CS6450- Visual Computing-Presentation 1

LayerCAM : Exploring Hierarchical Class Activation Maps for Localization

Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei

IEEE Transactions on Image Processing (Volume: 30),June 2021

Presented on March 7,2022

Guide: Prof.C.Krishna Mohan, Dept.of CSE, IITH

Teaching Assistant: Mr.Udaya Kumar

Presented by: Prasanna Kumar R [SM21MTECH14001]

Motivation

- Current deep models are very successful in Computer Vision, NLP etc.
- One problem with these networks is the explainability of the results it produces.
- We need explainable or interpretable deep models to gain human trust.
- One of the works which could explain these deep models is the LayerCAM

Previous Works: CAM, GradCAM and GradCAM++

- **Class Activation Mapping (CAM)[1]** localizes the discriminative regions used by a particular task despite not being trained for them.
- **Gradient -weighted Class Activation Mapping (GradCAM)[2]** uses gradient of target concept flowing into final convolutional layer.
- GradCAM is a generalization of CAM for any CNN architecture.
- **GradCAM++ [3]** is same as GradCAM except for different weightage for each pixel in the feature map

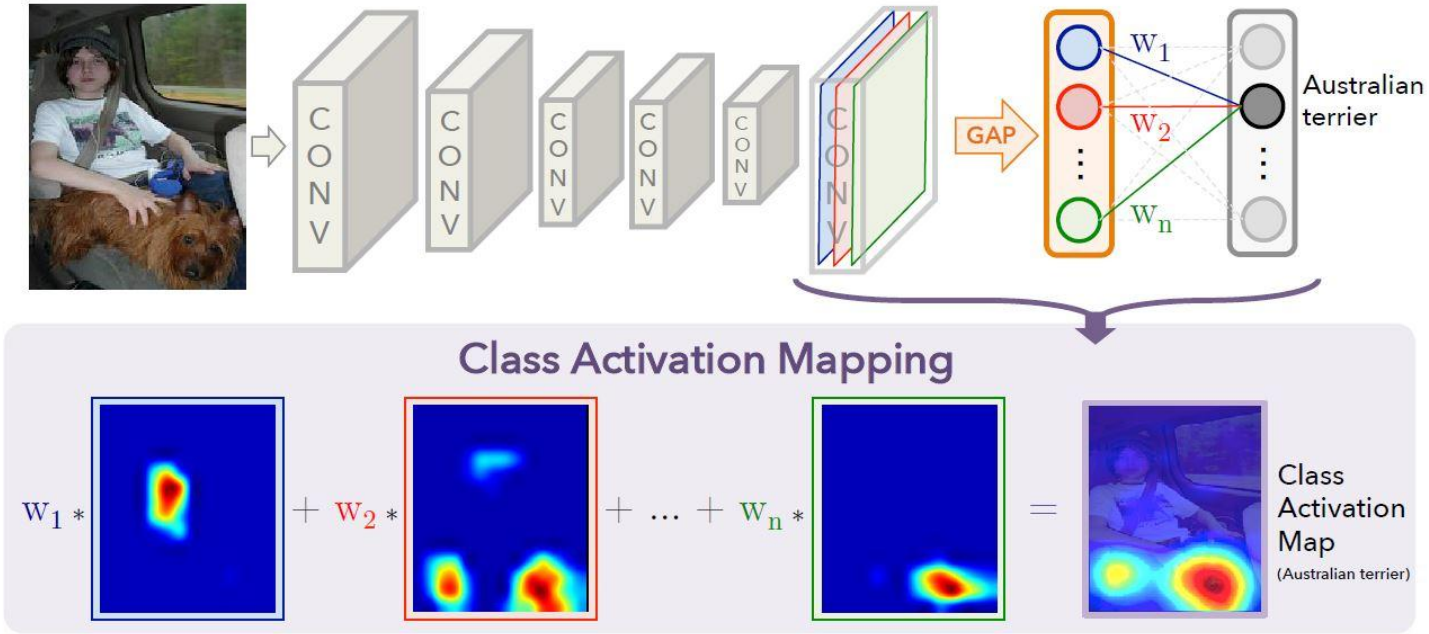
[1]*B. Zhou et al, "Learning deep features for discriminative localization," CVPR 2016*

[2]*R. R. Selvaraju, et al "Grad-cam: Visual explanations from deep networks via gradient-based localization," ICCV 2017*

[3]*A. Chattopadhay et al "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," WACV,2018*

