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

Sam Sattarzadeh[1], Mahesh Sudhakar[1], Anthony Lem[2], Shervin Mehryar[1], K. N. Plataniotis[1], Jongseong Jang[3], Hyunwoo Kim[3], Yeonjeong Jeong[3], Sangmin Lee[3], Kyunghoon Bae[3]

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

Overview of the presentation

- Explainable AI: Motivation, Applications
- Problem statement
- Our proposed method: Semantic Input Sampling for Explanation (SISE)
  1. Block-wise Feature aggregation
  2. Attribution-based perturbation
- 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.

Problem statement

Aim to address 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

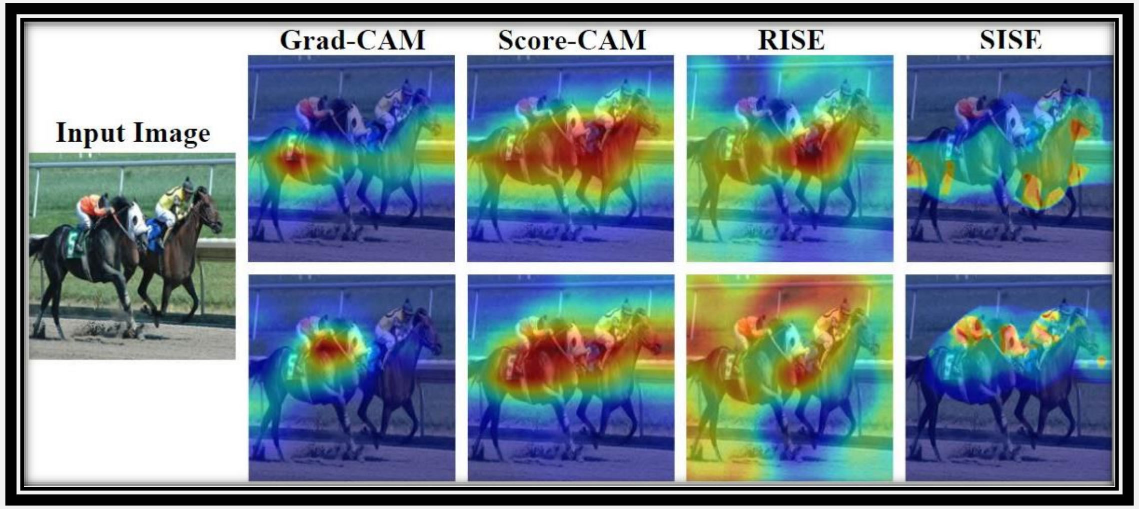
Model-specific : Specialized for a certain type of AI-based models, using assumptions regarding their architecture and properties

[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.

Limitations of Previous Works

Visual explanation algorithms:

- **Backpropagation-based methods:** Running by 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*).



Underestimation of global sensitivity

Low-resolution, noisy explanation map

Slow run-time of perturbation approaches

Goal of the proposal method

Goal:

- **Explanation completeness and faithfulness**: Correlation of the explanation maps with the model's behavior.
- **Visual quality**: The clarity of the generated explanations for the end-users (avoiding noise and blurring, high spatial resolution, and object localization ability).
- **+ Acceptable Run-time**

- We propose a novel attribution method which runs by visualizing the features detected in multiple layers of a CNN, and fusing this information in a unique explanation map.
- We discuss a simple strategy to select the minimum number of layers in each network to visualize in order to provide a concrete explanation for the whole CNN.
- By conducting thorough experiments on various models, we show that our proposed method offers more complete explanation maps and visualizes the features extracted by the target CNN more clearly, in comparison with the state-of-the-art attribution methods.

Our proposed method is perturbation-based.
However, it has common characteristics with the two other groups of the methods as well.

Semantic Input Sampling for Explanation (SISE)

Our Approach

Our proposed method (SISE):

- Inspired by **Randomized Input Sampling**[2] **(RISE)**
- **Model-specific solution** for CNNs to overcome the limitations of RISE
- **Idea:** Use feed-forwarding masked copies of a test image (called **attribution masks**) to the target model instead of **random masks**

Novelty

- The first to discuss and propose a logical layer selection strategy to get the most spatial and semantic information from a CNN by probing the minimum number of layers.
- The first to propose a fusion framework that aggregates the visualization maps from multiple layers in a factual manner, to improve the resolution of the explanation maps while retaining the class distinctiveness of the represented features

Related works on aggregating visualization maps from multiple layers

- (Rebuffi et al. 2020): Only combined multiple layer maps via simple operations such as addition or multiplication.
- (Wang et al.2019) : not address lack of class discriminability in the set of 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).

Developing SISE: Main ideas

Major questions:

1. What layers of a given CNN should we select to be visualized in the first 3 phases?

Idea. Block-wise Feature Explanation

2. How should we perturb the image to visualize each layer?.

Idea. Attribution-Based Perturbation

Block-Wise Feature Explanation

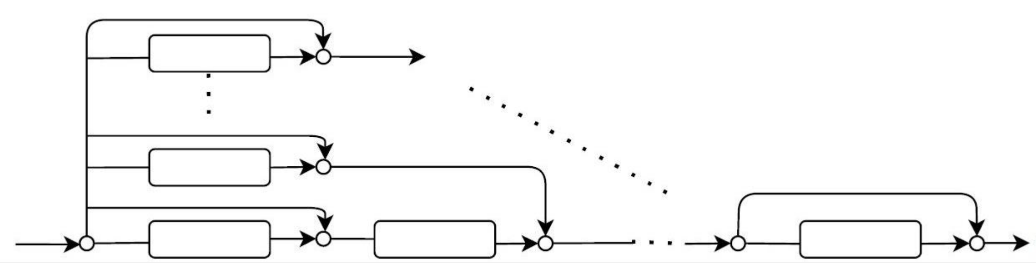
Pooling layers:

- Decreasing computational complexity in convolutional neural networks.
- Reducing dependency of the feature maps on local transmissions.
- Higher-level features can be interpreted as “presence of complex shapes, objects, and textures”.

Convolutional blocks:

- In shallow non-residual networks, they can be represented by a plain architecture.
- Residual networks are modelled with an unraveled architecture[3].

The unraveled architecture can be generalized to all CNNs, either residual or non-residual.

