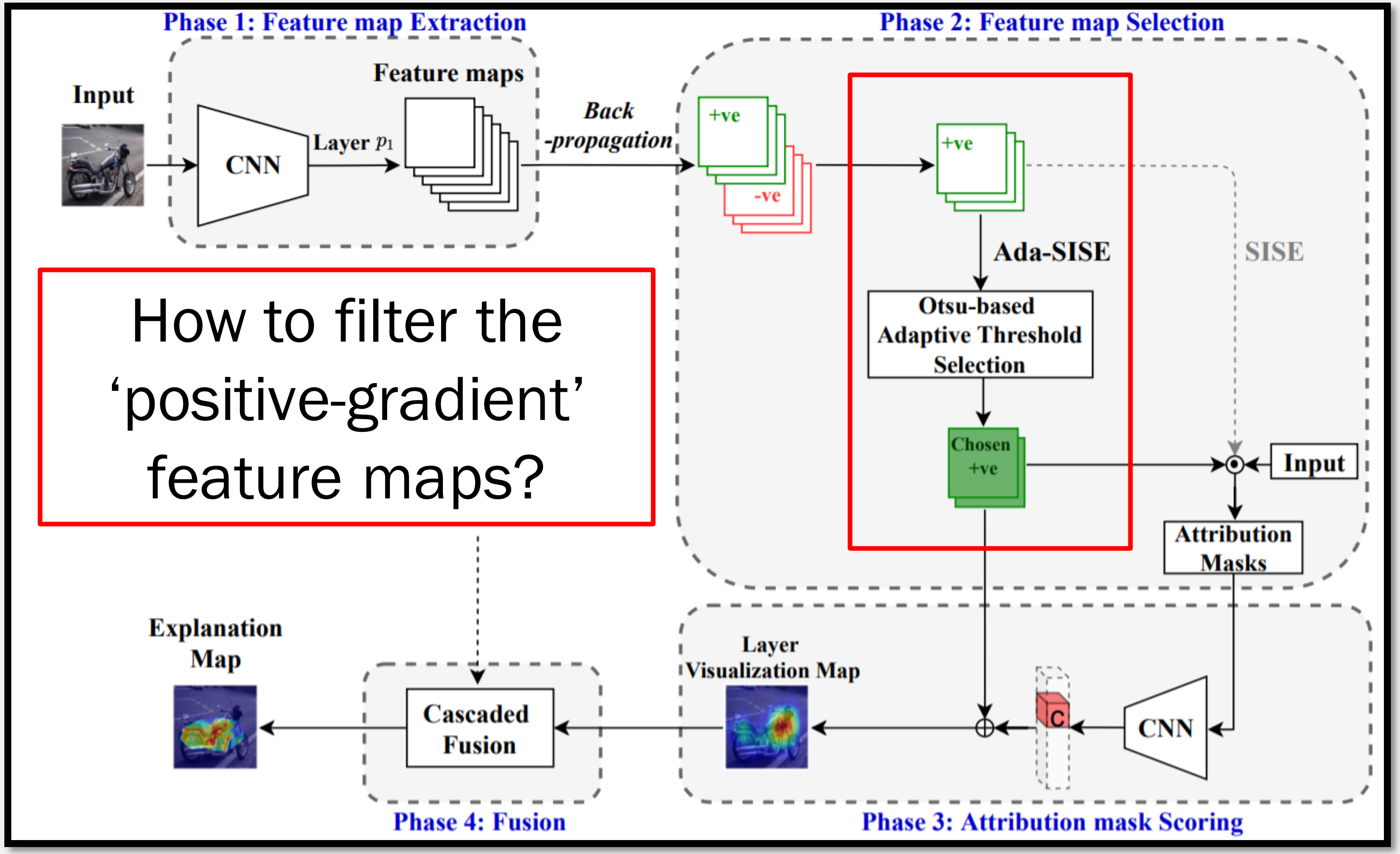
LG AI Research

The Edward S. Rogers Sr. Department of Electrical & Computer Engineering
UNIVERSITY OF TORONTO

FACULTY OF APPLIED SCIENCE & ENGINEERING

Our Approach

Ada-SISE vs. SISE:



Our Approach

Notations:

