IEEE CVPR WORKSHOP ON
FAIR, DATA EFFICIENT AND TRUSTED COMPUTER VISION

Score-CAM:

Score-Weighted Visual Explanations for Convolutional Neural Networks

Haofan Wang[1] Zifan Wang[1] Mengnan Du[2] Fan Yang[2] Zijian Zhang[3] Sirui Ding[3] Piotr Mardziel[1] Xia Hu[2]

[1]*Carnegie Mellon University* [2]*Texas A&M University* [3]*Wuhan University*

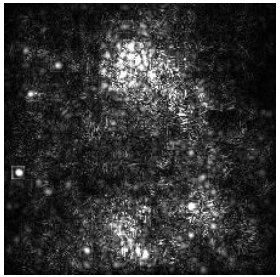
Carnegie Mellon University

TEXAS A&M UNIVERSITY

武漢大學 Wuhan University

Introduction

Common visual explanation methods

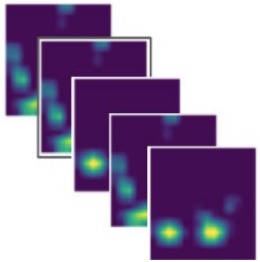


Gradient-based

- Vanilla gradient
- Guided backpropagation
- Integrated gradient
- Smooth gradient
- ...

Perturbation-based

- Randomly sampling
- Monte Carlo sampling
- Optimizing
- ...



CAM-based

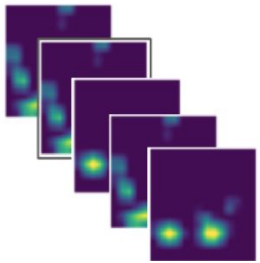
- CAM
- Grad-CAM
- Grad-CAM++
- ...

Introduction

Our contribution

- We propose a new CAM-based visual explanation method, **Score-CAM,** as a solution to existing issues in gradient-based CAMs, e.g. GradCAM.

- Proposed method outperforms several baseline methods under different metrics.



CAM-based

- CAM
- Grad-CAM
- Grad-CAM++
- ...

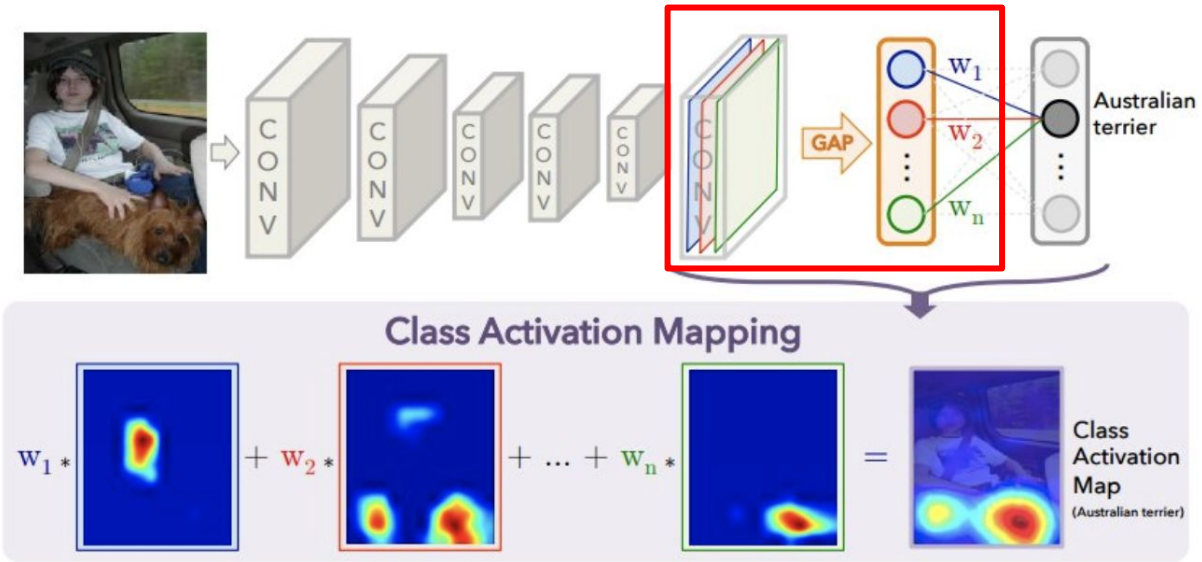
Background

Class Activation Mapping (CAM)