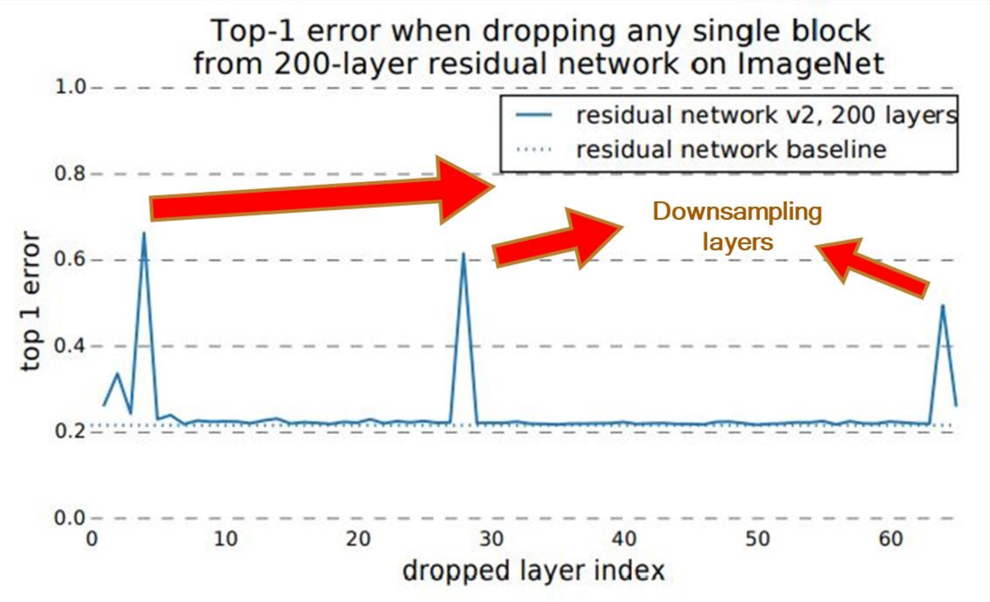


Unraveled architecture[3]

[3] Veit, A.; Wilber, M. J.; and Belongie, S. 2016. Residual networks behave like ensembles of relatively shallow networks. In Advances in neural information processing systems, 550– 558.

Block-Wise Feature Explanation

[3] shows both in theory and empirically that removing convolutional layers individually does not affect the network, but downsampling layers play a more important role.



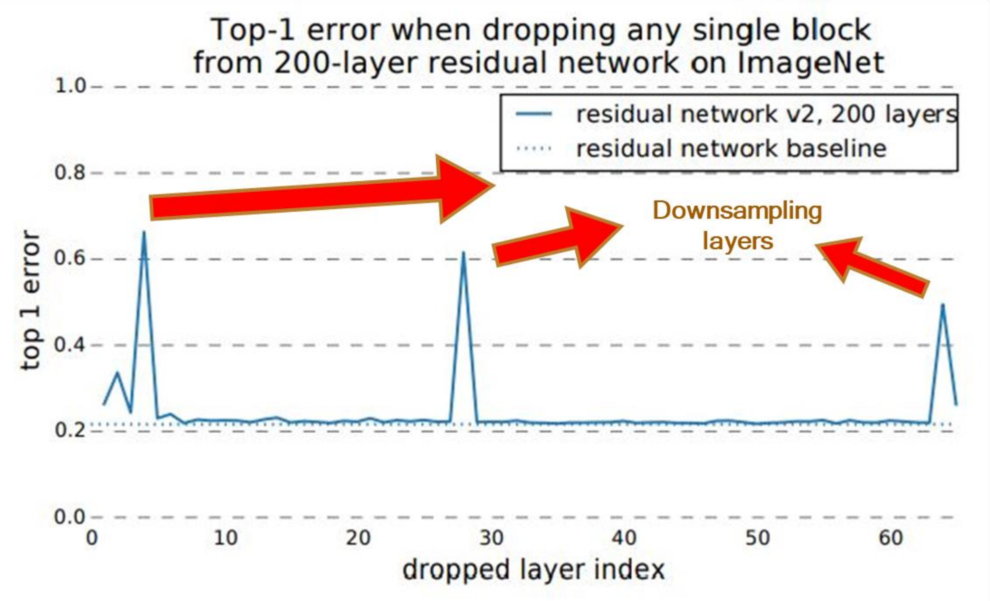
Insight 1: During a forward/backward pass, the information may be processed in a convolutional layer or propagated without any changes (e.g., from skip-connection layers).

Insight 2: However, regarding pooling operation, since the dimension of the layer's output is reduced, the implication above is not applied.

[3] Veit, A.; Wilber, M. J.; and Belongie, S. 2016. Residual networks behave like ensembles of relatively shallow networks. In Advances in neural information processing systems, 550– 558.

Block-Wise Feature Explanation

[3] shows both in theory and empirically that removing convolutional layers individually does not affect the network, but downsampling layers play a more important role.



Implication 1: All signals represented in each convolutional block can be traced from the input of their corresponding pooling (downsampling layer).

Implication 2: By visualizing the last convolutional layers in each convolutional block, representing the features captured by the CNN is achievable.

[3] Veit, A.; Wilber, M. J.; and Belongie, S. 2016. Residual networks behave like ensembles of relatively shallow networks. In Advances in neural information processing systems, 550– 558.

How Perturbation-based Methods Work

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

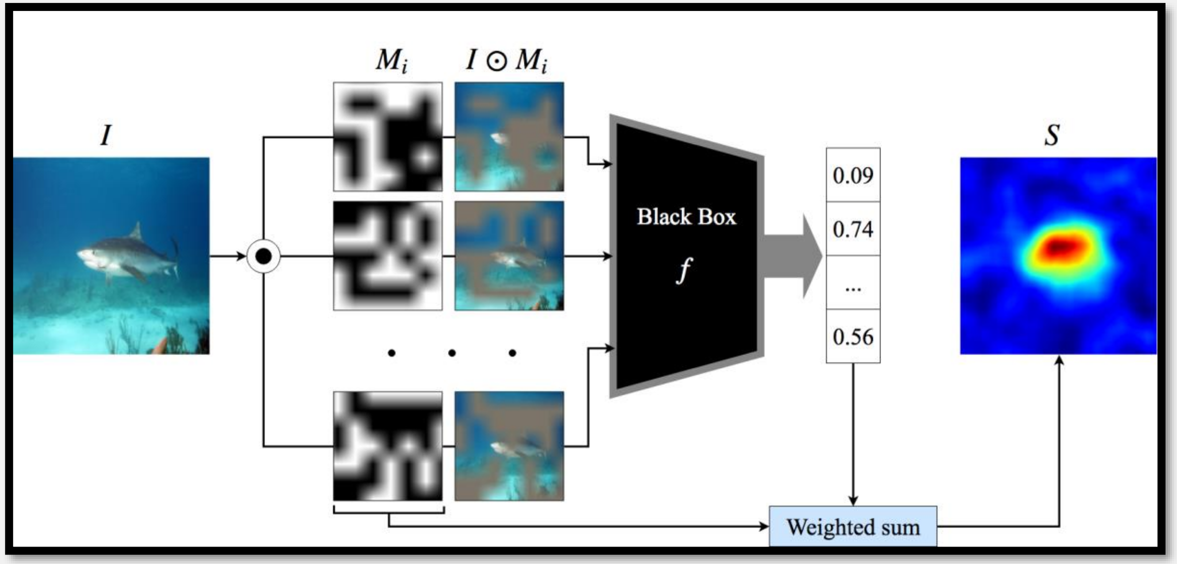


Image credit: [2]

Pros:

- Applicability of the method to the AI models beyond the family of CNNs.
- Shows the superior preciseness of perturbation rather than backpropagation, in forming explanation map.