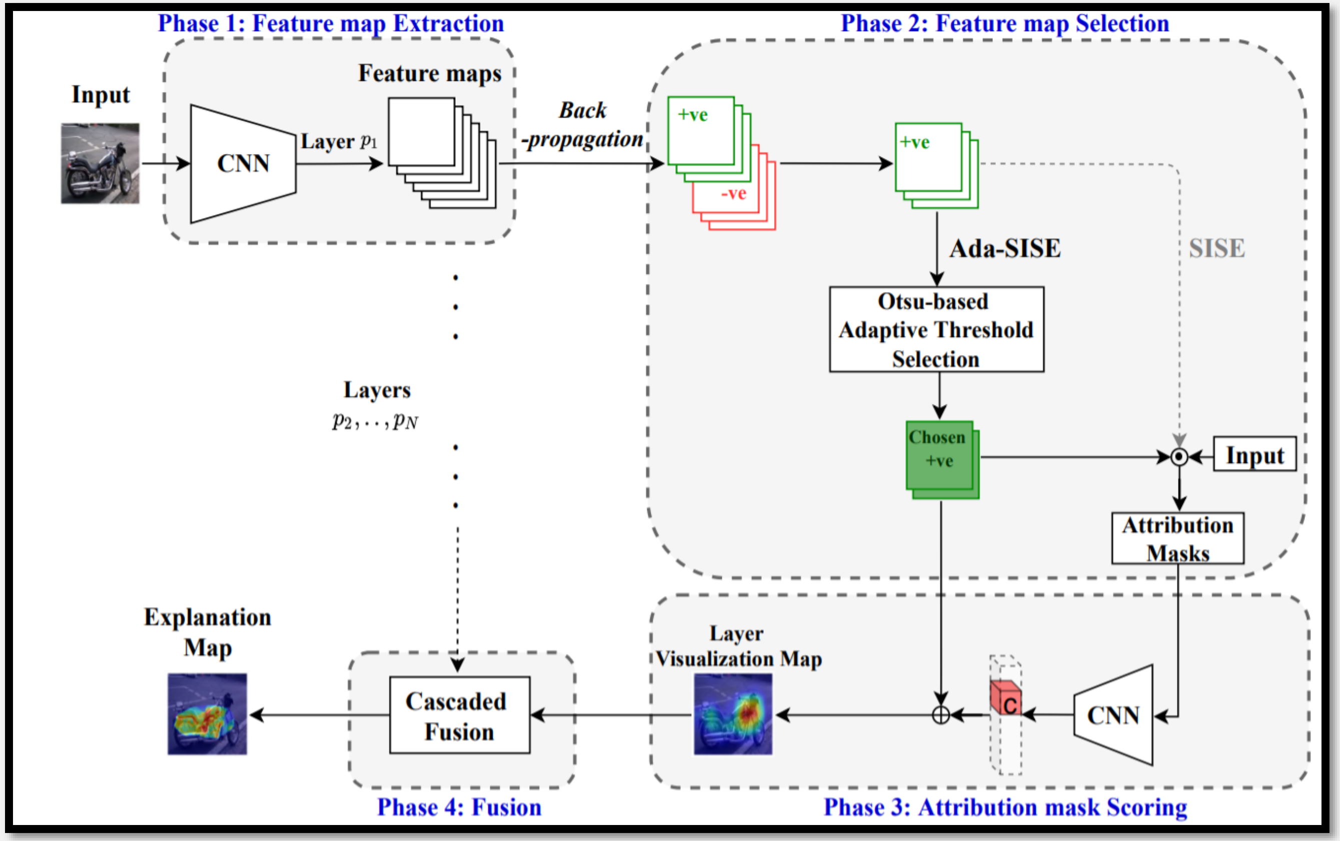
Assumption: the layer $[p]$ from the CNN model $\psi(.)$ contains M^p feature maps.

In the first phase, the input image I is passed through the model $\psi(.)$.

