CAM-Based Methods Can See through Walls

Magamed Taimeskhanov[1,3], **Ronan Sicre**[2], **Damien Garreau**[3]

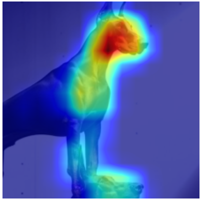
[1]Université Côte d'Azur, Laboratoire J.A. Dieudonné, CNRS, Nice, France

[2]Centrale Méditerranée, Aix-Marseille Univ., CNRS, LIS, Marseille, France

[3]Julius-Maximilians Universität, CAIDAS, Würzburg, Germany

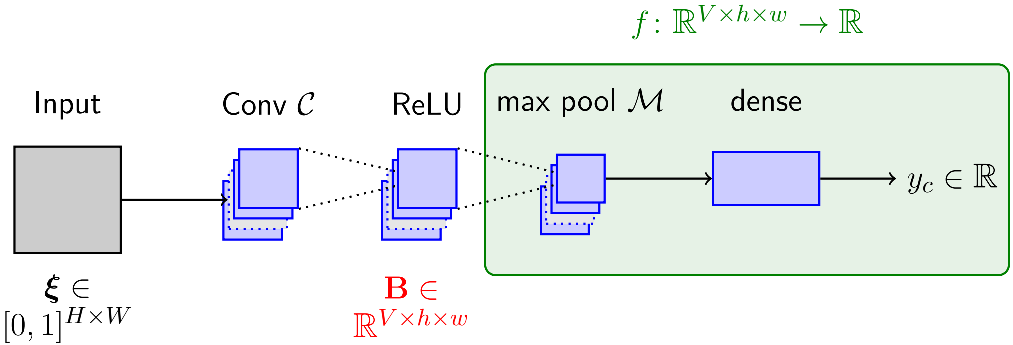
Introduction

- **Setting:** image classification using CNNs
- **Explainable AI** for image classification:
  - identifying key factors influencing predictions
  - post-hoc explanation: saliency maps

$\mathrm{CNN}\Big($  $\Big) \simeq 0.98 \xrightarrow{\text{explanation}}$ 

- **This talk** $=$ pinpointing CAM-based method problem

Simple CNN description



- $\mathbf{B}$ = activation maps
- look at one class score y_c

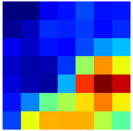
GradCAM in one slide

- **Importance weights α:**

$$\forall i \in [\![V]\!], \qquad \alpha_i := \mathrm{GAP}\left(\nabla_{\mathbf{B}^{(i)}} f\left(\mathbf{B}\right)\right) \in \mathbb{R}\,.$$

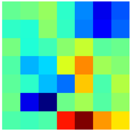
- **Intuition:** influence of a map $\mathbf{B}^{(i)}$ on prediction score
- **GradCAM on previous simple CNN:**

$\mathrm{ReLU}\Big(\alpha_1 \times$



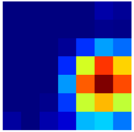
$\mathbf{B}^{(1)}$

$+ \cdots + \alpha_V \times$



$\mathbf{B}^{(V)}$

$\Big) =$



$[\mathbf{GC}]$

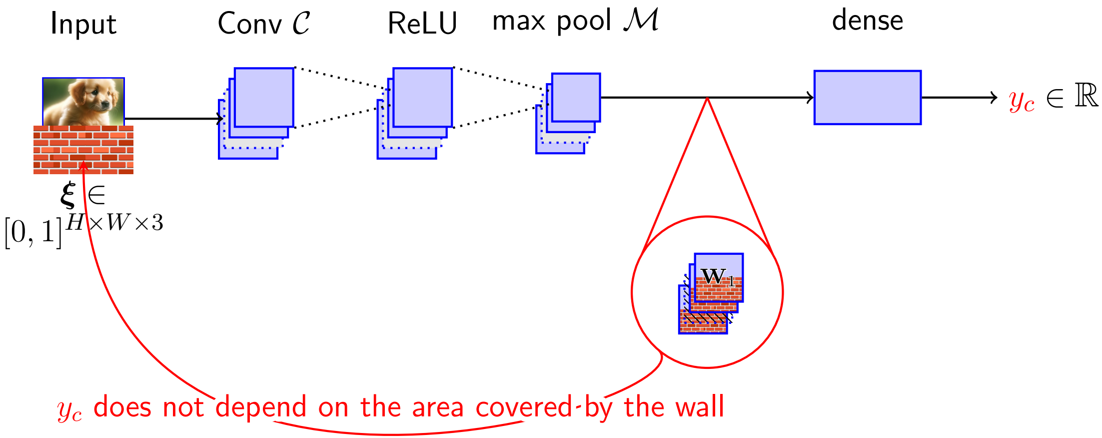
- then **upscale** to input size

Related work

- **Other CAM-based methods:**
  - *Seminal work:* CAM [Zhou et al., 2015]
  - *Extensions:*
    - GradCAM [Selvaraju et al., 2017]
    - GradCAM++ [Chattopadhay et al., 2018]
    - XGradCAM [Fu et al., 2020]
    - ScoreCAM [Wang et al., 2020]
    - AblationCAM [Desai et al., 2020]
    - EigenCAM [Muhammad et al., 2020]
    - HiResCAM [Draelos et al., 2020]
    - Opti-CAM [Zhang et al., 2024]
- **Other limitations of saliency maps:**
  - Adebayo et al., *Sanity Checks for Saliency Maps*, NeurIPS, 2018
  - Ghorbani et al., *Interpretation of Neural Networks Is Fragile*, AAAI, 2019
  - Kindermans et al., *The (Un)reliability of saliency methods*, Springer, 2019

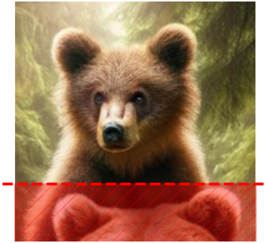
A partially blind model

- CNN with **zeroed out** weights in the first fully-connected layer



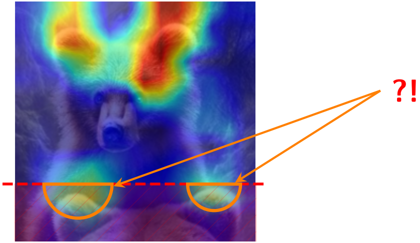
The problem: GradCAM can see through walls

Input image



(a) CNN does not see the red area...

