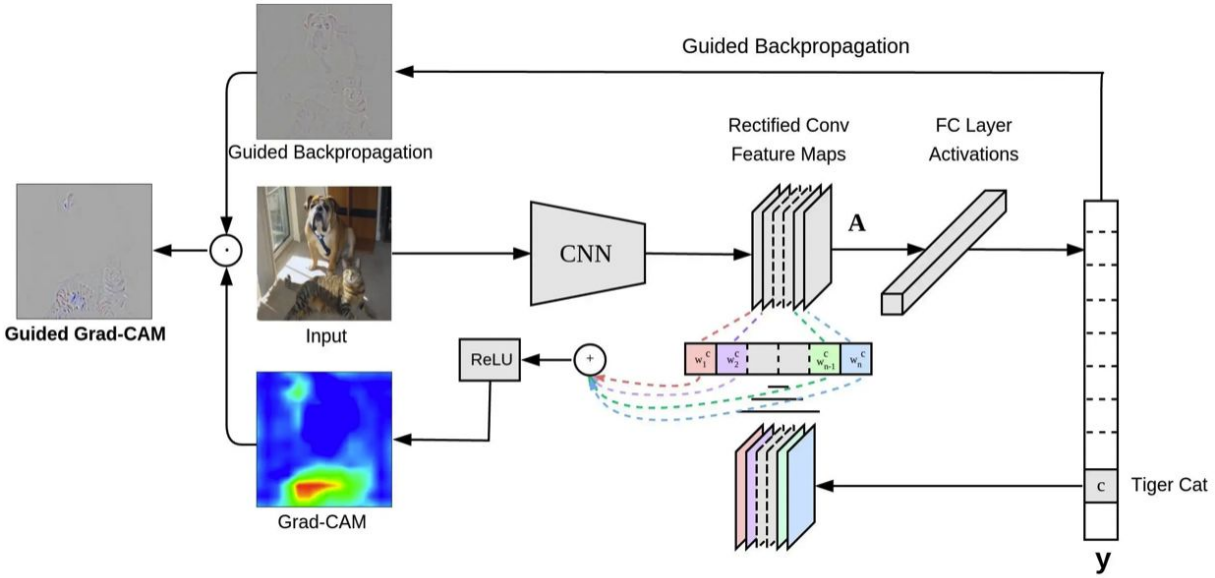
Shap-CAM: Visual Explanations for Convolutional Neural Networks based on Shapley Value *by Zheng et al.*