The set of feature maps extracted from the layer $[p]$: $F_k^{[p]}\ \forall\ k \in \{1, \ldots, M^p\}$

$$F_k^{[p]} = F_k^{[p]}: \Lambda^{[p]} \rightarrow \mathbb{R}$$

The set of locations in $F_k^{[p]}$ :: $\Lambda^{[p]}$



Our Approach

Average Gradient scores:

$$\sigma_k^{[p]} = \sum_{\lambda \in \Lambda^{[p]}} \frac{\partial \psi(I)}{\partial F_k^{[p]}(\lambda)}$$

$$\rho^{[p]} = \max_{k \in \{1,\ldots,M^p\}} \sigma_k^{[p]}$$

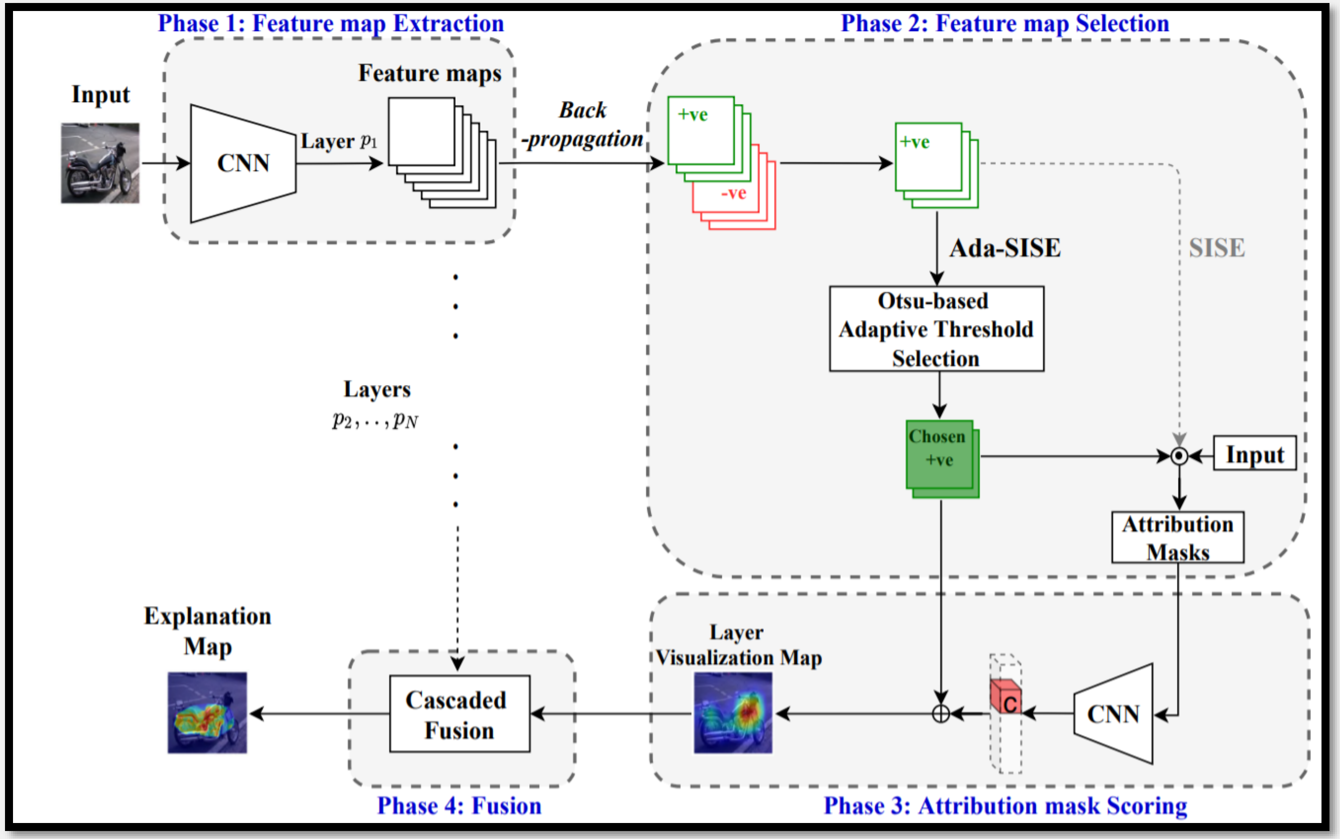
Normalized Average Gradient scores:

$$\upsilon_k^{[p]} = \frac{\sigma_k^{[p]}}{\rho^{[p]}}$$

The set of 'positive-gradient' feature maps:

$$F_k^{[p]+} = \{F_k^{[p]} | \upsilon_k^{[p]} > 0 , k \in \{1, \ldots, M^p\}\}$$

SISE



Our Approach

The set of ‘positive-gradient’ feature maps:

