INTEGRATED GRAD-CAM: SENSITIVITY-AWARE VISUAL EXPLANATION OF DEEPCONVOLUTIONAL NETWORKS VIA INTEGRATED GRADIENT-BASED SCORING

Sam Sattarzadeh[1], Mahesh Sudhakar[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

Overview of the presentation

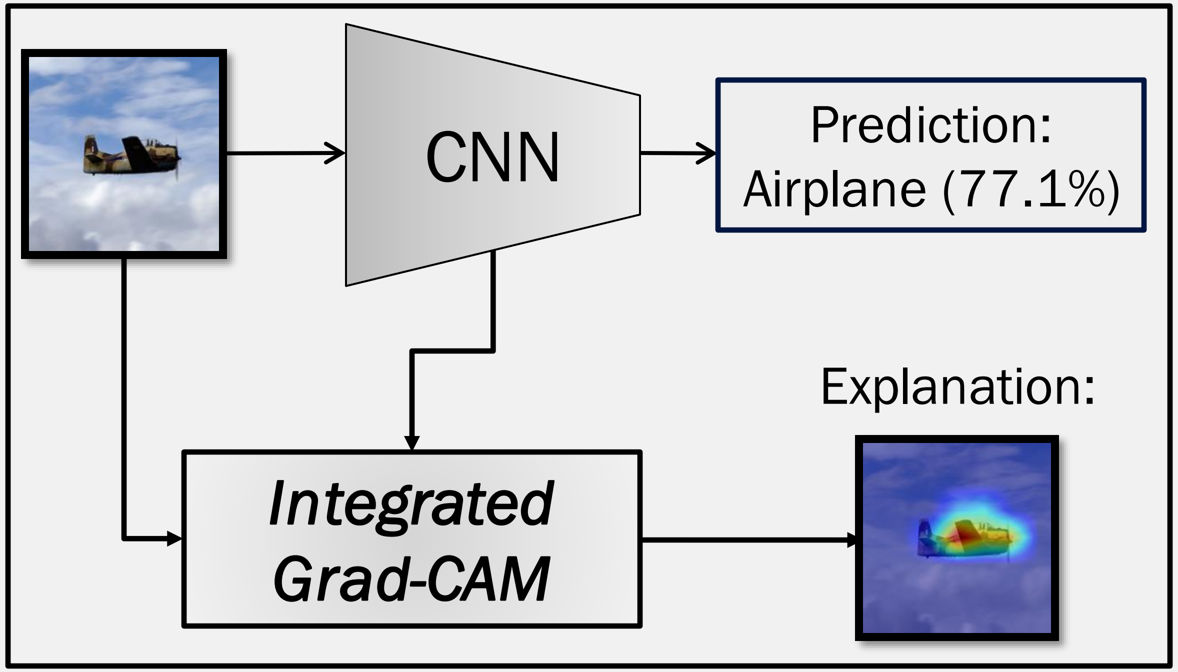
- Explainable AI: Motivation, Applications
- Problem statement
- Our proposed method: Integrated Grad-CAM (IG-CAM)
- Empirical results
- Conclusion
- References

Motivation

Explainable AI (XAI):
Understanding Convolutional Neural Networks (CNNs) is crucial for high-impact and high-risk applications in computer vision[1,2].

CNN-specific attribution methods:
Visualizing the input features responsible for CNN prediction.
(A branch of *post-hoc* and *local* XAI algorithms)

Impactful in:
- **Industrial Applications:**
Medicine, Autonomous Driving, Criminal Justice, Finance
- **Research Fields:**
Object Recognition, Semantic Segmentation, Model Debugging, Dataset Bias Detection, etc.



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

[1] https://ai.googleblog.com/2018/12/providing-gender-specific-translations.html
[2] 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.

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, SmoothGrad)*.

- **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, Occlusion)*.

Our focus: CAM-based methods
Specialized for CNNs, utilized for interpretation and high-level feature visualization.

Problem Statement

CAM-based techniques for CNN interpretation:

- **Grad-CAM**[3] Feature map-wise Gradient-based Weighting.
- **Grad-CAM++** [4] **:** Pixel-wise Gradient-based Weighting.
- **XGrad-CAM** [5]**:** Feature map-wise Axiom-based Weighting.

