Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks

Vision Lab @ JHU
http://www.vision.jhu.edu

Aditya Chattopadhyay[1] Anirban Sarkar[2] Prantik Howlader[2] Vineeth N Balasubramanian[2]

Center for Imaging Science, Johns Hopkins University, Baltimore, USA[1], Department of Computer Science & Engineering, Indian Institute of Technology, Hyderabad, India[2]

Motivation

- Develop explainable Convolutional Neural Network (CNN) models for proper understanding of their internal functioning.
- Provide good visual explanations of CNN decisions which are both faithful to the model as well as helps inculcate human trust in the model.

Contributions

- Building on recently proposed methods, CAM[2] & Grad-CAM[1], we propose a generalization called Grad-CAM++ that can provide better visual explanations of CNN model predictions.
- The proposed method exhibits better object localization as well as explains occurrences of multiple object instances in a single image, when compared to Grad-CAM.
- Our extensive experiments and evaluations, both subjective and objective, on standard datasets showed that Grad-CAM++ provides promising human-interpretable visual explanations for a given CNN architecture across multiple tasks including classification, image caption generation and 3D action recognition; as well as in new settings such as knowledge distillation.

Methodology

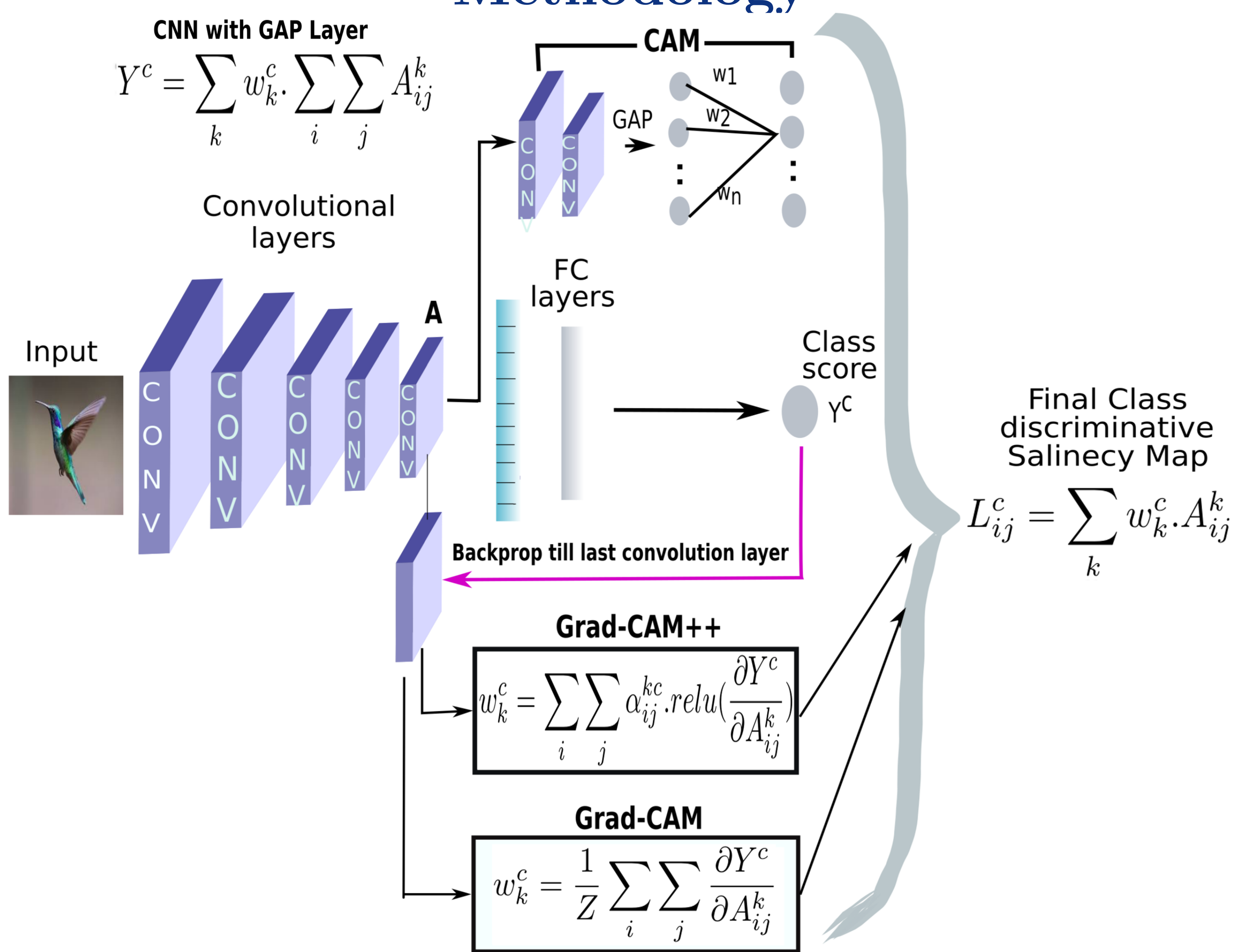

Fig. 1: An overview of all the three methods CAM, Grad-CAM, Grad-CAM++

$$Y^c = \sum_k [\sum_i \sum_j \{\sum_a \sum_b \alpha^{kc}_{ab} . relu(\frac{\partial Y^c}{\partial A^k_{ab}})\} A^k_{ij}]$$

$$\frac{\partial Y^c}{\partial A^k_{ij}} = \sum_a \sum_b \alpha^{kc}_{ab} . \frac{\partial Y^c}{\partial A^k_{ab}} + \sum_a \sum_b A^k_{ab} \{\alpha^{kc}_{ij} . \frac{\partial^2 Y^c}{(\partial A^k_{ij})^2}\}$$

$$\frac{\partial^2 Y^c}{(\partial A^k_{ij})^2} = 2.\alpha^{kc}_{ij} . \frac{\partial^2 Y^c}{(\partial A^k_{ij})^2} + \sum_a \sum_b A^k_{ab} \{\alpha^{kc}_{ij} . \frac{\partial^3 Y^c}{(\partial A^k_{ij})^3}\}$$

$$\alpha^{kc}_{ij} = \frac{\frac{\partial^2 Y^c}{(\partial A^k_{ij})^2}}{2\frac{\partial^2 Y^c}{(\partial A^k_{ij})^2} + \sum_a \sum_b A^k_{ab} \{\frac{\partial^3 Y^c}{(\partial A^k_{ij})^3}\}}$$