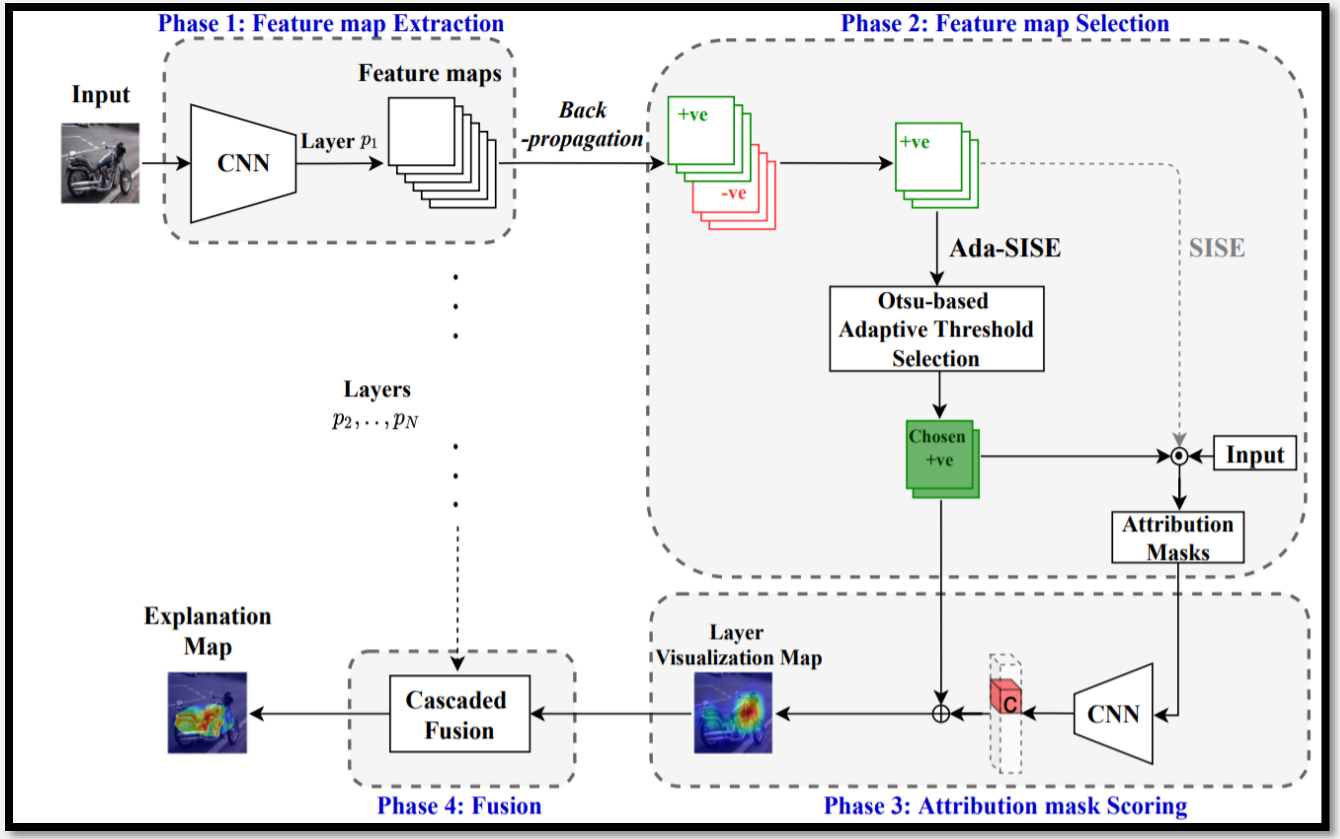
$$F_k^{[p]+} = \{F_k^{[p]} | v_k^{[p]} > 0 \,, k \in \{1, \ldots, M^p\}\}$$

The set of ‘positive-gradient’ values:

$$\Upsilon^{[p]} = \{v_k^{[p]} | v_k^{[p]} > 0 \,, k \in \{1, \ldots, M^p\}\}$$

Assumption: the set $\Upsilon^{[p]}$ is sorted increasingly.