CAM

Helps us to know what the network is looking at while predicting the class



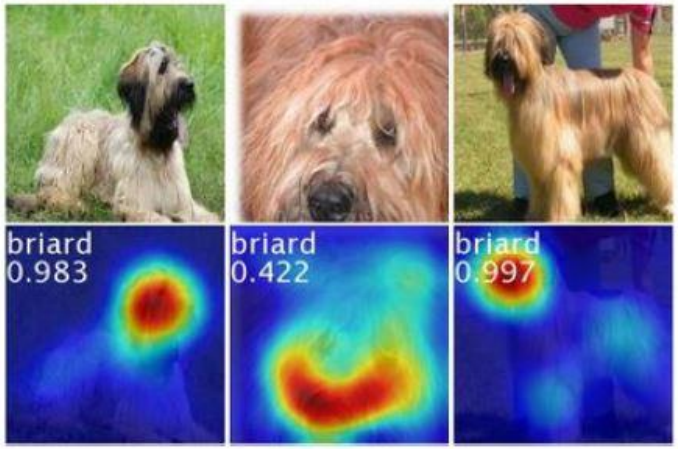
Picture Credits: *B. Zhou et al, "Learning deep features for discriminative localization," CVPR 2016*

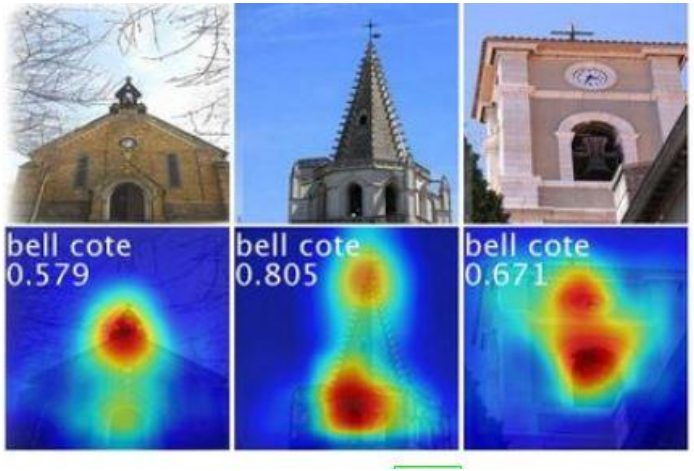
CAM:

Mathematical Expression:

$$Y^c = \sum_k w_k^c \cdot \frac{1}{Z} \sum_i \sum_j A_{ij}^k$$

Examples:





Picture Credits: *B. Zhou et al, "Learning deep features for discriminative localization," CVPR 2016*

CAM: Merits and Demerits

Merits:

- Can localize objects without positional supervision.

Demerits:

- Retraining needed to explain trained models.
- Constrained on CNN architecture.
- Model may trade off accuracy for interpretability.

Grad-CAM

- Applicable to wide variety of CNN-model families
- No Global Average Pooling(GAP) is needed.
- The Class feature weights are gradient themselves. No retraining needed!

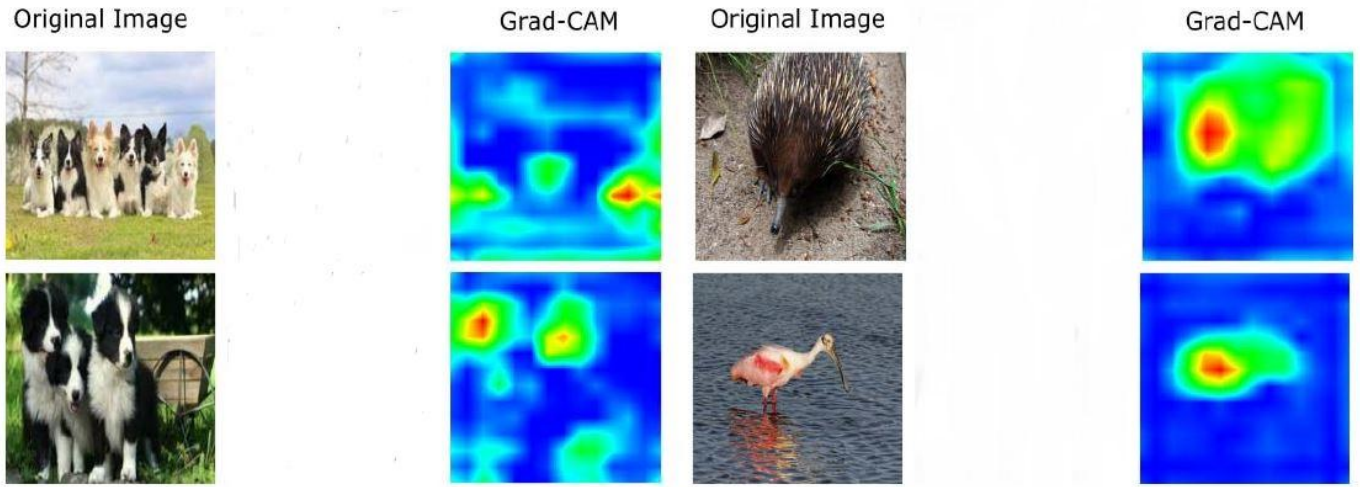
$$w_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k}$$