When there is no GAP layer, we need alternative way to compute the weight vector

→ **Gradient-based approximation:**

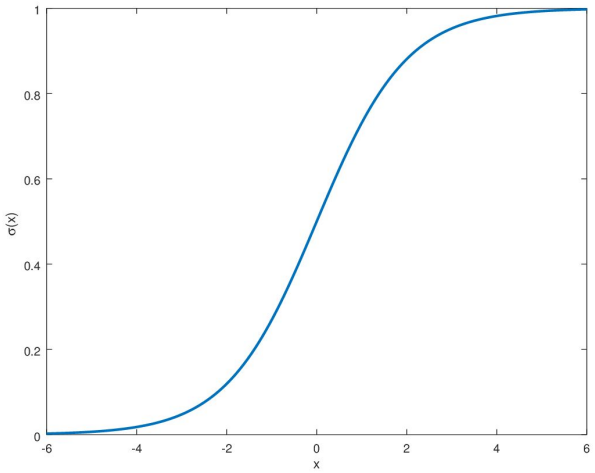
GradCAM,
GradCAM++ …



[Zhou et al. 2016]

Issues of gradient

- Gradient Saturation



- False Confidence

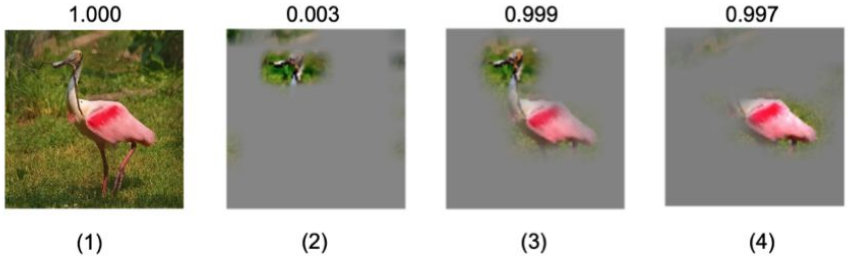


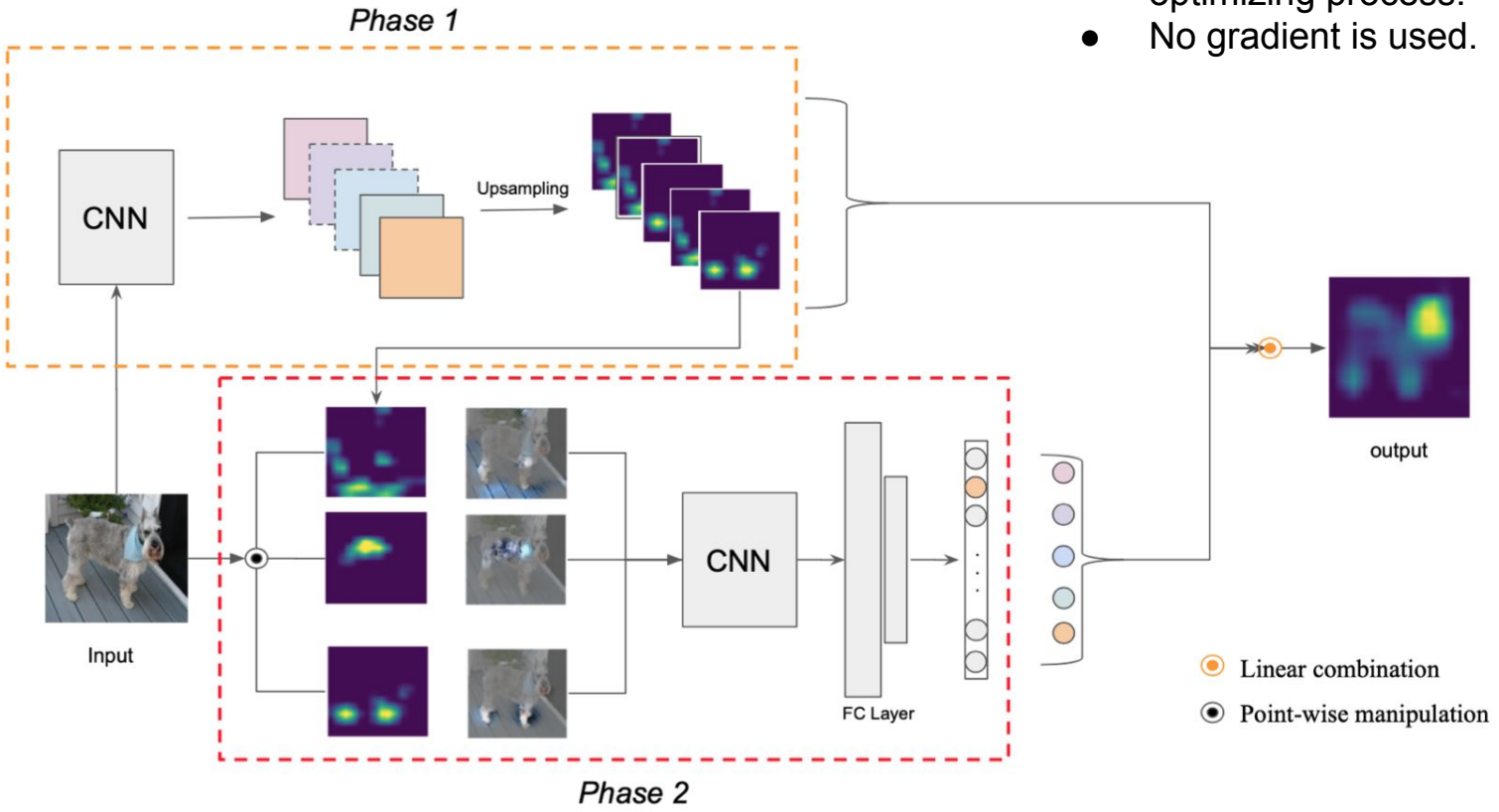
Figure 2. (1) is the input image, (2)-(4) are generated by masking input with upsampled activation maps. The weights for activation maps (2)-(4) are 0.035, 0.027, 0.021 respectively. The values above are the increase on target score given (1)-(4) as input. As shown in this example, (2) has the highest weight but cause less increase on target score.