Cons:

- Low visual quality of RISE explanation maps.
- Increase of failure chance, while dealing with small object instances
- Slow runtime, as it passes numerous (4000-8000) masked images through a model.

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

Attribution-based Perturbation

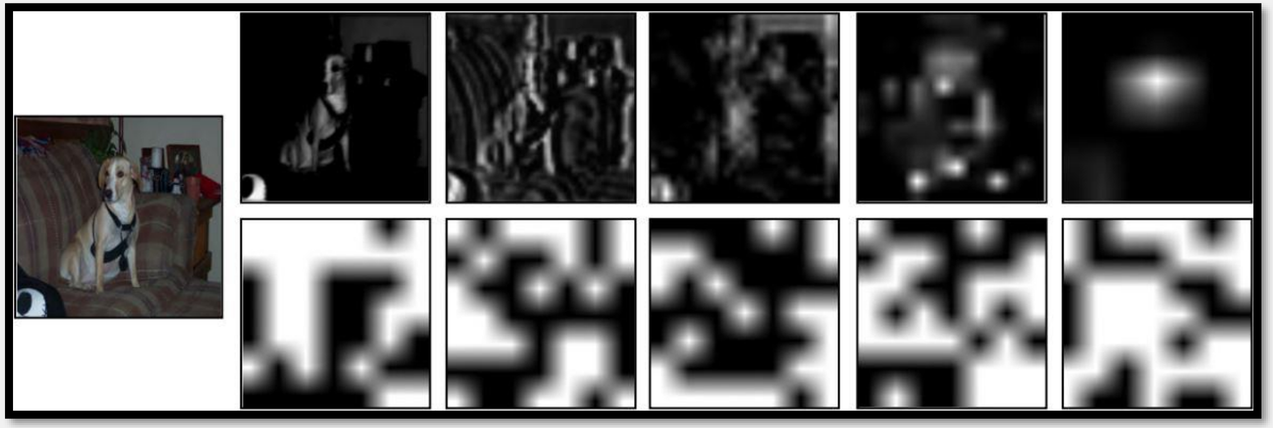
Semantic Input Sampling for Explanation (SISE):

Idea:

- We get attribution masks from the feature maps in the feature extractor part of the CNN.
- Since feature maps might contain class-indiscriminative attributions, we use backpropagation to select the most class-discriminative feature maps to be converted to attribution masks.

Pros:

- Depicting the attributions captured by the model.
- Ignoring the background and outliers ignored by the model.
- Reducing computational time (lower number of masks required than random masks).



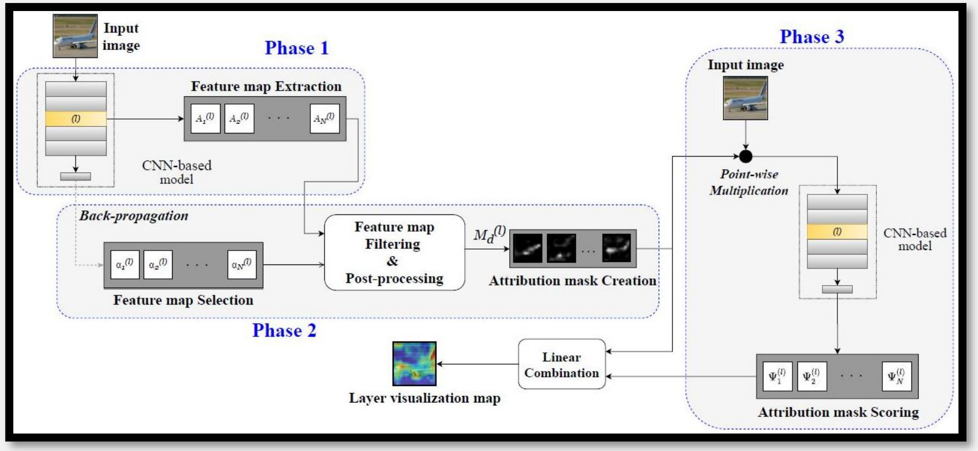
Top: Attribution masks – Bottom: Random masks[2].

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

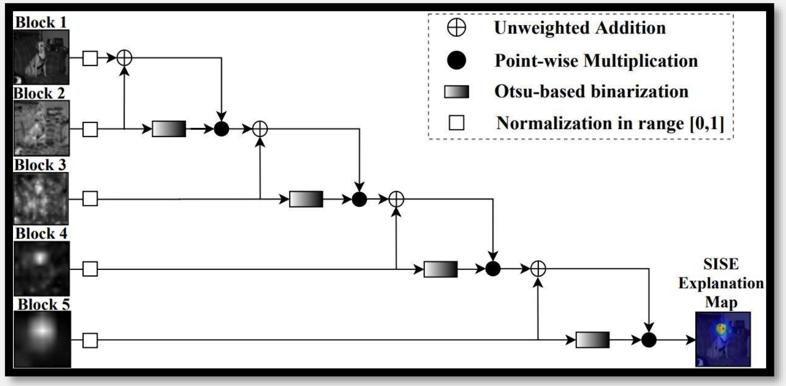
Our Approach

Our proposed method (SISE):

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



Input image
Phase 1
Feature map Extraction
$A_1^{(l)}$ $A_2^{(l)}$ · · · $A_N^{(l)}$
CNN-based model
Back-propagation
$\alpha_1^{(l)}$ $\alpha_2^{(l)}$ · · · $\alpha_N^{(l)}$
Feature map Selection
Phase 2
Feature map Filtering & Post-processing
$M_d^{(l)}$
Attribution mask Creation
Phase 3
Input image
Point-wise Multiplication
CNN-based model
$\Psi_1^{(l)}$ $\Psi_2^{(l)}$ · · · $\Psi_N^{(l)}$
Attribution mask Scoring
Linear Combination
Layer visualization map



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.