- After finding the weights, the Grad-CAM heat map is given by:

$$L_{Grad-CAM}^c = ReLU\left(\sum_k w_k^c A^k\right)$$

Grad-CAM: Demerits

- Grad-CAM fails to properly localize objects in an image if it contains multiple occurrences of the same class.
- Unsatisfactory localization performance, especially under occlusion.



Picture Credits: *A. Chattopadhay et al "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," WACV,2018*

Grad-CAM++:

- As we have seen, Grad-CAM considers all pixel gradients equally when computing importance weights of activation/feature maps.

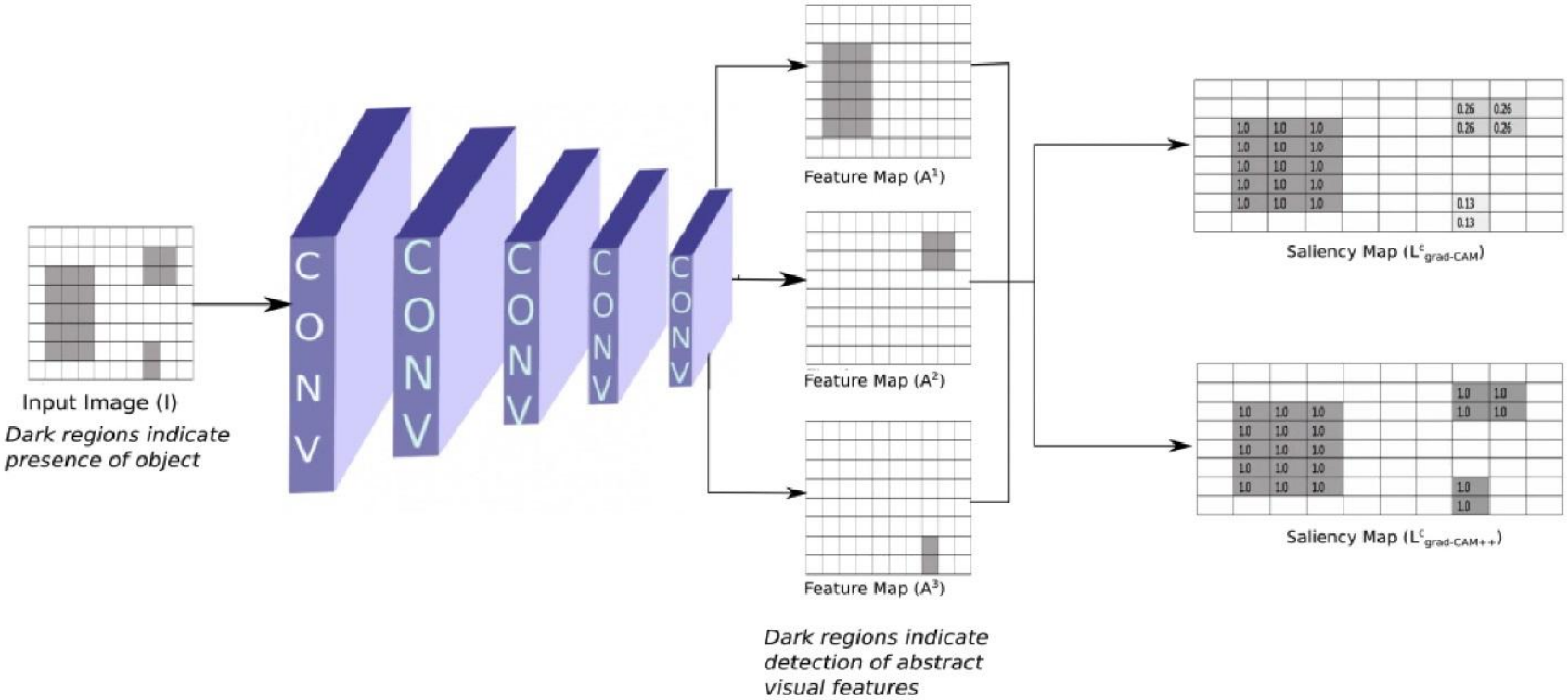
$$w_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k}$$

- This can suppress activation maps with comparatively lesser spatial footprint.