Intuition

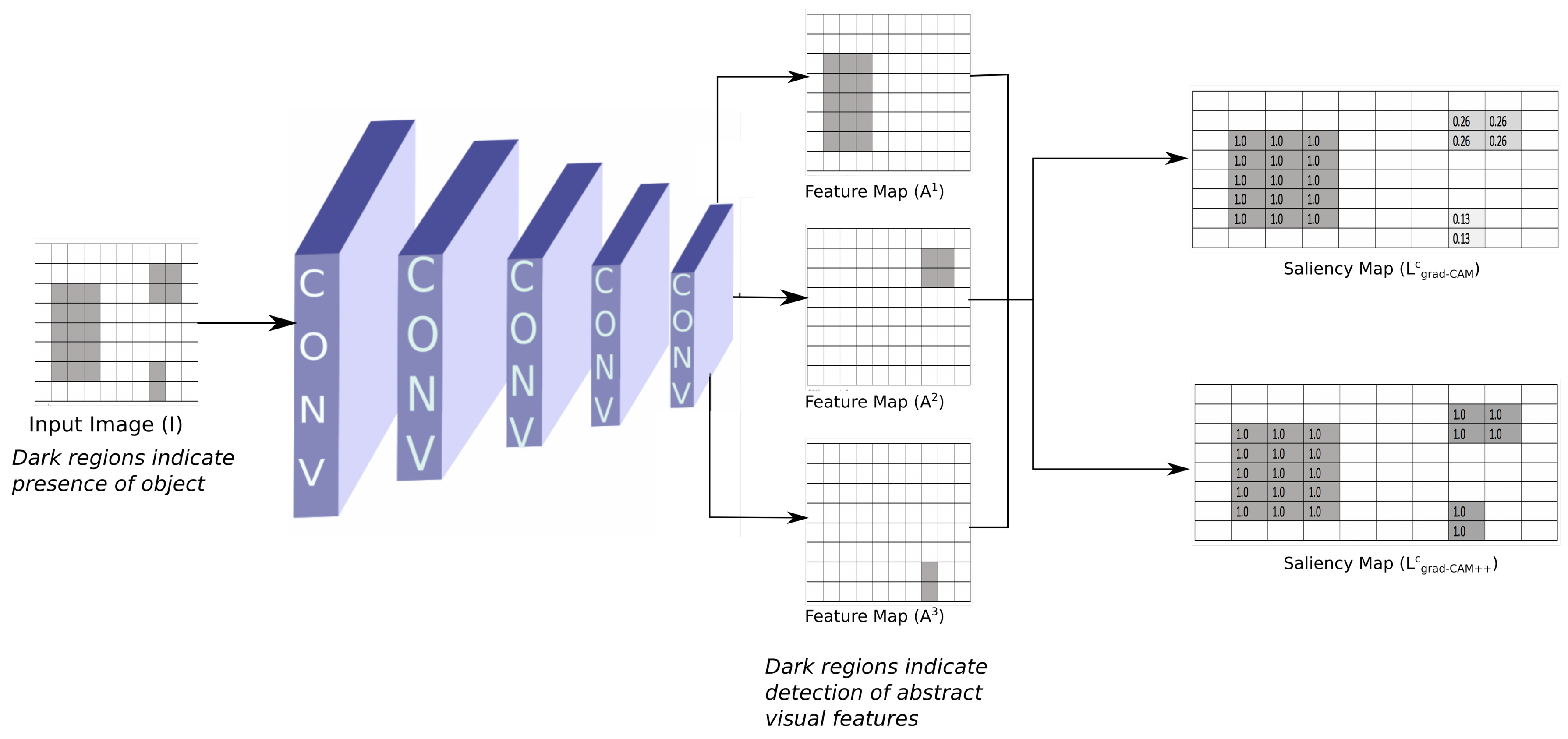

Fig. 2: The CNN task here is binary object classification. Clearly taking a weighted combination of gradients $L^c_{gradCAM++}$ provides better salient features (all the spatially relevant regions of the input image are equally highlighted) than its unweighted counterpart $L^c_{gradCAM}$ (some parts of the object are paled out in the saliency map). The values in the pixels of each saliency map indicates the intensity at that point.

Quantitative Results

Method	Grad-CAM++	Grad-CAM
Avg Drop %	**36.84**	46.56
% Incr in Conf	**17.05**	13.42
Win %	**70.72**	29.28

Table 1: Results for objective evaluation of the explanations generated by Grad-CAM++ and Grad-CAM on ILSVRC2012 val set for VGG-16.

Method	Grad-CAM++	Grad-CAM
Avg Drop %	**19.53**	28.54
% Incr in Conf	**18.96**	21.43
Win %	**61.47**	39.44

Table 2: Results for objective evaluation of the explanations generated by Grad-CAM++ and Grad-CAM on PASCAL VOC 2007 val set for VGG-16.

- Table 1 & 2 support our claim that Grad-CAM++ explanations are more faithful to the underlying model.
- Evaluated human interpretability of our method via mechanical turk experiments.
- 43.88% people preferred Grad-CAM++ visualizations, 22.43 favored Grad-CAMM, while 33.69% were neutral.

Learning From Explanations: Knowledge Distillation

Loss function used	mAP (% increase)
$L_{exp\_student}$**(Grad-CAM++)**	**0.42 (35.5%)**
$L_{cross\_ent} + L_{KD}$	0.34 (9.7%)
L_{cross_ent} [Baseline]	0.31 (0.0%)

Table 3: Results for training a student network with explanations from the teacher (VGG-16 fine-tuned) and with knowledge distillation on PASCAL VOC 2007 dataset. The % increase is with respect to the baseline loss L_{cross_ent}.