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

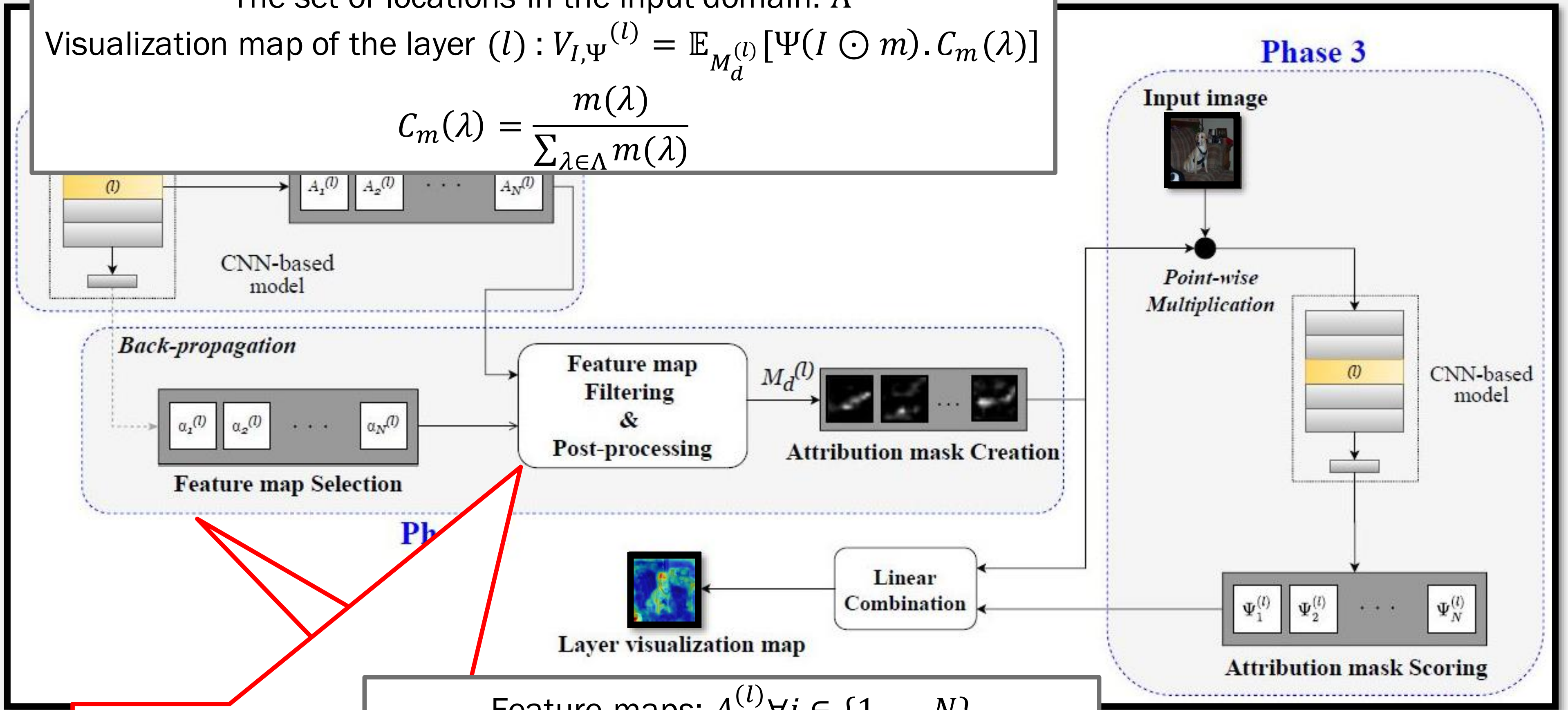
Methodology

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

The set of locations in the input domain: Λ

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

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


Phase 3
Input image
Point-wise Multiplication
CNN-based model
(l)
Attribution mask Scoring
$\Psi_1^{(l)}$ $\Psi_2^{(l)}$ . . . $\Psi_N^{(l)}$
Linear Combination
Layer visualization map
(l)
$A_1^{(l)}$ $A_2^{(l)}$ . . . $A_N^{(l)}$
CNN-based model
Back-propagation
$\alpha_1^{(l)}$ $\alpha_2^{(l)}$ . . . $\alpha_N^{(l)}$
Feature map Selection
Feature map Filtering & Post-processing
$M_d^{(l)}$
Attribution mask Creation
Ph

The feature map

(“μ” is a threshold

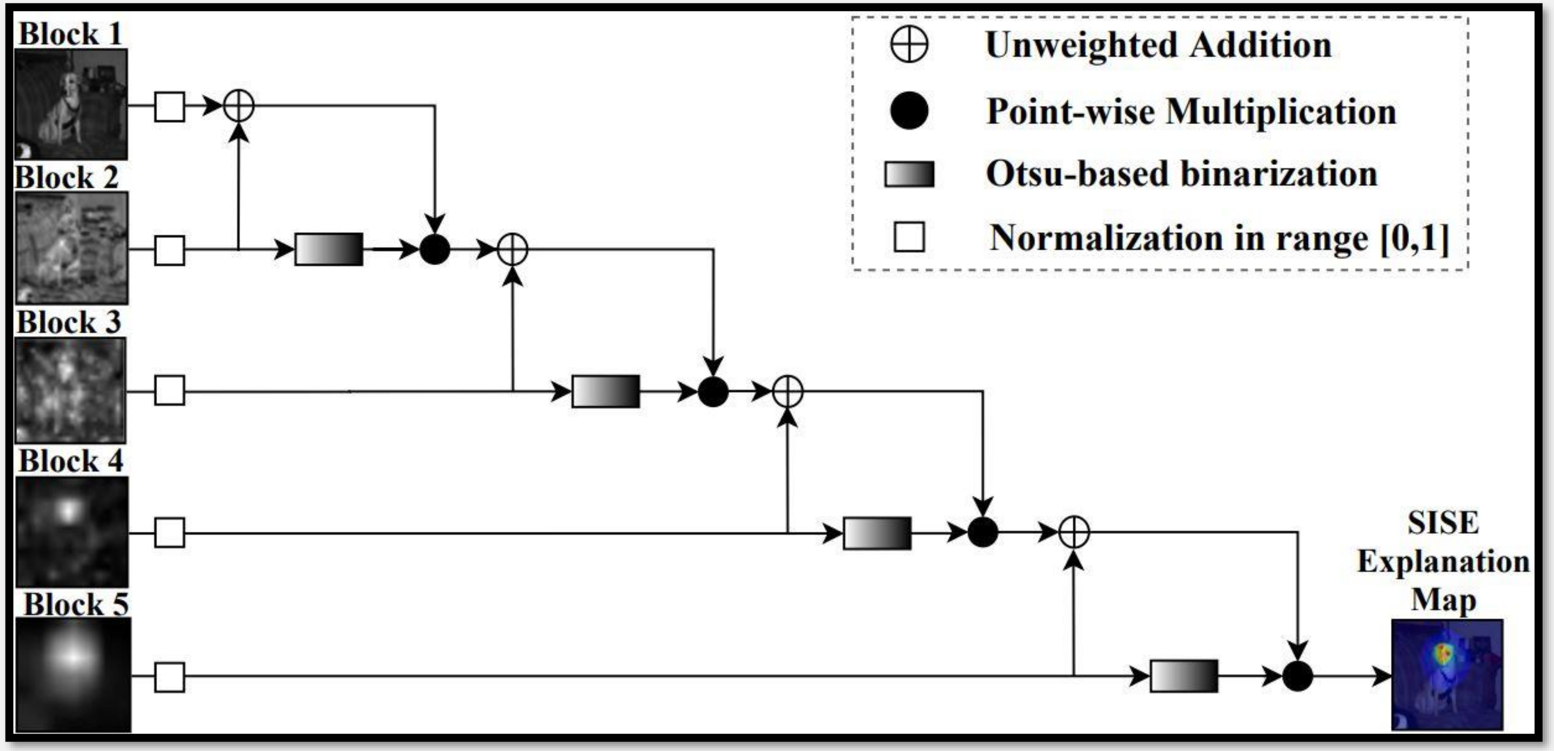
Feature maps: $A_i^{(l)} \forall i \in \{1, \dots, N\}$