Our approach: Integrated Grad-CAM

➢ Addressing the limitations of backpropagation in explaining non-linear models.

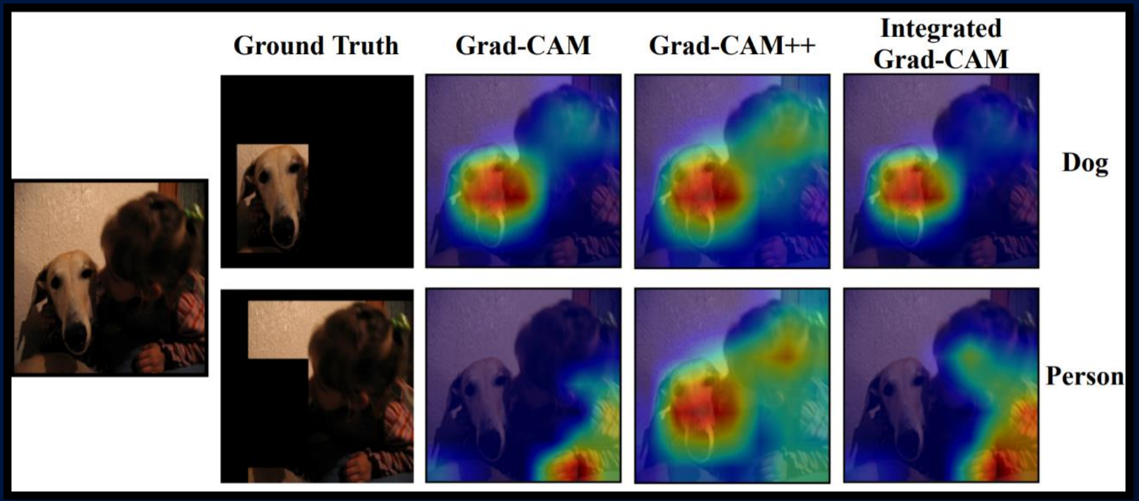
➢ Solving the gradient limitations by employing gradients!

[3] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017.

[4] Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018.

[5] Fu, Ruigang, et al. "Axiom-based grad-cam: Towards accurate visualization and explanation of cnns." arXiv preprint arXiv:2008.02312 (2020).

Integrated Grad-CAM



LG AI Research

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

FACULTY OF APPLIED SCIENCE & ENGINEERING

Integrated Grad-CAM: Intuition

Integrated Gradients[6]:

- **Addressing the issues in the method "Vanilla Backpropagation".**

- **Guarantees the Sensitivity axiom:**

  "For each pair of input and baseline differing only in one feature, an attribution method should highlight this difference by assigning different values corresponding to that feature."

- **Idea:** Calculating the integral of gradient values in a path that links a specific baseline to the input.

- **Takeaways:**

  Enhanced clarity of explanations.

  Improved estimation of the features' contribution in the model's prediction.

[6] Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." International Conference on Machine Learning. PMLR, 2017.

Path Integral

Path Information:

$$\gamma(\alpha) = I' + f(\alpha) \times (I - I') \qquad (0 \leq \alpha \leq 1)$$

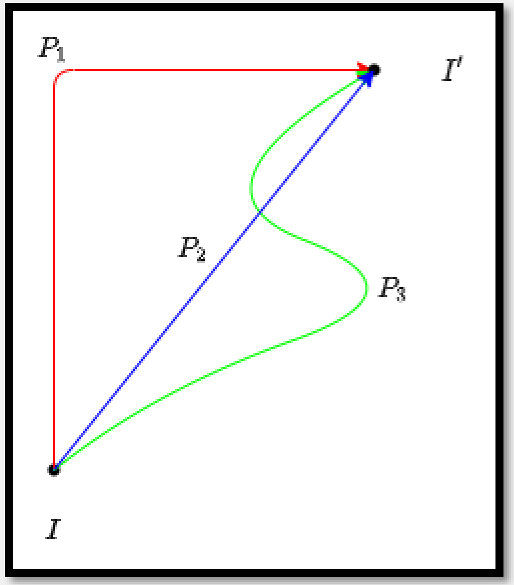
$f(\alpha)$: $\mathbb{R} \longrightarrow \mathbb{R}$: Differentiable & Monotonically Increasing

$$f(0) = 0 \quad \& \quad f(1) = 1$$

Integral Gradients:

For each pair of functions $(h(.), g(.))$:

$$PathIG_{h,g}(I) = \int_{\alpha=0}^{1} \frac{\partial h(\gamma(\alpha))}{\partial g(\gamma(\alpha))} \left[g(\gamma(\alpha)) - g(I')\right] d\alpha$$



Some paths linking I and I' in the image domain

(Input: I - Baseline: I')

Integrated Grad-CAM: Intuition

How "Integrated Gradients" can estimate the features' importance more accurately than "Vanilla Gradient"?

Example: $i_1 = i_2 = 1 \rightarrow y = 1$

Importance of i_1 in the model's prediction ($S(i_1)$):

Vanilla Gradient: $S(i_1) = \frac{\partial y}{\partial i_1}|_{i_1=i_2=1} = \mathbf{0}$

Integrated Gradients: $\begin{cases} \gamma(\alpha) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \alpha \times \left(\begin{bmatrix} i_1 \\ i_2 \end{bmatrix} \right) = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} \\ S(i_1) = \int_{\alpha=0}^{1} \frac{\partial y}{\partial \gamma_1(\alpha)} \gamma_1(\alpha) d\alpha = \mathbf{0.5} \end{cases}$

The same idea can be proposed to improve Grad-CAM!

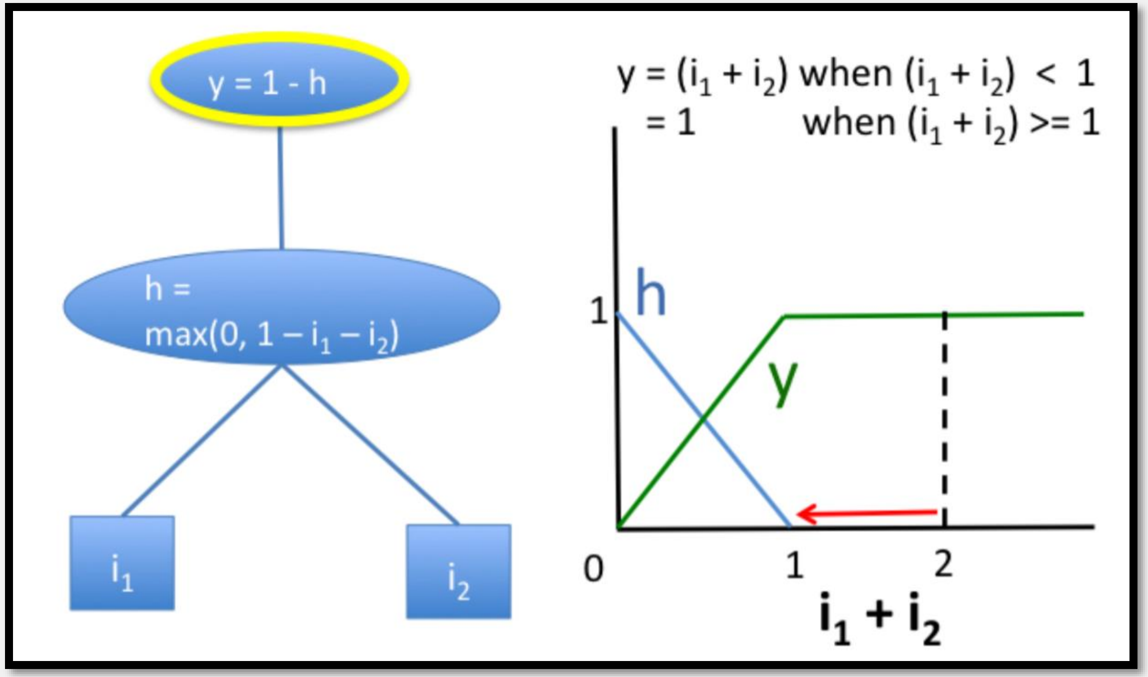


Image Credit: [7]

[6] Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje. "Learning important features through propagating activation differences." International Conference on Machine Learning. PMLR, 2017.