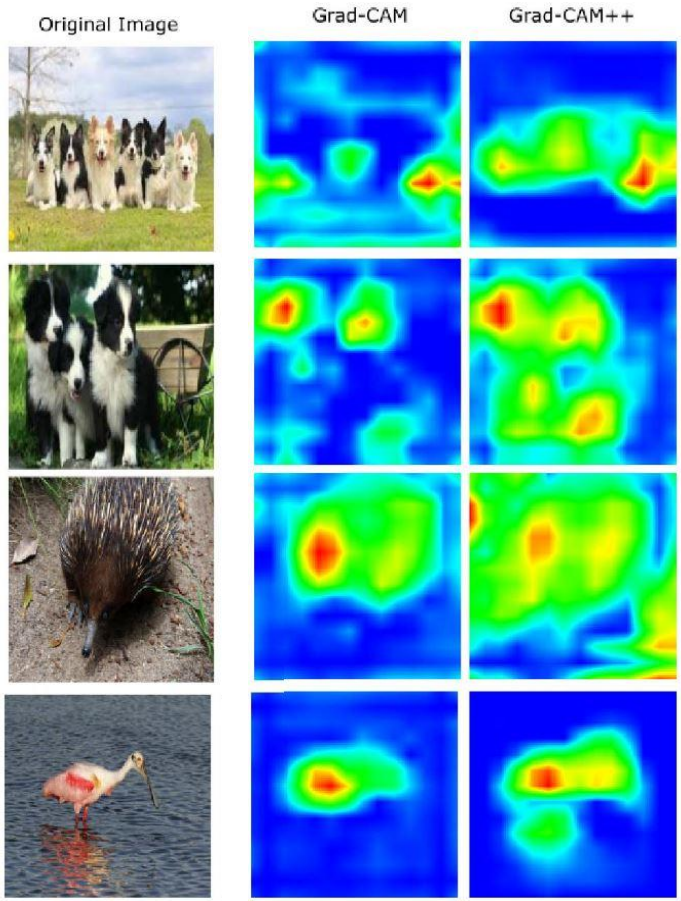
- The above limitations are overcome by Grad-CAM++ by taking a weighted combination of positive partial derivatives instead of a global average.

$$w_k^c = \sum_i \sum_j \alpha_{ij}^{kc} . relu\left(\frac{\partial Y^c}{\partial A_{ij}^k}\right) \quad \text{where,} \quad \alpha_{ij}^{kc} = \frac{\frac{\partial^2 Y^c}{(\partial A_{ij}^k)^2}}{2\frac{\partial^2 Y^c}{(\partial A_{ij}^k)^2} + \sum_a \sum_b A_{ab}^k \left\{\frac{\partial^3 Y^c}{(\partial A_{ij}^k)^3}\right\}}$$