The set of locations in the feature maps: $\Lambda^{(l)}$

Average gradient scores: $\alpha_i^{(l)} = \sum_{\lambda^{(l)} \in \Lambda^{(l)}} \frac{\partial \Psi(I)}{\partial A_i^{(l)}(\lambda^{(l)})}$

Methodology

Phase 4:
Fusion block



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].

MS COCO 2014[6]:

- Purpose: Multi-label image classification, Object Detection
- Containing over 40,000 validation images in 80 classes, Segmentation masks provided.
- A VGG-16 model and a ResNet-50 model trained on this dataset are utilized[4].

PAO Severstal[7]:

- Purpose: Anomaly Segmentation (we recast it to an image classification dataset)
- Containing test images from 4 defective and one normal classes
- Only correct labels provided.
- A balanced subset of images used (containing 4381 images).
- A ResNet-101 model trained on this dataset is utilized.

[4] 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.

[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.

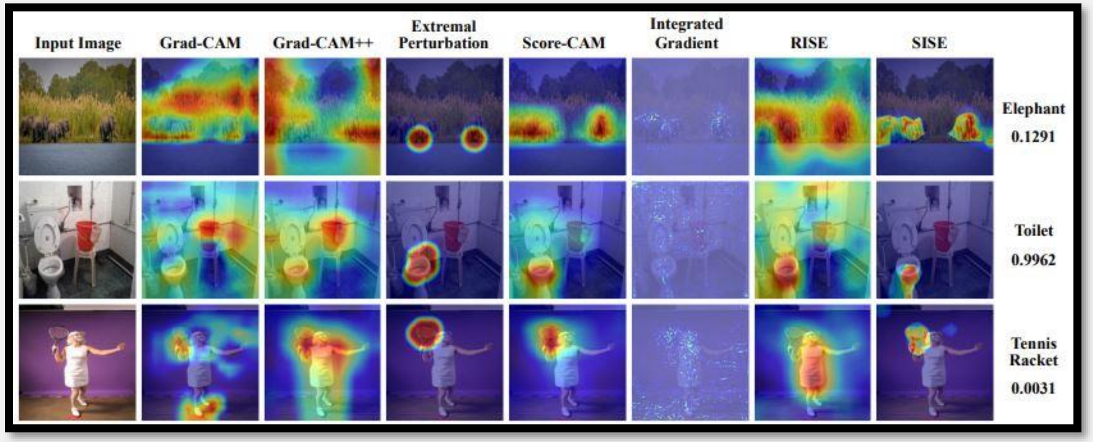
[6] Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; and Zitnick, C. L. 2014. Microsoft ´ coco: Common objects in context. In European conference on computer vision, 740–755. Springer.

[7] PAO Severstal. 2019. Severstal: Steel Defect Detection on Kaggle Challenge.

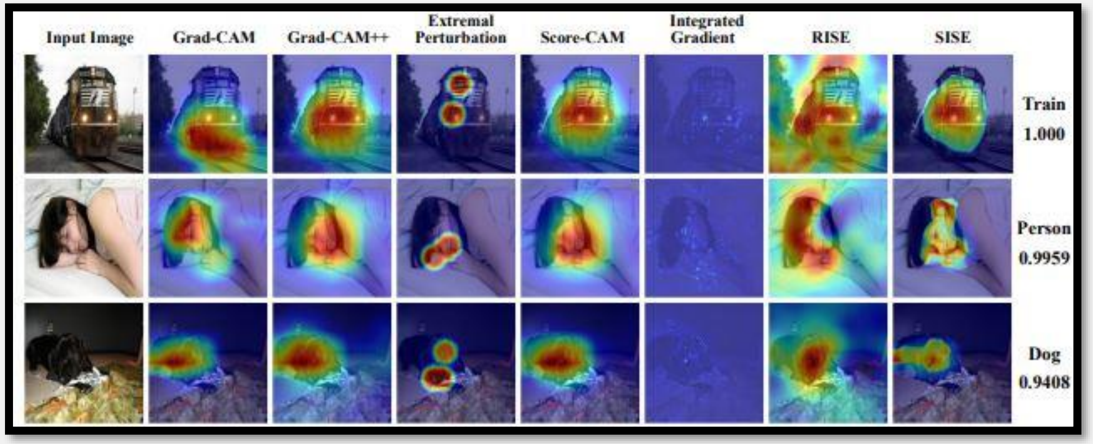
Qualitative results

<u>Advantages</u>

1. Improved spatial resolution of the explanation maps
2. Highlighting the mid-level and low-level features extracted by the target CNN.
3. More accurate explanations for the smaller instances.
4. Ignoring class-indistinctive features in the explanations.
5. The ability to provide concrete outputs while dealing with multiple instances from different classes (see the next slide).