Motivation

- **Perturbation-based Approximation.**
- The importance of channels (activation maps) are determined by model's response to corresponding perturbations.

Approach: Score-CAM

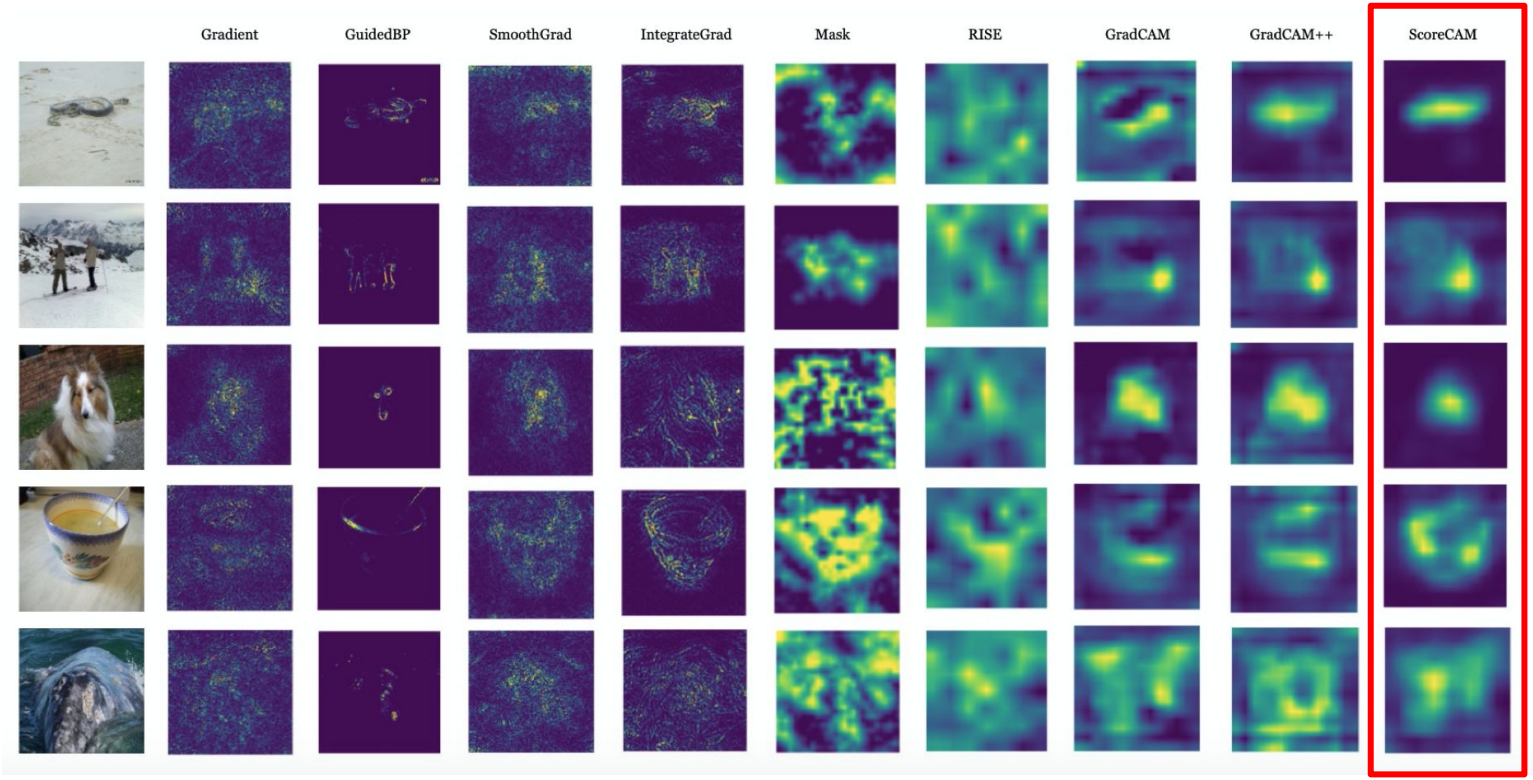
- No re-training process or modification of network architecture (don't require a GAP layer).
- Not require mask sampling or optimizing process.
- No gradient is used.