Loss function used	Test error rate
L_{cross_ent}	6.78
$L_{exp\_student}$**(Grad-CAM++)**	**6.74**
$L_{exp\_student}$(Grad-CAM)	6.86
$L_{cross\_ent} + L_{KD}$	5.68
$L_{exp\_student}$**(Grad-CAM++)**$+L_{KD}$	**5.56**
$L_{exp\_student}$(Grad-CAM)$+L_{KD}$	5.8

Table 4: Results for knowledge distillation to train a student (WRN-16-2) from a deeper teacher network (WRN-40-2).

$$L_{exp\_student}(c, W_s, W_t, I) = L_{cross\_ent}(c, W_s(I)) + \alpha(L_{interpret}(c, W_s, W_t, I))$$

where $L_{interpret}$ is defined as:

$$L_{interpret}(c, W_s, W_t, I) = ||L^c_s(W_s(I)) - L^c_t(W_t(I))||^2_2$$

Qualitative Results

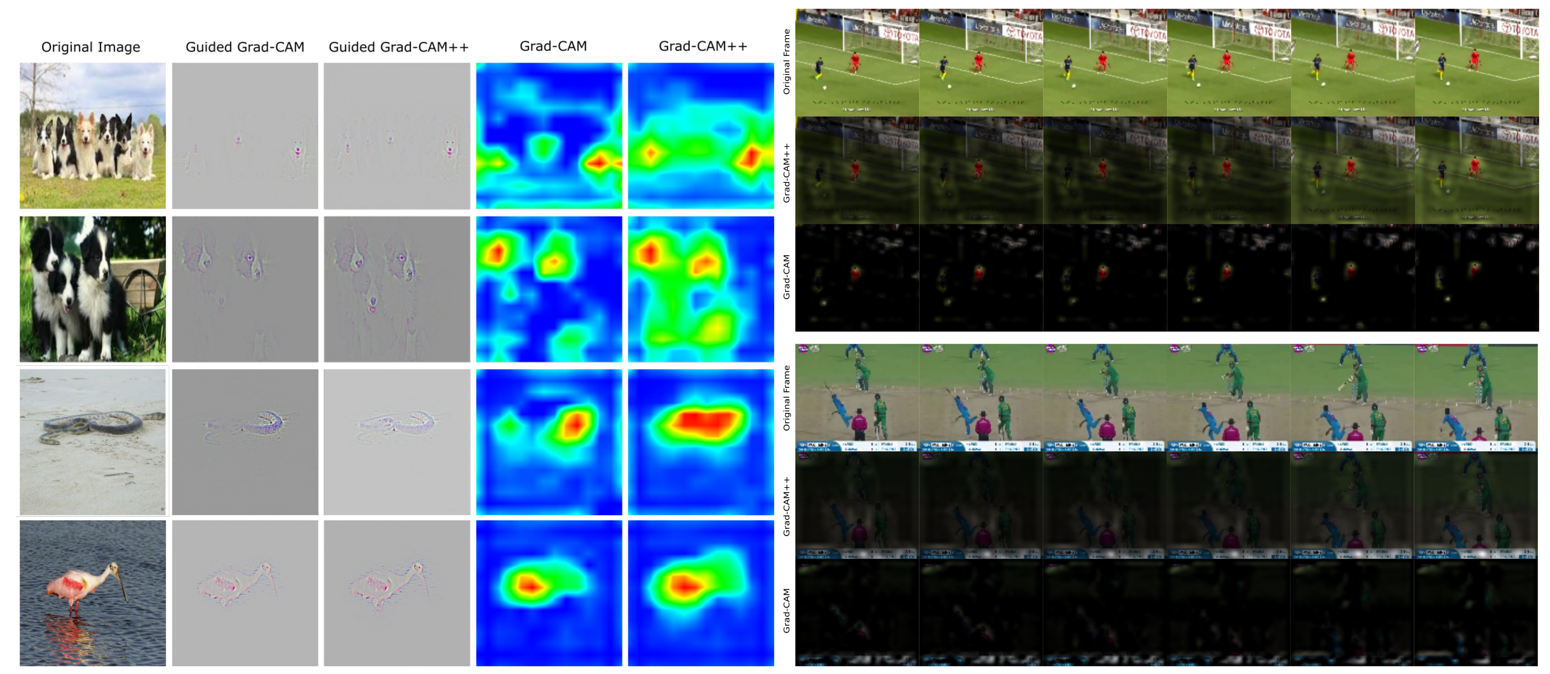

Fig. 3: From left to right: Cols 1-5 highlight effectiveness of Grad-CAM++ in identifying salient regions of images in object classification tasks over Grad-CAM. Cols 6-11 Results for action recognition tasks by 3D-CNNs.