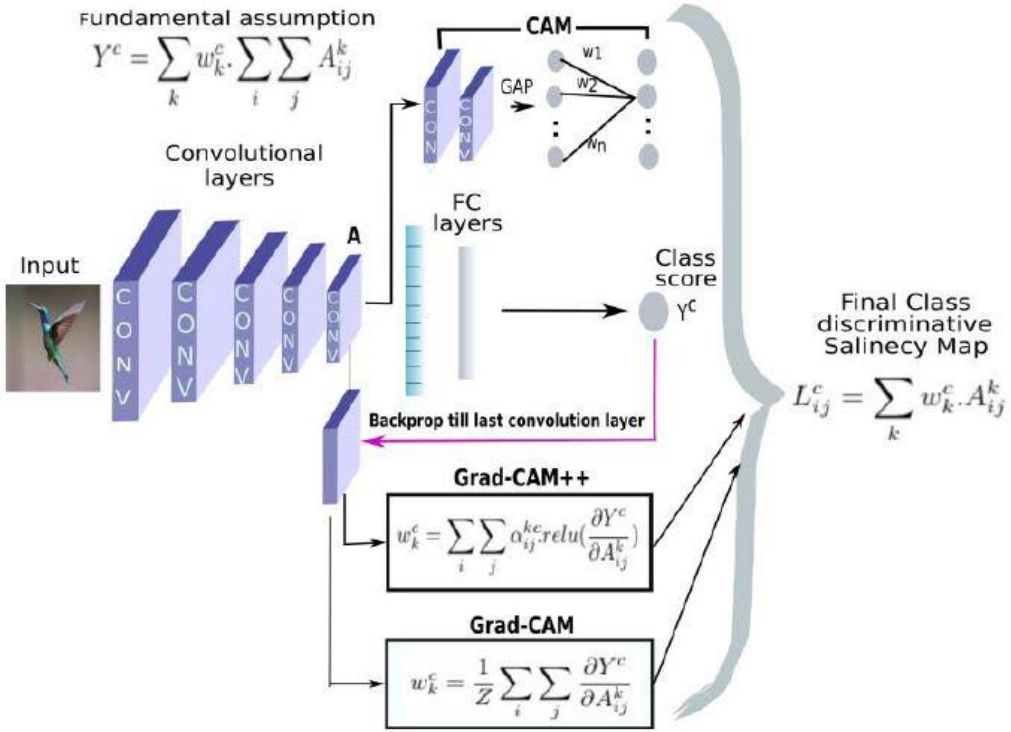
Grad-CAM++: Intuition



Grad-CAM++: Performance



Overview of discussed methods:



Picture Credits: *A. Chattopadhay et al "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," WACV,2018*

LayerCAM:Motivation

- Although Grad-CAM and Grad-CAM++ can generate reliable class activation maps from the final convolutional layer, the located object regions are often small and coarse.

- We hope to find more fine-grained localization information to remedy the class activation maps.

- Can we use the other layers of CNN where there is more spatial information?

LayerCAM:

- It rethinks the relationship between feature map and their corresponding gradients.
- It can produce reliable class activation maps for different layers of CNN instead of only the final convolutional layer.
- This property enables us to collect object localization information from coarse (rough spatial localization) to fine (precise fine-grained details) levels.

Credits: *P. -T. Jiang et al, "LayerCAM: Exploring Hierarchical Class Activation Maps for Localization," in IEEE Transactions on Image Processing, vol. 30, 2021.*

LayerCAM:

Q. Why can't we apply Grad-CAM or Grad-CAM++ to shallow convolutional layers?

- At the last stage, the variances* corresponding to the most feature maps tends to zero.
- Therefore, the global weights used by Grad-CAM and Grad-CAM++ can represent the importance of each spatial location in the feature map.
- However, at the shallow layers, the variances corresponding to most feature maps are very large.
- The global weight cannot represent the importance of different locations in the feature maps on the target category.

*For Grad-CAM, we compute the variance of gradients g^{kc}. And for Grad-CAM++, we compute the variance of $\alpha^{kc} * RELU(g^{kc})$

LayerCAM:

- Instead of a global weight, we generate a separate weight for each spatial location in a feature map.
- Formally, the weight of the spatial location (i, j) in the k-th feature map can be written as

$$w_{ij}^{kc} = relu(g_{ij}^{kc}). \quad \text{where,} \quad g_{ij}^{kc} = \frac{\partial y^c}{\partial A_{ij}^k}$$

- To obtain the class activation map for a certain layer, Layer-CAM first multiplies the activation value of each location in the feature map by a weight:

$$\hat{A}_{ij}^k = w_{ij}^{kc} \cdot A_{ij}^k.$$

LayerCAM:

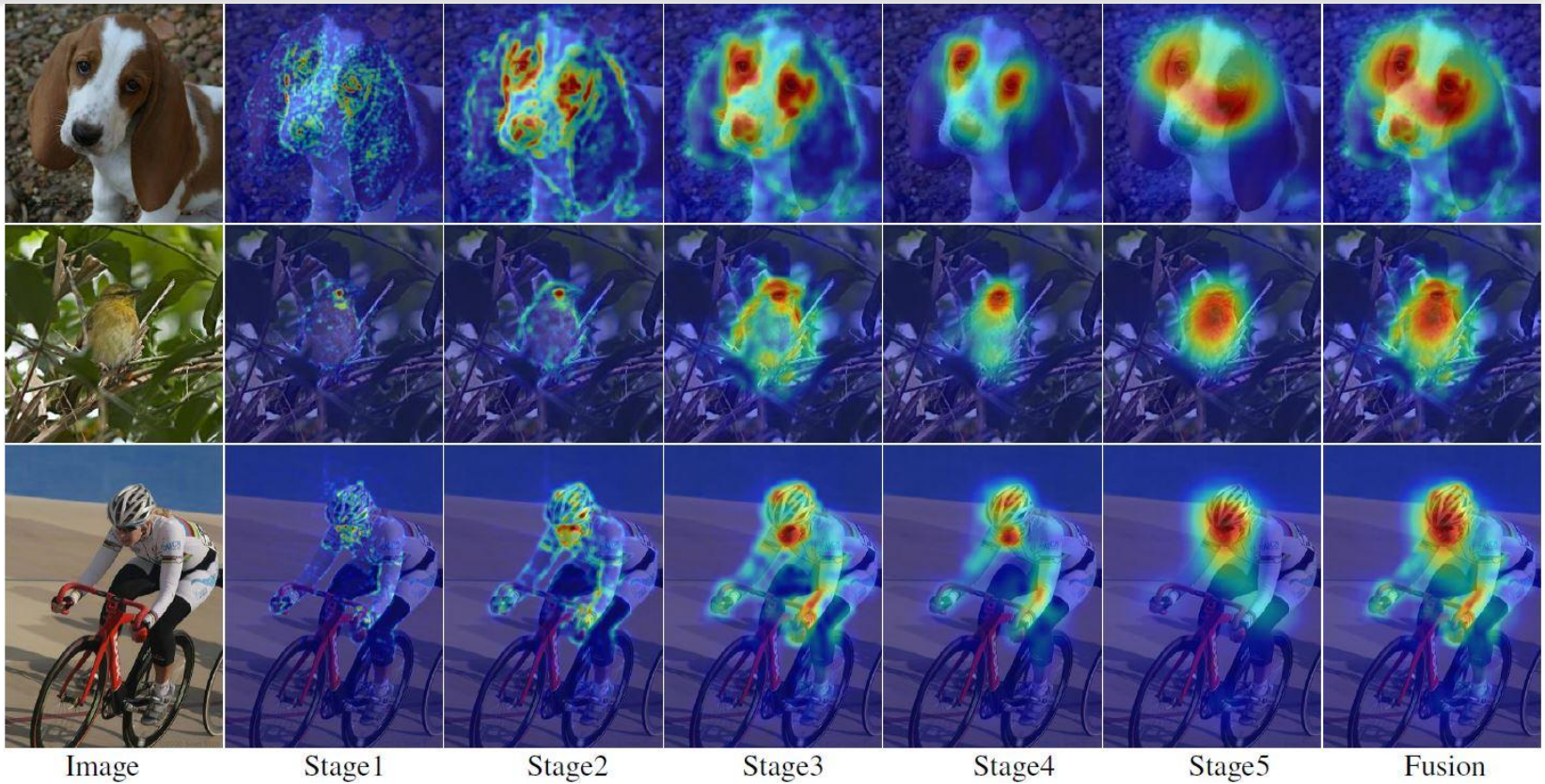
- Finally, the results $\hat{A_k}$ are linearly combined along the channel dimension to obtain the class activation map, which is formulated as follows:

$$M^c = \text{ReLU}\left(\sum_k \hat{A}^k\right)$$

- The class activation maps generated from the shallow layers can capture reliable fine grained object localization information.