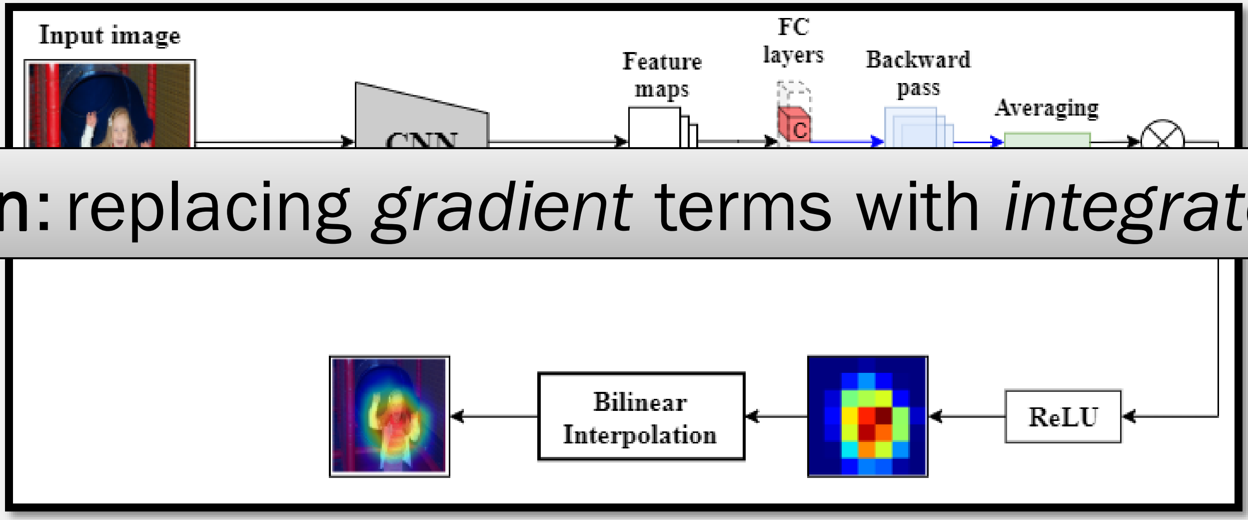
Grad-CAM formulation

While feeding the CNN with the input image "$\boldsymbol{I}$":

Feature maps in the convolutional layer "l":

$$\{A^{1l}(I), A^{2l}(I), \dots, A^{Nl}(I)\}$$

Model's confidence score for class "c“: $y_c(I)$



Our modification: replacing *gradient* terms with *integrated gradient* terms

Grad-CAM Explanation map: $M^{c}_{Grad-CAM} = \mathrm{ReLU}(\sum_{k=1}^{N}(\sum_{i,j}\frac{\partial y_c(I)}{\partial A^{kl}_{i,j}(I)})A^{kl}(I))$