Experiments

- Qualitative Evaluation via Visualization
- Faithfulness Evaluation via Image Recognition
- Localization Evaluation
- Sanity Check
- Applications

Results: visual comparison



Focused

Less noises

Results: faithfulness evaluation

- **Experiment setting:**
  - Pre-trained **VGG16** model from Pytorch model zoo.
  - **2000** images are randomly selected from **ILSVRC2012** val.
- **Evaluation metrics:**
  - Average Drop: **With only the explanation map region, the average score drop on the target class.**
  - Average Increase: **With only the explanation map region as input,** the percent of samples that have score increase on target class
  - Deletion AUC: **Removing the ordered highlighted region** step by step, the area under the curve of score on target class.
  - Insertion AUC: **Inserting the ordered highlighted region** step by step, the area under the curve of score on target class.
- **Experiment goal:**
  - Quantifying the **relevance** of features highlighted by explanations.

Results: faithfulness evaluation

Table 1. Evaluation results on Recognition (lower is better in Average Drop, higher is better in Average Increase).

Method	Mask	RISE	GradCAM	GradCAM++	ScoreCAM
Average Drop(%)	63.5	47.0	47.8	45.5	**31.5**
Average Increase(%)	5.29	14.0	19.6	18.9	**30.6**

Table 3. Comparative evaluation in terms of deletion (lower is better) and insertion (higher is better) scores.

	Grad-CAM	Grad-CAM++	Score-CAM
Insertion	0.357	0.346	**0.386**
Deletion	0.089	0.082	**0.077**

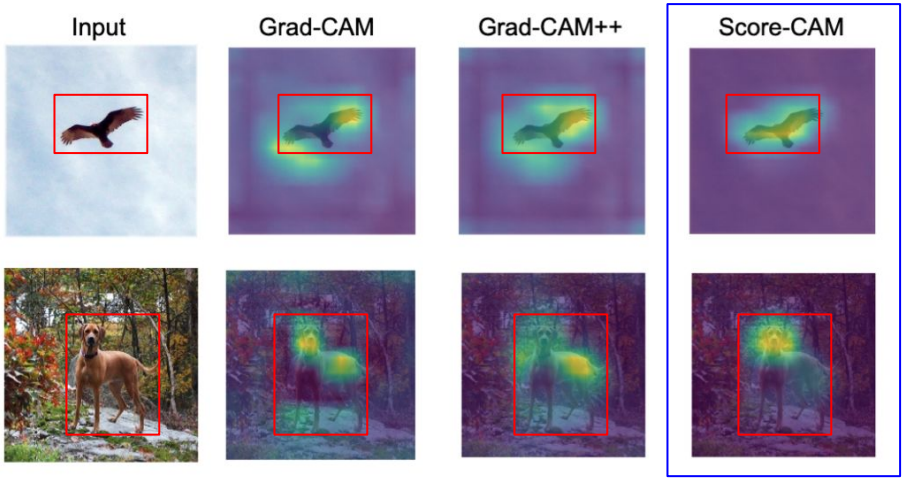
Score-CAM **highlights the most necessary** & **sufficient** features compared with other works, which means that **removing** the region will **cause the largest drop** while **keeping** the region will **bring the highest confidence**.