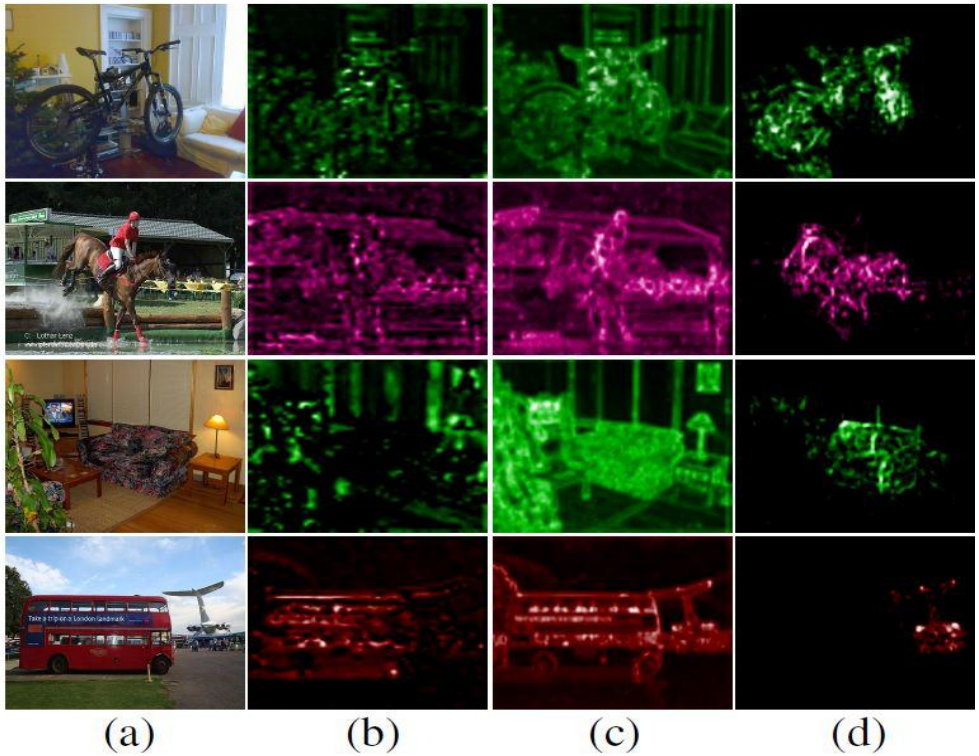
- The separate weight for each location can reflect the importance of different locations in the feature maps related to the target categories.

LayerCAM:



Credits: *P. -T. Jiang et al, "LayerCAM: Exploring Hierarchical Class Activation Maps for Localization," in IEEE Transactions on Image Processing, vol. 30, 2021.*

LayerCAM:



Our LayerCAM that assigns a separate weight for each location in the spatial dimension can consider the different importance to the class of interest, which can keep reliable object localization information while removing the background noise.

Visualization of class activation maps. (a) Source images. (b-d) Class activation maps from the 'pool2' layer of VGG16 by Grad-CAM, Grad-CAM++ and LayerCAM.

Credits: *P. -T. Jiang et al, "LayerCAM: Exploring Hierarchical Class Activation Maps for Localization," in IEEE Transactions on Image Processing, vol. 30, 2021.*

LayerCAM: Experimental Analysis

1. To verify the localization ability of LayerCAM, weakly-supervised object localization experiment is conducted.

2. To test the reliability of the general localization ability of CAMs from the final convolutional layer, the image occlusion experiment is conducted.

3. To analyze the fine grained localization from shallow layers, the surface defect detection experiment is conducted.

4. Finally, we demonstrate that the combination of class activation maps from different stages is beneficial to the weakly-supervised semantic segmentation.

LayerCAM: Experimental Analysis

1.Weakly-supervised object localization experiment:

Dataset: ILSVRC validation set that has 50000 images.

Metrics: The localization accuracy is measured by the loc1 and loc5 metrics.

Results:

COMPARISON OF THE LOCALIZATION ACCURACY OF THE CLASS ACTIVATION MAPS FROM DIFFERENT STAGES. THE 'S' IN THE FIRST ROW DENOTES 'STAGE' IN VGG16. **S5-S1** DENOTES THE LAST CONVOLUTIONAL LAYER OF EACH STAGE IN VGG16.

Method	Metric (%)	S5	S4	S3	S2	S1
Grad-CAM	*loc1*	43.62	18.32	8.87	19.59	13.95
	loc5	53.99	22.70	11.05	23.85	17.27
Grad-CAM++	*loc1*	45.44	41.11	35.33	31.70	31.32
	loc5	56.42	50.97	43.86	39.40	38.90
LayerCAM	*loc1*	**46.62**	**44.05**	**41.83**	**43.18**	**43.71**
	loc5	**57.83**	**55.02**	**52.28**	**53.60**	**54.34**

LayerCAM: Experimental Analysis

2. Image Occlusion experiment:

Dataset: ILSVRC validation set that has 50000 images.

Procedure:
1. Select out the images that are predicted correctly by VGG16.
2. For these truly predicted images, we occlude them with the threshold of 0.7 and then input them to the network.

Results:

COMPARISON OF THE CLASSIFICATION ACCURACY ON THE IMAGE OCCLUSION EXPERIMENT. **CONFIDENCE**: DENOTES THE AVERAGE PREDICTED SCORES OF THE GROUND-TRUTH CATEGORY. LOWER IS BETTER.