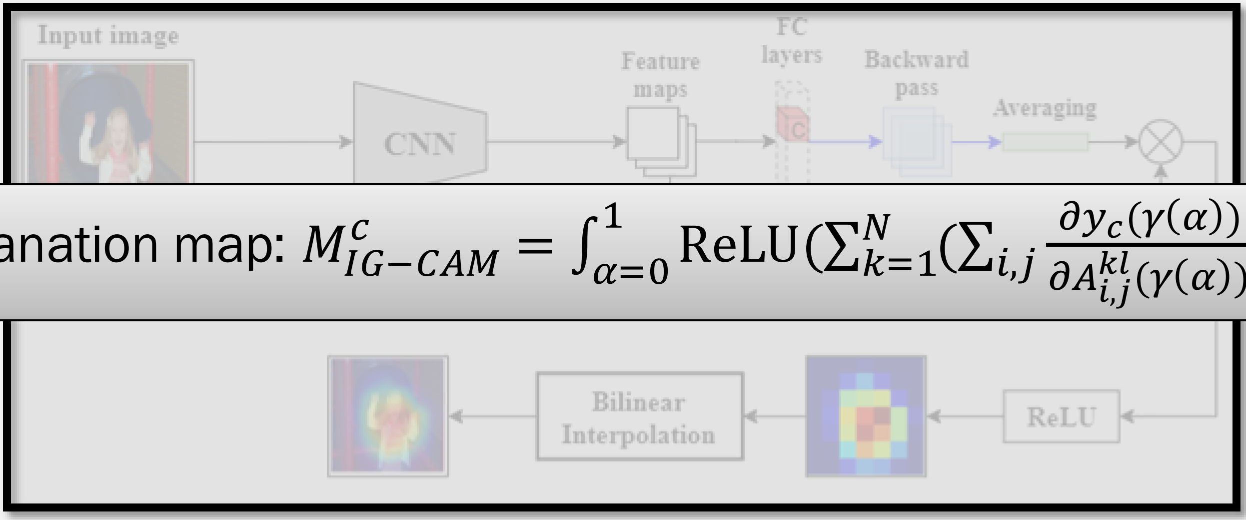
Integrated Grad-CAM formulation

While feeding the CNN with the input image "I":

Feature maps in the convolutional layer "l":

$$\{A^{1l}(I), A^{2l}(I), \dots, A^{Nl}(I)\}$$

Model's confidence score for class "c": $y_c(I)$



Integrated Grad-CAM Explanation map: $M^{c}_{IG-CAM} = \int_{\alpha=0}^{1} \mathrm{ReLU}(\sum_{k=1}^{N}(\sum_{i,j}\frac{\partial y_c(\gamma(\alpha))}{\partial A^{kl}_{i,j}(\gamma(\alpha))})(A^{kl}(\gamma(\alpha)) - (A^{kl}(I'))))d\alpha$

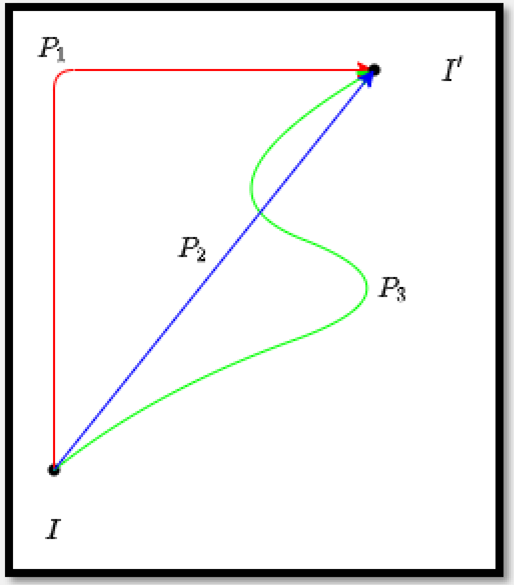
Grad-CAM Explanation map: $M^{c}_{Grad-CAM} = \mathrm{ReLU}(\sum_{k=1}^{N}(\sum_{i,j}\frac{\partial y_c(I)}{\partial A^{kl}_{i,j}(I)})A^{kl}(I))$

Integrated Grad-CAM formulation

Limitation of our equation:
The equation below is hard to implement.

Solution:
Approximating our equation with a summation.

- For simplicity, select a linear path between the input and the baseline.
- Use *Reimann's Approximation.*



Some paths linking I and I' in the image domain

(Input: I - Baseline: I')

Path P_2: $\gamma(\alpha) = I' + f(\alpha) \times (I - I') \qquad (0 \leq \alpha \leq 1)$

$$f(\alpha) = \alpha$$

Integrated Grad-CAM Explanation map: $M^{c}_{IG-CAM} = \int_{\alpha=0}^{1} \mathrm{ReLU}(\sum_{k=1}^{N}(\sum_{i,j} \frac{\partial y_c(\gamma(\alpha))}{\partial A^{kl}_{i,j}(\gamma(\alpha))})(A^{kl}(\gamma(\alpha)) - (A^{kl}(I'))))d\alpha$

Integrated Grad-CAM implementation

Path P_2: $\gamma(\alpha) = I' + \alpha \times (I - I') \qquad (0 \leq \alpha \leq 1)$