The i-th minimum value in the set $\Upsilon^{[p]}$ :: $\Upsilon^{[p]}(i)$



Our Approach

The set of 'positive-gradient' feature maps:

$$F_k^{[p]+} = \{F_k^{[p]} | v_k^{[p]} > 0 \,, k \in \{1, \dots, M^p\}\}$$

The set of 'positive-gradient' values:

$$\Upsilon^{[p]} = \{v_k^{[p]} | v_k^{[p]} > 0 \,, k \in \{1, \dots, M^p\}\}$$

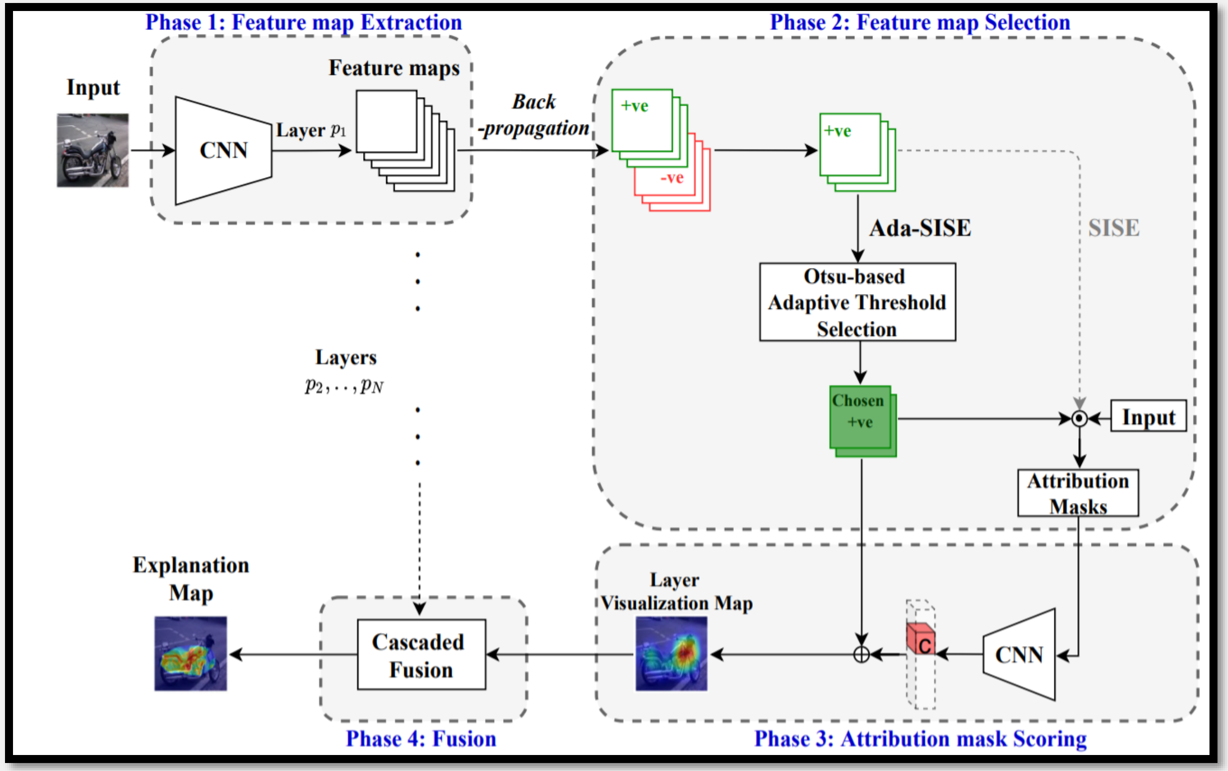
Assumption: the set $\Upsilon^{[p]}$ is sorted increasingly.

Lower mean function:

$$\omega_L^{[p]}(i) = \frac{\sum_{j=1}^{i}(\Upsilon^{[p]}(i))}{\left(i / |\Upsilon^{[p]}|\right)}$$

Upper mean function:

$$\omega_H^{[p]}(i) = \frac{\sum_{j=i}^{|\Upsilon^{[p]}|}(\Upsilon^{[p]}(i))}{\left(\left(|\Upsilon^{[p]}| - i\right) / |\Upsilon^{[p]}|\right)}$$