Method	original	Grad-CAM	Grad-CAM++	LayerCAM
Top-1 Acc (%)	68.74	50.36	50.07	**48.26**
Top-5 Acc (%)	88.57	75.62	75.26	**73.43**
Confidence (%)	68.64	50.24	49.99	**48.12**

LayerCAM achieves a lower classification accuracy than Grad-CAM and Grad-CAM++, demonstrating that the CAMs generated by LayerCAM from the final convolutional layer can discover more important spatial object regions for the target category

LayerCAM: Experimental Analysis

3. Industry surface defect localization experiment:

Dataset: The DAGM-2007 defect dataset, which contains 3550 training images and 400 test images.

Architecture: We apply our LayerCAM, Grad-CAM, and Grad-CAM++ to layer3 of ResNet-50 to generate class activation maps.

Metric: mean Intersection over union(m-IoU).

Results:

COMPARISON OF DIFFERENT METHODS. THE SEGNET AND REFINENET ARE THE FULLY-SUPERVISED METHODS, WHILE THE OTHERS ARE WEAKLY-SUPERVISED METHODS.

Methods	mIoU (%)	FPS
SegNet	21.95*	17.92*
RefineNet	**32.90***	31.05*
Grad-CAM	0.35	60.97
Grad-CAM++	6.46	60.24
LayerCAM	**27.26**	60.61

LayerCAM: Experimental Analysis

4. Weakly supervised semantic segmentation:

To further test the quality of our class activation maps, we apply them to the WSSS tasks that needs more pixel accurate information.

Dataset: PASCAL VOC 2012.

Metric: mean Intersection over union(m-IoU)

Results:

WEAKLY-SUPERVISED SEGMENTATION RESULTS ON THE PASCAL VOC DATASET. 'WEAK' MEANS THE APPROACHES WITH ONLY IMAGE-LEVEL SUPERVISION. FOR A FAIR COMPARISON, OUR APPROACH IS ALSO BASED ON THE DEEPLAB-LARGEFOV SEGMENTATION MODEL.

Methods	val (%)	test (%)
Grad-CAM	55.6	56.3
Grad-CAM++	55.5	56.1
LayerCAM (Ours, VGG16)	**60.8**	**61.4**
LayerCAM (Ours, ResNet101)	**63.0**	**64.5**

COMPARISONS OF THE MIOU SCORES ON THE PASCAL VOC VALIDATION SET WHEN COMBINING CLASS ACTIVATION MAPS FROM DIFFERENT STAGES.

S5	S4	S3	S2	S1	mIoU (%)
✔					55.6
	✔				55.0
		✔			50.8
			✔		50.5
				✔	46.0
✔	✔				57.1
✔	✔	✔			60.4
✔	✔	✔	✔		**60.8**
✔	✔	✔	✔	✔	60.2

LayerCAM: Conclusion

- In this paper, we propose an attention method, LayerCAM, which can generate reliable class activation maps from different layers of the CNN effectively.
- The class activation maps from deep layers can locate the general location of objects, and the maps from shallow layers can generate fine-grained object localization information.
- The combination of class activation maps from different layers can find more object locations, which is beneficial for improving the performance of the weakly-supervised tasks.
- Experiments show that LayerCAM has better object localization ability than current attention methods.

Future work

- Implementation of the LayerCAM and compare the results with the existing methods.

References:

1. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in IEEE Conf. Comput. Vis. Pattern Recog., 2016
2. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” Int. J. Comput. Vis., vol. 128, no. 2, pp. 336–359, 2020.
3. A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks,” in IEEE Winter Conf. Appl. Comput. Vis., 2018, pp. 839–847
4. J. Choe and H. Shim, “Attention-based dropout layer for weakly supervised object localization,” in IEEE Conf. Comput. Vis. Pattern Recog., 2019, pp. 2219–2228.
5. W. Wang, J. Shen, and H. Ling, “A deep network solution for attention and aesthetics aware photo cropping,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 7, pp. 1531–1544, 2018.
6. W. Wang, J. Shen, X. Lu, S. C. Hoi, and H. Ling, “Paying attention to video object pattern understanding,” IEEE Trans. Pattern Anal. Mach. Intell., 2020.
7. H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, “Score-cam: Score-weighted visual explanations for convolutional neural networks,” in IEEE Conf. Comput. Vis. Pattern Recog. Worksh., 2020, pp. 24–25.

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Thank You!

Appendix:

- The loc1 metric denotes that the estimated result falls into the correct category if the intersection over union (IoU) between the estimated bounding box and the ground-truth bounding box is greater than or equal to 0.5 and meanwhile the top 1 predicted class is correct. The loc5 metric is for the top 5 predicted categories