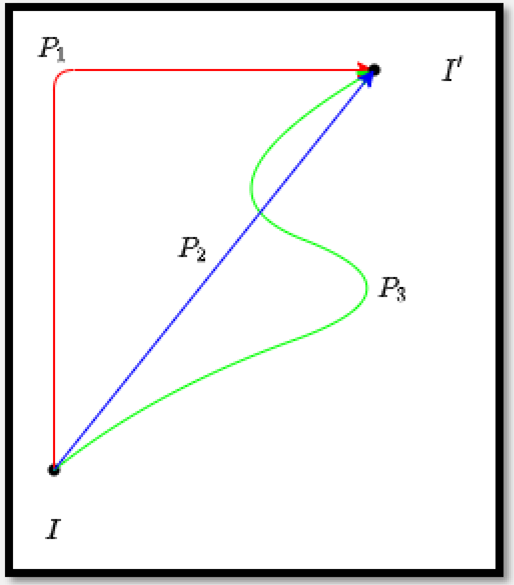
Reimann's Approximation:
Sample m points along the path P_2 with a constant interval.

Interval step: $\frac{1}{m} \quad (m \in \mathbb{N})$

Sampled points: $\alpha \in \{\frac{t}{m} | t = \{1, \dots, m\}\}$

$$M^c_{IG-CAM} = \int_{\alpha=0}^{1} \mathrm{ReLU}(\sum_{k=1}^{N}(\sum_{i,j} \frac{\partial y_c(\gamma(\alpha))}{\partial A^{kl}_{i,j}(\gamma(\alpha))})(A^{kl}(\gamma(\alpha)) - (A^{kl}(I'))))d\alpha$$

$$M^c_{IG-CAM} \cong \sum_{t=1}^{m} \mathrm{ReLU}(\frac{1}{m}\sum_{k=1}^{N}(\sum_{i,j} \frac{\partial y_c(\gamma\left(\frac{t}{m}\right))}{\partial A^{kl}_{i,j}(\gamma\left(\frac{t}{m}\right))})(A^{kl}\left(\gamma\left(\frac{t}{m}\right)\right) - (A^{kl}(I'))))d\alpha$$