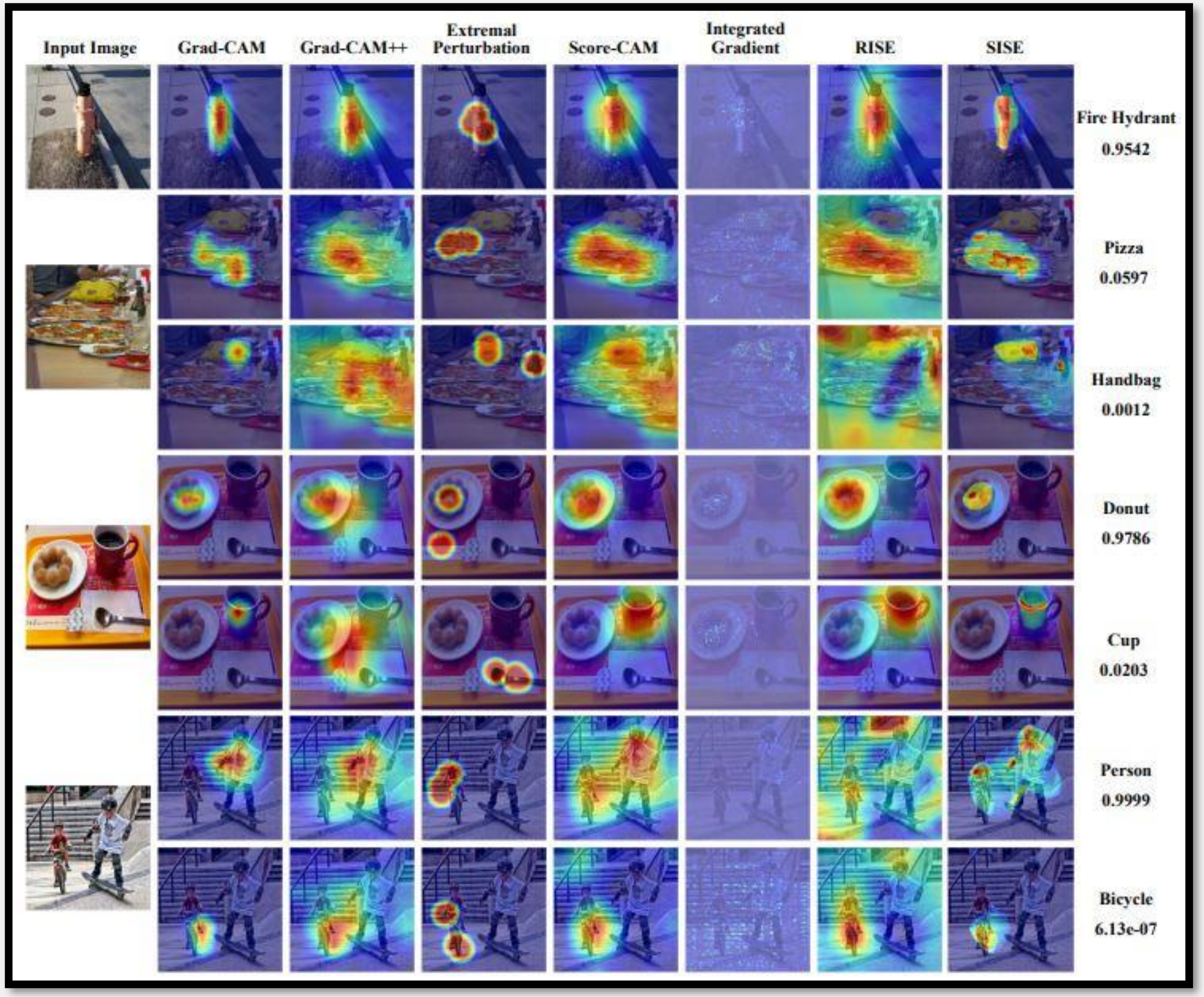


Dataset: PASCAL VOC 2007, Model: ResNet-50



Dataset: MS COCO 2014, Model: VGG-16

Qualitative results



Dataset: MS COCO 2014, Model: ResNet-50

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 mIoU).
- mIoU (comparing the top 20% pixels of explanation maps with ground truth.)

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).

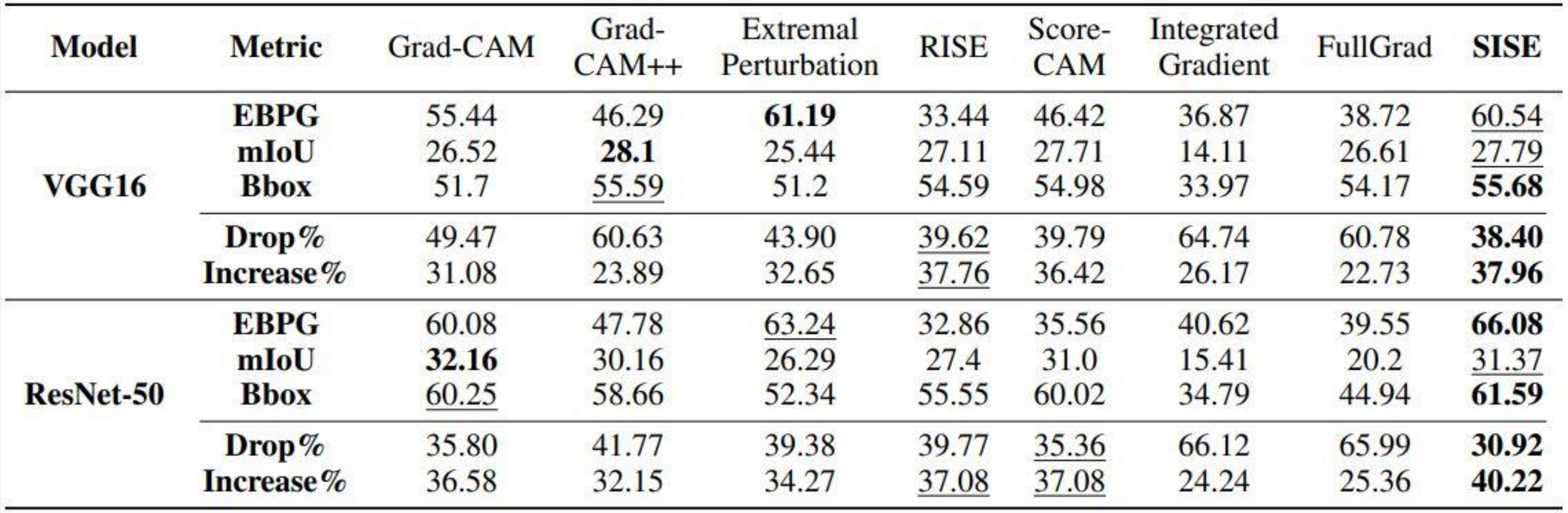
[8] Wang, H.; Wang, Z.; Du, M.; Yang, F.; Zhang, Z.; Ding, S.; Mardziel, P.; and Hu, X. 2020. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 24–25.

[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.