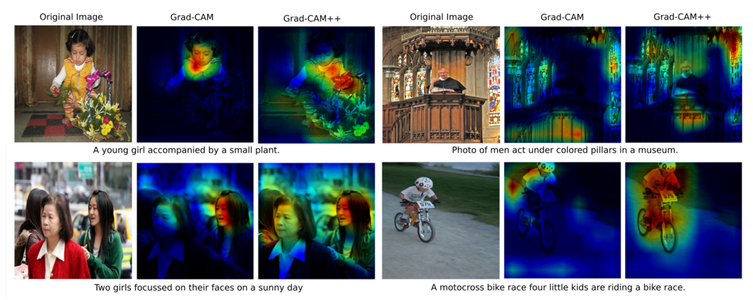

Fig. 4: Results for image-captioning tasks by CNNs.

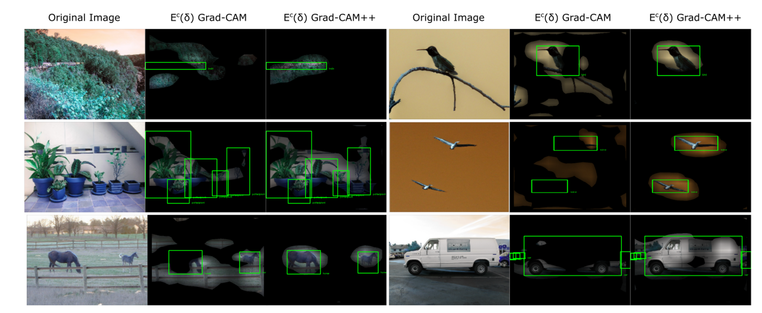

Fig. 5: Results for object localization capabilities of Grad-CAM++.

Does Grad-CAM++ do well because of larger maps?

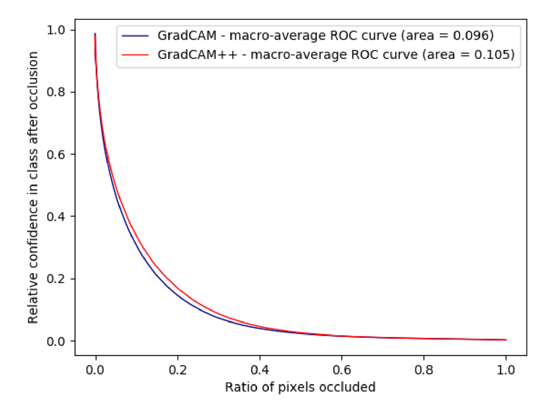

Fig. 6: ROC curve to study the relationship between spatial extents of visual explanations and the corresponding relative confidence when the visual explanation region is provided as input to the model.

- In general, we expect a lower drop in classification score if the explanation map region provided as input to the model for a given image I and class c has greater area.
- A threshold parameter θ (quantile) was varied from 0 to 1 at equally-spaced discrete intervals to generate the curve.
- Observe that at each θ, Grad-CAM++ highlights regions that are as faithful or more to the underlying model than Grad-CAM, irrespective of the spatial extents.

References

[1] Selvaraju et al., Grad-cam: Visual explanations from deep networks via gradient-based localization. ICCV'17.
[2] Zhou et al., Learning deep features for discriminative localization CVPR'16.

Acknowledgements

We thank the Ministry of Human Resource Development, India for financial assistance and NVIDIA for donation of K40 GPU through their Academic Hardware Grant program.

For code: https://github.com/adityac94/Grad_CAM_plus_plus