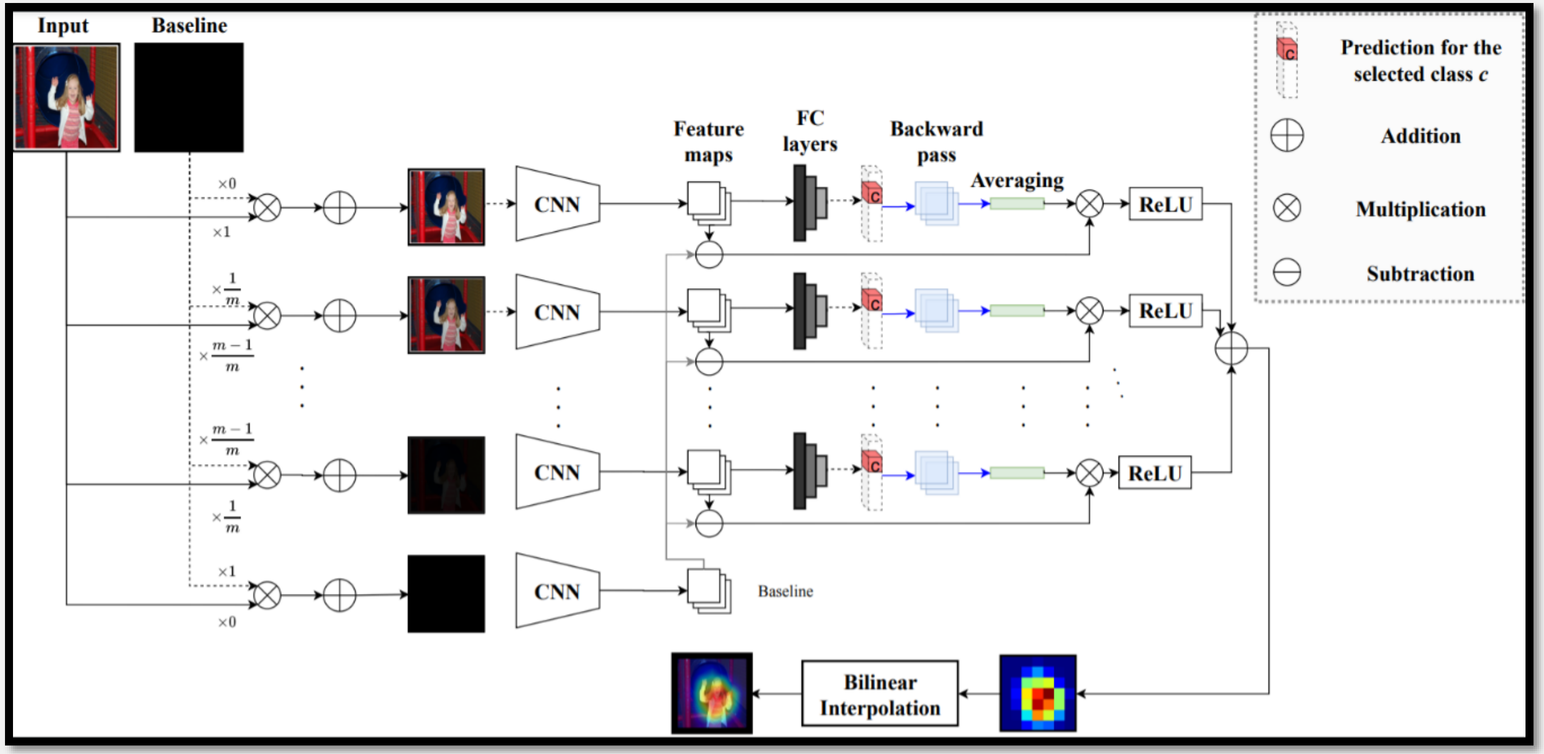


Some paths linking I and I' in the image domain

(Input: I - Baseline: I')

Integrated Grad-CAM implementation

IG-CAM can be modelled by applying Grad-CAM to translated copies of the input image.