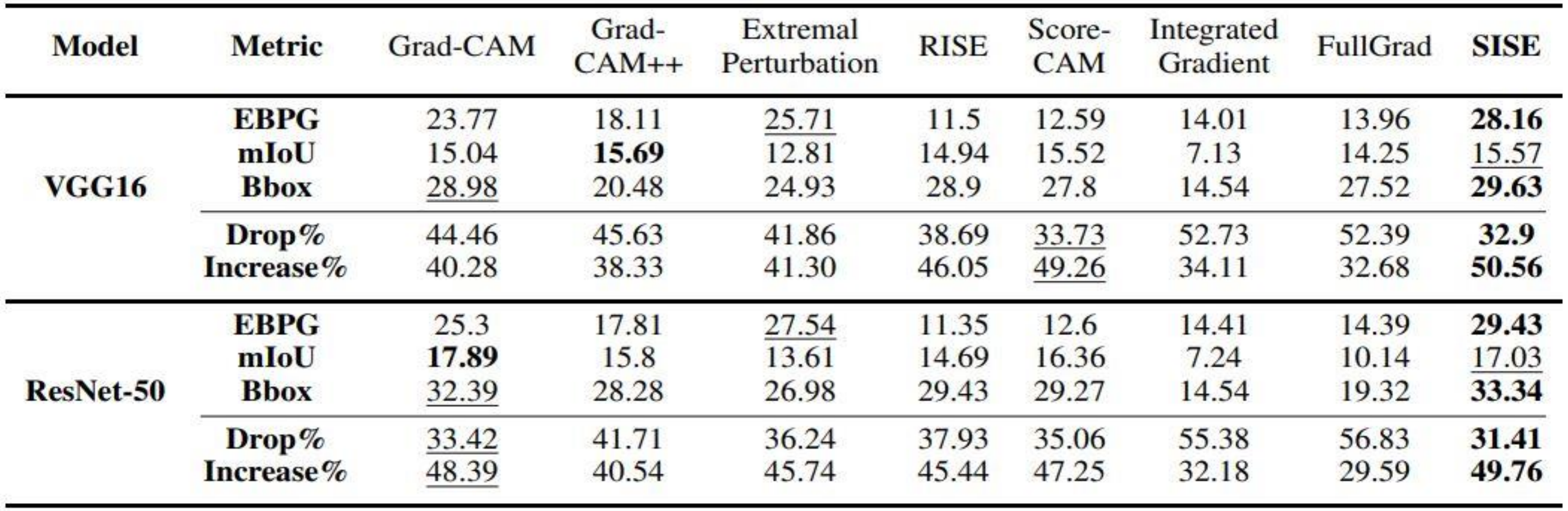
[11] Ramaswamy, H. G.; et al. 2020. Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradientfree Localization. In The IEEE Winter Conference on Applications of Computer Vision, 983–991

Empirical Results



Model	**Metric**	Grad-CAM	Grad-CAM++	Extremal Perturbation	RISE	Score-CAM	Integrated Gradient	FullGrad	**SISE**
VGG16	**EBPG**	55.44	46.29	**61.19**	33.44	46.42	36.87	38.72	60.54
	mIoU	26.52	**28.1**	25.44	27.11	27.71	14.11	26.61	27.79
	Bbox	51.7	55.59	51.2	54.59	54.98	33.97	54.17	**55.68**
	Drop%	49.47	60.63	43.90	39.62	39.79	64.74	60.78	**38.40**
	Increase%	31.08	23.89	32.65	37.76	36.42	26.17	22.73	**37.96**
ResNet-50	**EBPG**	60.08	47.78	63.24	32.86	35.56	40.62	39.55	**66.08**
	mIoU	**32.16**	30.16	26.29	27.4	31.0	15.41	20.2	31.37
	Bbox	60.25	58.66	52.34	55.55	60.02	34.79	44.94	**61.59**
	Drop%	35.80	41.77	39.38	39.77	35.36	66.12	65.99	**30.92**
	Increase%	36.58	32.15	34.27	37.08	37.08	24.24	25.36	**40.22**

Dataset: PASCAL VOC 2007



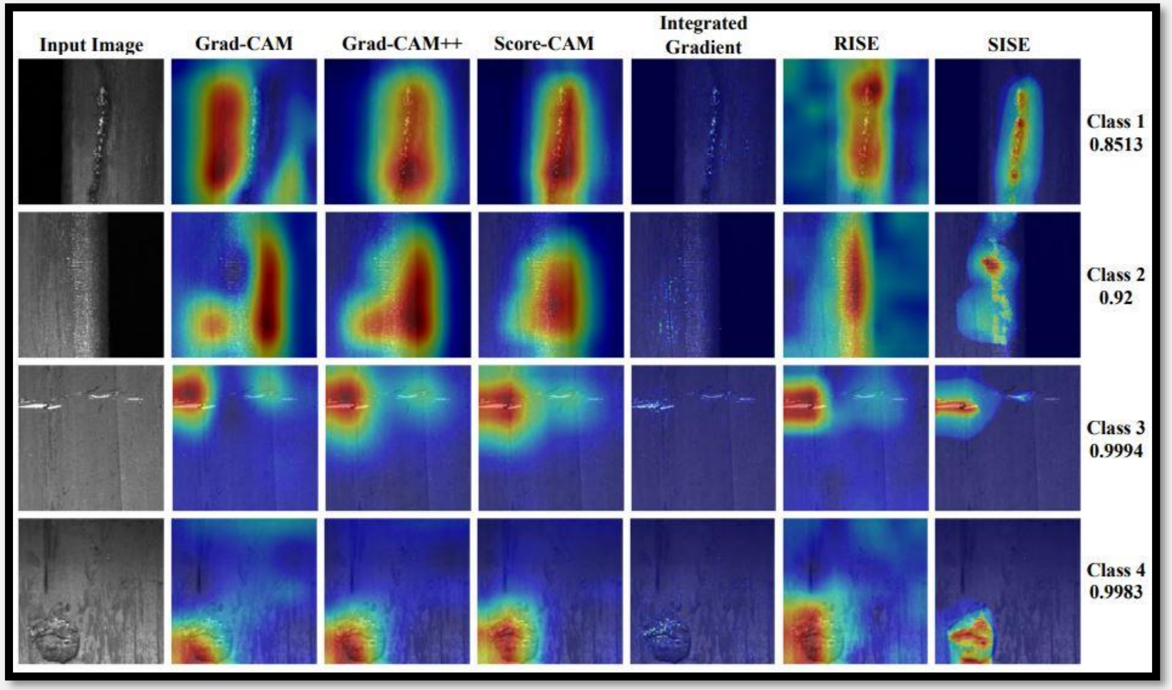
Model	**Metric**	Grad-CAM	Grad-CAM++	Extremal Perturbation	RISE	Score-CAM	Integrated Gradient	FullGrad	**SISE**
VGG16	**EBPG**	23.77	18.11	25.71	11.5	12.59	14.01	13.96	**28.16**
	mIoU	15.04	**15.69**	12.81	14.94	15.52	7.13	14.25	15.57
	Bbox	28.98	20.48	24.93	28.9	27.8	14.54	27.52	**29.63**
	Drop%	44.46	45.63	41.86	38.69	33.73	52.73	52.39	**32.9**
	Increase%	40.28	38.33	41.30	46.05	49.26	34.11	32.68	**50.56**
ResNet-50	**EBPG**	25.3	17.81	27.54	11.35	12.6	14.41	14.39	**29.43**
	mIoU	**17.89**	15.8	13.61	14.69	16.36	7.24	10.14	17.03
	Bbox	32.39	28.28	26.98	29.43	29.27	14.54	19.32	**33.34**
	Drop%	33.42	41.71	36.24	37.93	35.06	55.38	56.83	**31.41**
	Increase%	48.39	40.54	45.74	45.44	47.25	32.18	29.59	**49.76**