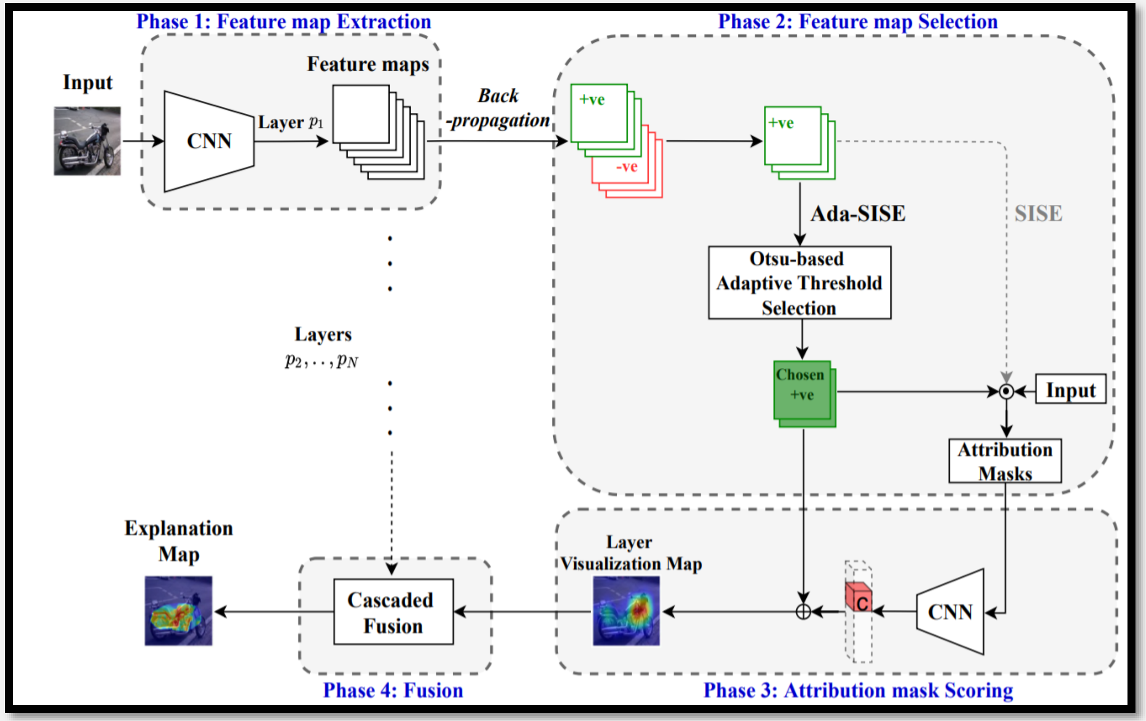
Our Approach

Lower mean function:	Upper mean function:
$\omega_L^{[p]}(i) = \dfrac{\sum_{j=1}^{i}(\Upsilon^{[p]}(i))}{(i/\lvert\Upsilon^{[p]}\rvert)}$	$\omega_H^{[p]}(i) = \dfrac{\sum_{j=i}^{\lvert\Upsilon^{[p]}\rvert}(\Upsilon^{[p]}(i))}{((\lvert\Upsilon^{[p]}\rvert - i)/\lvert\Upsilon^{[p]}\rvert)}$

Inter-class variance

$$\tau^{[p]}(i) = \omega_L^{[p]}(i) \times \omega_H^{[p]}(i) \times [\frac{\lvert\Upsilon^{[p]}\rvert - i}{\lvert\Upsilon^{[p]}\rvert} - \frac{i}{\lvert\Upsilon^{[p]}\rvert}]^2$$

$$\tau^{[p]}(i) = \omega_L^{[p]}(i) \times \omega_H^{[p]}(i) \times [\frac{\lvert\Upsilon^{[p]}\rvert - 2i}{\lvert\Upsilon^{[p]}\rvert}]^2$$



Maximizing the **inter-class variance** between the Average-Gradient values of the lower and upper class:

$$\mu^{[p]}(i) = \Upsilon^{[p]}(\operatorname*{argmax}_{j\in\{1,\ldots,\lvert\Upsilon^{[p]}\rvert\}} \tau^{[p]}(j))$$