Karol Pustelnik & Krystian Król

Grad-CAM - overview

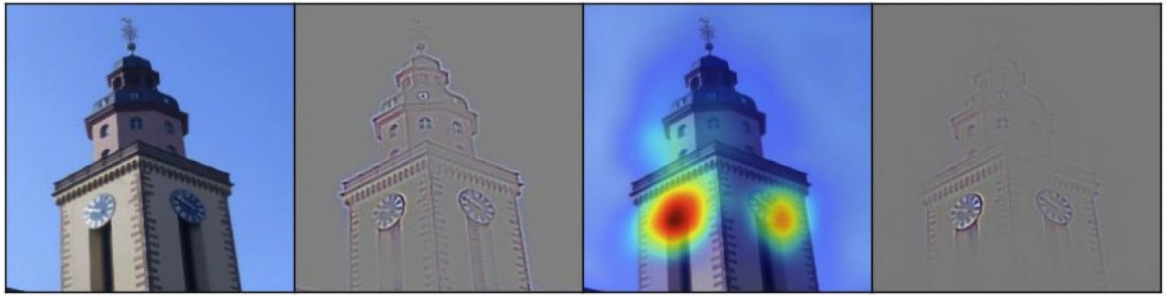


$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

$$L^c_{Grad-CAM} = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right)$$

Grad-CAM - examples

A man is holding a hot dog in his hand



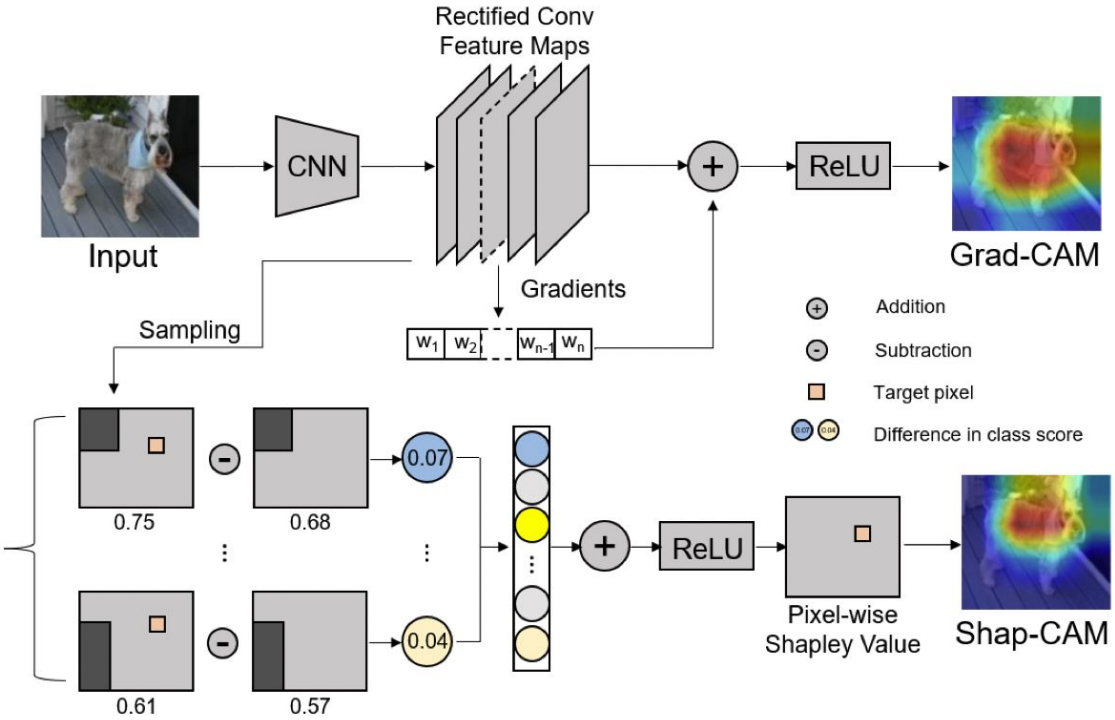
A large clock tower with a clock on the top of it

Shapley value

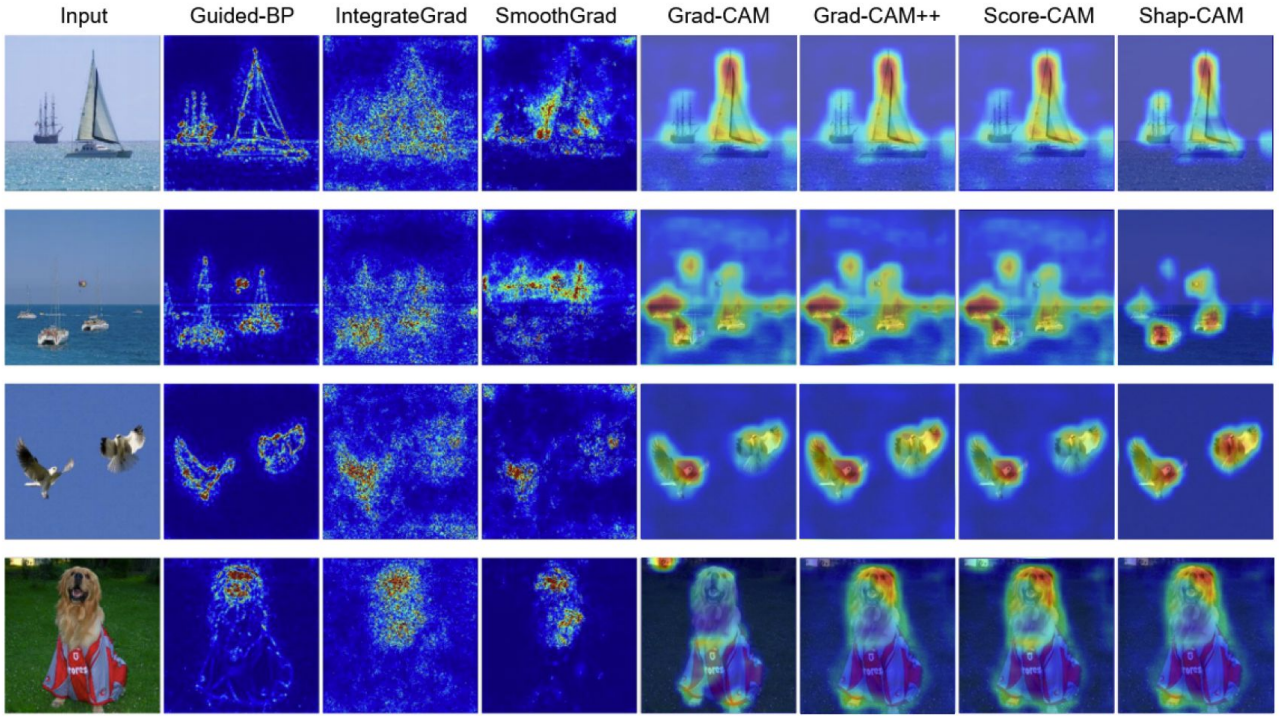
Shapley value is the only way of assigning attributions to players that satisfies the following four properties:

- Null player.
- Symmetry.
- Linearity.
- Efficiency.

Shap-CAM - overview



Shap-CAM - examples



Experiments

Faithfulness evaluation

Faithfulness: An explanation is faithful to the model if it represents the true decision/reasoning process of the model

Average drop: $$\sum_{i=1}^{N} \frac{\max(0, Y_i^c - O_i^c)}{Y_i^c} \cdot 100\%$$

Increase in confidence metric: $$\sum_{i=1}^{N} \frac{sign(Y_i^c < O_i^c)}{N} \times 100\%$$

Table 1. Recognition evaluation results on the ImageNet (ILSVRC2012) validation set (lower is better in Average Drop, higher is better in Average Increase).

Method	Mask	RISE	GradCAM	GradCAM++	ScoreCAM	ShapCAM
Avr. Drop(%)	63.5	47.0	47.8	45.5	31.5	**28.0**
Avr. Increase(%)	5.29	14.0	19.6	18.9	30.6	**31.8**

Table 2. Recognition evaluation results on the PASCAL VOC 2007 validation set (lower is better in Average Drop, higher is better in Average Increase).

Method	Mask	RISE	GradCAM	GradCAM++	ScoreCAM	ShapCAM
Avr. Drop(%)	45.3	31.3	28.5	19.5	15.6	**13.2**
Avr. Increase(%)	10.7	18.2	21.4	19.0	28.9	**32.7**

Localization evaluation

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

Table 3. Localization Evaluations of Proportion (%) using Energy-based Pointing Game (Higher the better).

Method	Grad-CAM	Grad-CAM++	Score-CAM	Shap-CAM
VGG-16	39.95	40.16	40.10	**40.45**
ResNet18	40.90	40.85	40.76	**41.28**

Knowledge Distillation

Table 4. Test error rate (%) for knowledge distillation to train a student from a deeper teacher network. L_{CE} is the normal cross entropy loss function. The Column 2-6 refer to the modified loss function L_{stu} where the explanations for images are generated using the corresponding interpreter.

Loss function	L_{CE}	GradCAM	GradCAM++	ScoreCAM	ShapCAM
w/o L_{KD}	6.78	6.86	6.74	6.75	**6.69**
w/ L_{KD}	5.68	5.80	5.56	5.42	**5.37**

Our opinions / time for questions

- Authors don’t mention the dependency of Shapley-value on seed. Worst case scenario, which we don’t imply, is that they cherry-picked the seed to look good in the paper and tried to hide it in the paper.