Results: localization evaluation

- **Experiment setting:**
  - Pre-trained **VGG16** model from Pytorch model zoo.
  - **500** images are randomly selected from **ILSVRC2012** val (**with bbox**).
    - Object occupies **less than 50%** region of the whole image.
    - The object is of only **one bounding box** for simplicity.
- **Evaluation metrics:**
  - Energy-based point game
    - The percent of pixel values that fall into the bounding box.

$$Proportion = \frac{\sum L^c_{(i,j)\in bbox}}{\sum L^c_{(i,j)\in bbox} + \sum L^c_{(i,j)\notin bbox}}$$

Results: localization evaluation



Most of values are inside the bounding box

Table 2. Comparative evaluation on Energy-Based Pointing Game (higher is better).

	Grad	Smooth	Integrated	Mask	RISE	GradCAM	GradCAM++	ScoreCAM
Proportion(%)	41.3	42.4	44.7	56.1	36.3	48.1	49.3	**63.7**

Results: sanity check

Score-CAM **passes** sanity check as previous works, which indicates that it is sensitive to model parameters.

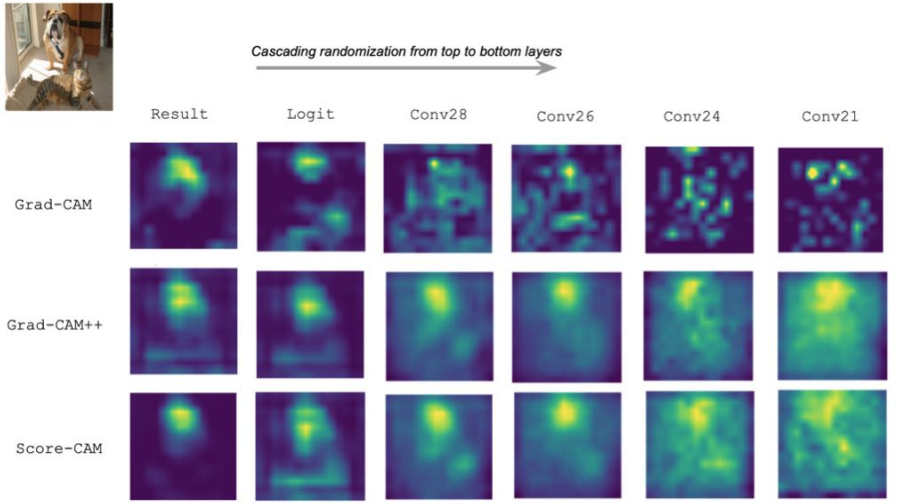


Figure 9. Sanity check results by randomization. The first column is the original generated saliency maps. The following columns are results after randomizing from top the layers respectively. The results show sensitivity to model parameters, the quality of saliency maps can reflect the quality of the model. All three types of CAM pass the sanity check.

Results: applications

- Model convergence

Figure 10. The left is generated by no-finetuning VGG16 with 22.0% classification accuracy , the right is generated by finetuning VGG16 with 90.1% classification accuracy. It shows that the saliency map becomes more focused as the increasing of classification accuracy.

- Dataset bias

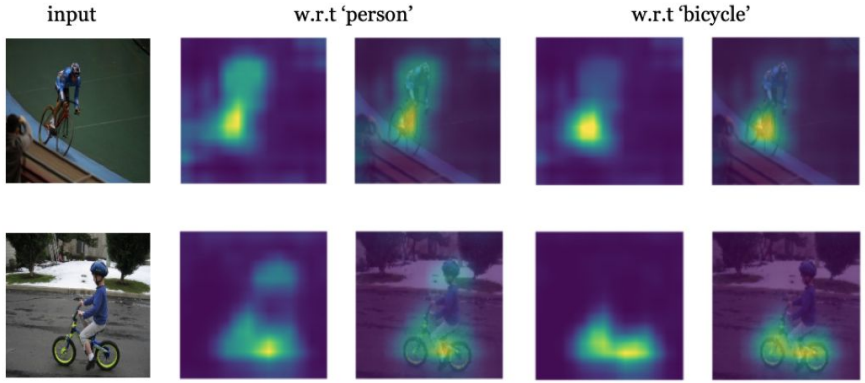


Figure 11. The left column is input example, middle is saliency map w.r.t predicted class (person), right is saliency map w.r.t target class(bicycle).

Thanks for listening!

https://github.com/haofanwang/Score-CAM