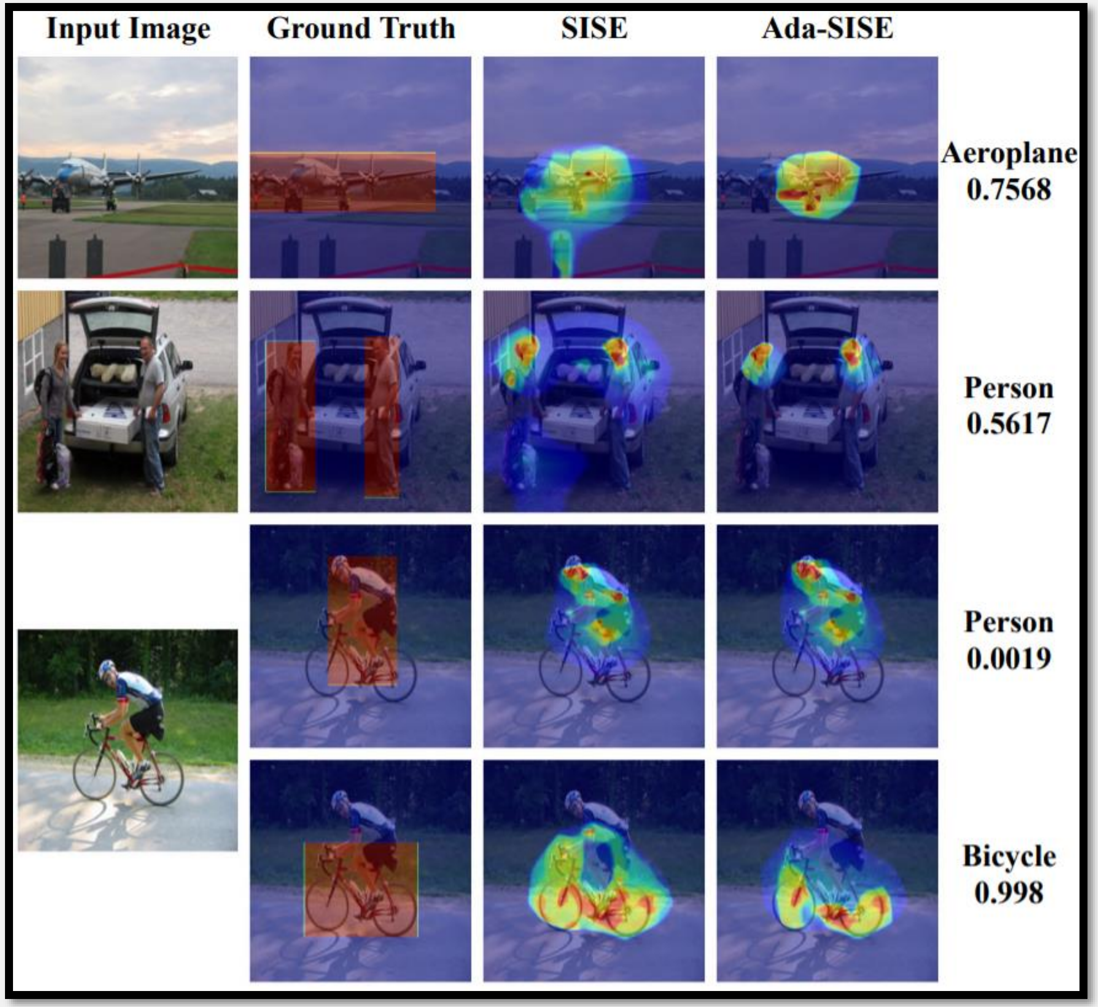
The number of selected feature maps:

$$\max_{j\in\{1,\ldots,\lvert\Upsilon^{[p]}\rvert\}} \tau^{[p]}(j)$$

Experiments: Datasets and Models

PASCAL VOC 2007[5]:

- 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[4].



[5] Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2007. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

Quantitative evaluation: metrics

Ground truth-based metrics

Verifying the meaningfulness of explanation methods, and their ability in feature visualization.

- Energy-based pointing game[8] (The fraction of energy inside am explanation map captured in a bounding box.)
- Bounding box[9] (Adaptive version of mean Intersection over Union (mIoU)).

Model truth-based metrics

Justifying the faithfulness and validity of the explanation maps from the perspective of the model.

- Drop rate[10] (Measuring the average drop in the model's confidence score (if drops), when only the top 15% of the pixels are retained).
- Increase rate[10] (Measuring the rate of increase in the model's confidence score, when only the top 15% of the pixels are retained).

[8] Wang, H.; Wang, Z.; Du, M.; Yang, F.; Zhang, Z.; Ding, S.; Mardziel, P.; and Hu, X. 2020. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 24–25.

[9] Schulz, K.; Sixt, L.; Tombari, F.; and Landgraf, T. 2020. Restricting the Flow: Information Bottlenecks for Attribution. In International Conference on Learning Representations. URL https://openreview.net/forum?id=S1xWh1rYwB.