Dataset: MS COCO 2014

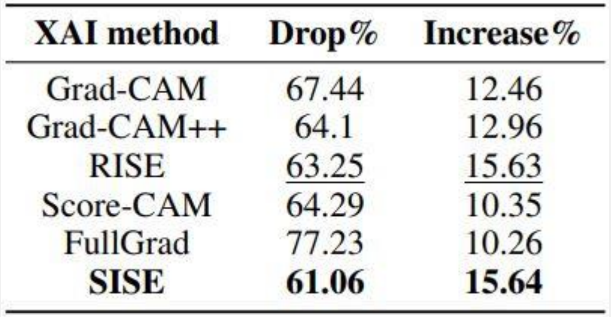
SISE in Visual Anomaly Inspection

Challenges

1. Class imbalance
2. Intra-class variance
3. Inter-class similarity
4. Abstract patterns representing each class.
5. Lack of segmentation masks to evaluate ground truth-based metrics.



Dataset: PAO Severstal, Model: ResNet-101



XAI method	Drop%	Increase%
Grad-CAM	67.44	12.46
Grad-CAM++	64.1	12.96
RISE	63.25	15.63
Score-CAM	64.29	10.35
FullGrad	77.23	10.26
SISE	**61.06**	**15.64**

Dataset: PAO Severstal, Model: ResNet-101

Additional experiments & Discussion

Discussion

1. SISE requires significantly less number of masked images to work, rather than similar methods (e.g., RISE, Score-CAM).

2. However, discarding trivial and manipulating masked images, is a great contribution provided by SISE.

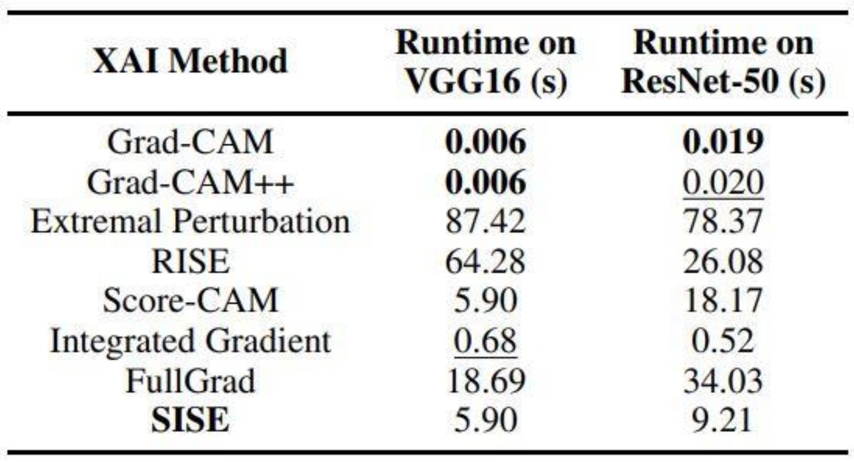
3. Our proposed method shows more accurate performance while dealing with smaller objects.

4. Grad-CAM/Grad-CAM++ are yet the fastest methods, since they rely on only a single forward pass and a single backward pass.

5. By increasing the threshold parameter μ, SISE runs faster, in turn with degrading its explanation ability.

6. The performance degradation in SISE is better to be quantified via model-truth based metrics.

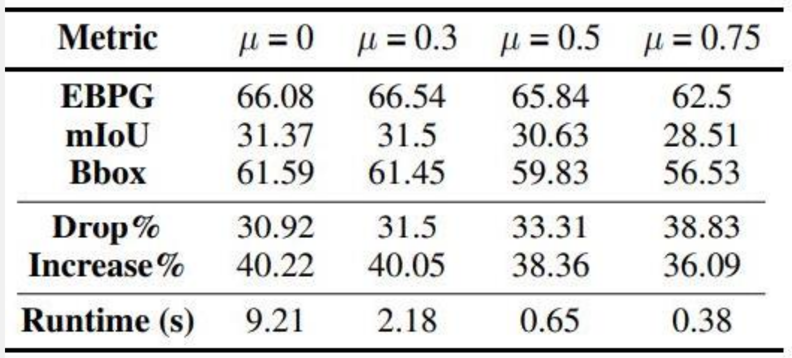
Complexity evaluation



XAI Method	Runtime on VGG16 (s)	Runtime on ResNet-50 (s)
Grad-CAM	**0.006**	**0.019**
Grad-CAM++	**0.006**	0.020
Extremal Perturbation	87.42	78.37
RISE	64.28	26.08
Score-CAM	5.90	18.17
Integrated Gradient	0.68	0.52
FullGrad	18.69	34.03
SISE	5.90	9.21

Dataset: PASCAL VOC 2007

Ablation study



Metric	$\mu = 0$	$\mu = 0.3$	$\mu = 0.5$	$\mu = 0.75$
EBPG	66.08	66.54	65.84	62.5
mIoU	31.37	31.5	30.63	28.51
Bbox	61.59	61.45	59.83	56.53
Drop%	30.92	31.5	33.31	38.83
Increase%	40.22	40.05	38.36	36.09
Runtime (s)	9.21	2.18	0.65	0.38

Dataset: PASCAL VOC 2007, Model; ResNet-50

Takeaways

SISE

1. Improving resolution of explanation maps by aggregating mid-level, and high-level features extracted by a target CNN.

2. Enhancing the explanation maps in terms of class-distinctiveness and completeness, by proposing a method to extract attribution masks.

3. Decreasing the computational overhead of the prior perturbation-based methods, besides strengthening the properties of SISE explanation maps that are crucial to gain users' trust in the target model.

4. Verifying the effectiveness of SISE by setting up extensive experiments using various model and datasets.

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.
- Veit, A.; Wilber, M. J.; and Belongie, S. 2016. Residual networks behave like ensembles of relatively shallow networks. In Advances in neural information processing systems, 550– 558.
- Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2007. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

Thank you. Questions?