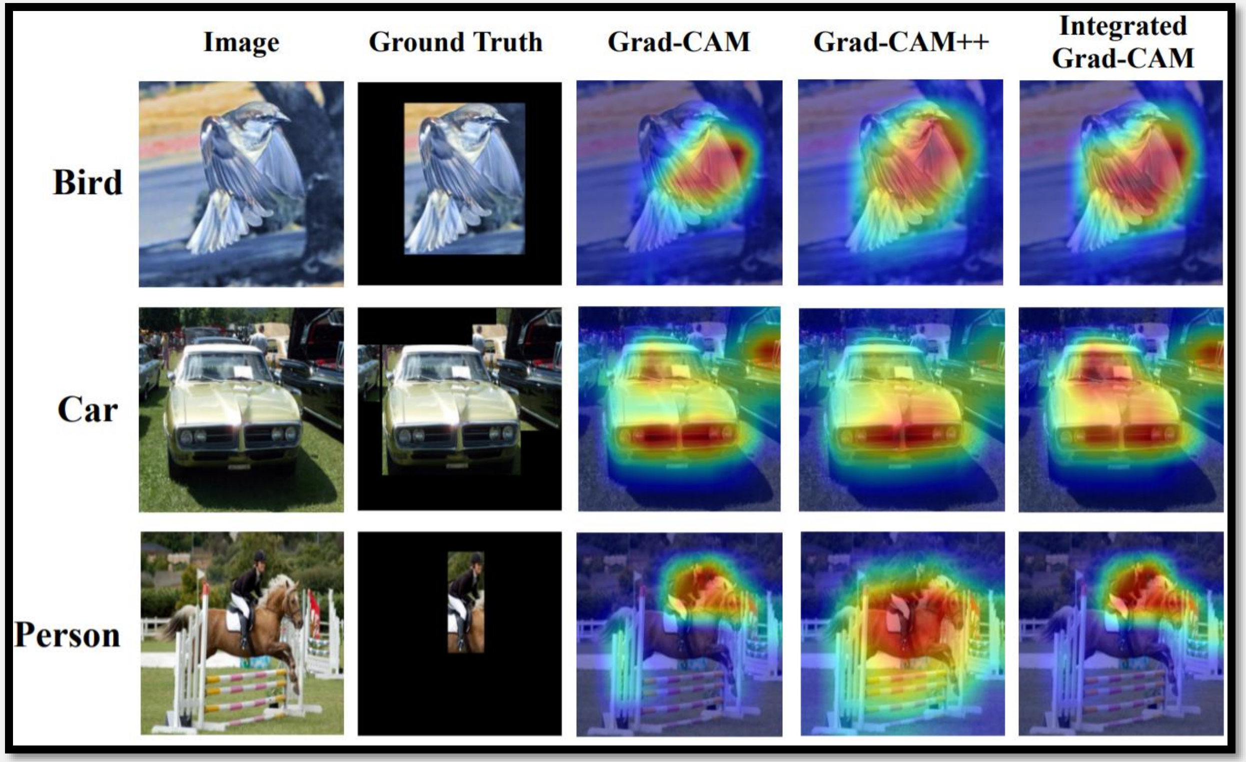


$$M^{c}_{IG-CAM} \cong \sum_{t=1}^{m} \mathrm{ReLU}\Big(\frac{1}{m}\sum_{k=1}^{N}\Big(\sum\nolimits_{i,j}\frac{\partial y_c(\gamma\left(\frac{t}{m}\right))}{\partial A^{kl}_{i,j}(\gamma\left(\frac{t}{m}\right))}\Big)\Big(A^{kl}\left(\gamma\left(\frac{t}{m}\right)\right) - (A^{kl}(I'))\Big)\Big)d\alpha$$

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].
- In out experiments, the number of intervals for IG-CAM is set to $m=20$.



[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

Ground truth-based metrics

Verifying the meaningfulness of explanation methods, and their preciseness 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)).

Dataset: PASCAL VOC 2007

	Metric	Grad-CAM	Grad-CAM++	Integrated Grad-CAM
VGG16	EBPG(%)	55.44	46.29	**55.94**
	Bbox(%)	51.7	55.59	**55.6**
	Drop(%)	49.47	60.63	**47.96**
	Increase(%)	31.08	23.89	**31.47**
ResNet-50	EBPG(%)	60.08	47.78	**60.41**
	Bbox(%)	60.25	58.66	**61.94**
	Drop(%)	35.80	41.77	**34.49**
	Increase(%)	36.58	32.15	**36.84**

For each metric, the best is shown in bold.

[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.

Empirical Results

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).

Dataset: PASCAL VOC 2007

	Metric	Grad-CAM	Grad-CAM++	Integrated Grad-CAM
VGG16	EBPG(%)	55.44	46.29	**55.94**
	Bbox(%)	51.7	55.59	**55.6**
	Drop(%)	49.47	60.63	**47.96**
	Increase(%)	31.08	23.89	**31.47**
ResNet-50	EBPG(%)	60.08	47.78	**60.41**
	Bbox(%)	60.25	58.66	**61.94**
	Drop(%)	35.80	41.77	**34.49**
	Increase(%)	36.58	32.15	**36.84**

For each metric, the best is shown in bold.

[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

Complexity Analysis

Dataset: PASCAL VOC 2007

Model	Grad-CAM	Grad-CAM++	IG-CAM (m=20)	IG-CAM (m=50)
ResNet-50	11.3 ms	12.2 ms	54.8 ms	108.08 ms

Average run-time on different models

Insights:

- The number of calls in IG-CAM ("*m*") does not improve its performance significantly, if increased from 20.
- Though IG-CAM runs slower rather than Grad-CAM and Grad-CAM++, the modifications in IG-CAM do not slow this method down considerably.
- Though some perturbation-based methods may outperform IG-CAM, the satisfying speed of our method makes it a desired choice for real-world real-time applications.

Takeaways

<u>IG-CAM</u>

1. Circumvented the underestimations in Grad-CAM and Grad-CAM++.

2. Addressed the issues caused by backpropagation in the methods above

3. Though slower than the conventional methods, offers an acceptable run-time to be used in real-world applications.

4. The takeaways of IG-CAM are verified through extensive experiments on the PASCAL VOC 2007 dataset.

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