[10] Chattopadhay, A.; Sarkar, A.; Howlader, P.; and Balasubramanian, V. N. 2018. Grad-CAM++: Generalized GradientBased Visual Explanations for Deep Convolutional Networks. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 839–847. doi:10.1109/WACV. 2018.00097.

[11] Ramaswamy, H. G.; et al. 2020. Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradientfree Localization. In The IEEE Winter Conference on Applications of Computer Vision, 983–991

Empirical Results

Dataset: PASCAL VOC 2007

	Metric	Grad-CAM	Grad-CAM++	Extremal Perturbation	Score-CAM	Integrated Gradients	FullGrad	RISE	SISE	Ada-SISE
VGG16	EBPG(%)	55.44	46.29	**61.19**	46.42	36.87	38.72	33.44	60.54	<u>60.79</u>
	Bbox(%)	51.7	55.59	51.2	54.98	33.97	54.17	54.59	<u>55.68</u>	**55.73**
	Drop(%)	49.47	60.63	43.90	39.79	64.74	60.78	39.62	**38.40**	<u>38.87</u>
	Increase(%)	31.08	23.89	32.65	36.42	26.17	22.73	37.76	<u>37.96</u>	**38.25**
ResNet-50	EBPG(%)	60.08	47.78	63.24	35.56	40.62	39.55	32.86	<u>66.08</u>	**66.4**
	Bbox(%)	60.25	58.66	52.34	60.02	34.79	44.94	55.55	<u>61.59</u>	**61.7**
	Drop(%)	35.80	41.77	39.38	35.36	66.12	65.99	39.77	**30.92**	**30.92**
	Increase(%)	36.58	32.15	34.27	37.08	24.24	25.36	37.08	<u>40.22</u>	**40.75**

For each metric, the best is shown in bold, and the second-best is underlined.

Complexity Analysis Results

Dataset: PASCAL VOC 2007

Model	RISE	SISE	Ada-SISE
VGG-16	64.28 s	5.96 s	**4.23 s**
ResNet-50	26.08 s	9.21 s	**6.29 s**

Average run-time on different models

Dataset: PASCAL VOC 2007

# of the conlvolutional block	p1	p2	p3	p4	p5	Total
RISE	N/A	N/A	N/A	N/A	N/A	8000
SISE	31	130	262	515	1008	1946
Ada-SISE	26	114	179	420	551	1290

Average number of required random/attribution masks for RISE/SISE/Ada-SISE to operate on a ResNet-50 Model

Ada-SISE reduces 33 percent of the computational load of SISE, without any performance degradation

Takeaways

<u>Ada-SISE</u>

1. Reducing 33% of the computational overhead in the bottleneck of SISE method.

2. Discarding the outlier information from the set of generated attribution masks.

3. The properties above are verified through qualitative and quantitative experiments on different model trained with the PASCAL VOC 2007 dataset.

4. Eliminating the need for hyperparameter-tuning; a noteworthy benefit in industry applications.

References

- Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).
- Wang, H.; Wang, Z.; Du, M.; Yang, F.; Zhang, Z.; Ding, S.; Mardziel, P.; and Hu, X. 2020. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 24–25.
- Fong, R.; Patrick, M.; and Vedaldi, A. 2019. Understanding deep networks via extremal perturbations and smooth masks. In Proceedings of the IEEE International Conference on Computer Vision, 2950–2958.
- Petsiuk, V.; Das, A.; and Saenko, K. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. In Proceedings of the British Machine Vision Conference (BMVC).
- Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, 3319–3328. JMLR. org.
- Chattopadhay, A.; Sarkar, A.; Howlader, P.; and Balasubramanian, V. N. 2018. Grad-CAM++: Generalized GradientBased Visual Explanations for Deep Convolutional Networks. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 839–847. doi:10.1109/WACV. 2018.00097.
- Srinivas, S.; and Fleuret, F. 2019. Full-gradient representation for neural network visualization. In Advances in Neural Information Processing Systems, 4126–4135.
- Lipton, Z. C. 2018. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. Queue 16(3): 31–57. ISSN 1542- 7730. doi:10.1145/3236386.3241340. URL https://doi.org/ 10.1145/3236386.3241340.
- Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2007. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.
- 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." arXiv preprint arXiv:2010.00672 (2020).

Thank you. Questions?

LG AI Research

The Edward S. Rogers Sr. Department of Electrical & Computer Engineering
UNIVERSITY OF TORONTO

FACULTY OF APPLIED SCIENCE & ENGINEERING