GradCAM map



(b)...but GradCAM highlights inside

Theory on simple CNN

- **Main result:** GradCAM expected behavior

Theorem (Taimeskhanov, Sicre, and Garreau, 2024)

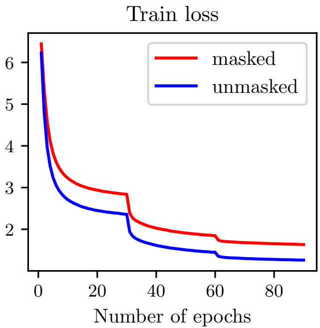
Let $\mathbf{m} := \boldsymbol{\xi}_{i:i+k-1,j:j+k-1}$ *be a patch with* (i,j) *pixel and* k *filter size. Assume that* $\mathbf{W}_{:,-\frac{h'}{2}:,:} = 0$ *and the parameters are* $\mathcal{N}\left(0,\tau^2\right)$ *(i.i.d.). Then,*

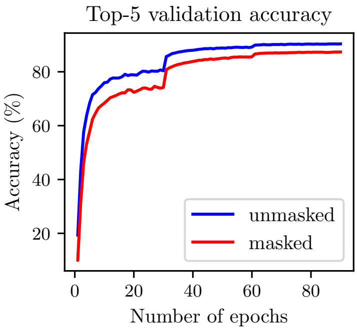
$$\mathbb{E}\left[[\mathbf{GC}]_{i,j}\right] = \mathbb{E}\left[\sigma\left(\sum_{v=1}^{V}\alpha_v \mathbf{B}_{i,j}^{(v)}\right)\right] \geq \frac{V-20}{\sqrt{V}}\sqrt{\frac{h'w'}{16\pi}}\frac{\tau^2}{hw}\left\|\mathbf{m}\right\|_2 .$$

- **Consequence:** GradCAM highlights an image area $\mathbf{m}$ if $\left\|\mathbf{m}\right\|_2 > 0$

Training a VGG16

- **Theory:** does it hold in practice?
- **masked-VGG16** trained to a reasonable accuracy
- **Baseline v.s. masked:** 71.5% v.s. 66.5% top-1



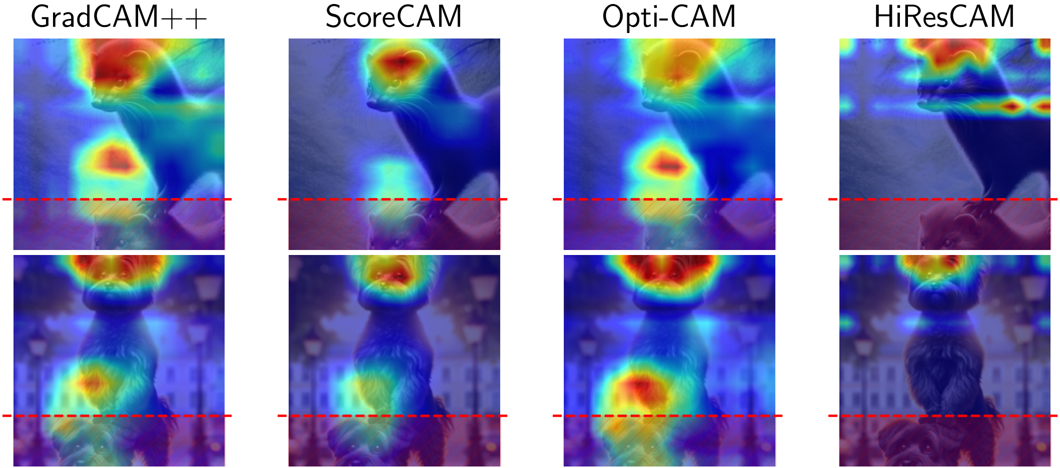


Two new datasets

- **Idea:** one animal at the top, the other at the bottom
- **STACK-MIX:**
  - 100 animal images from ImageNet-1k
  - created by mixing, *à la* cutmix[a]
- **STACK-GEN:**
  - 100 animal images generated by DALL·E 3
  - post-processing

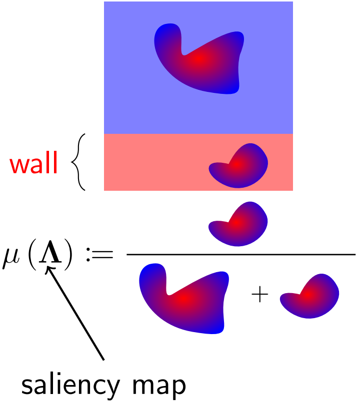
[a] Yun et al., *CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features*, ICCV, 2019

Qualitative results



Quantitative results

- **Metric** (% of ℓ^2-norm):



- **Activation behind the wall for VGG16:**

methods	STACK-MIX ↓	STACK-GEN ↓
GradCAM	22.7 ± 13.4	21.6 ± 11.6
GradCAM++	28.8 ± 8.1	28.5 ± 7.9
XGradCAM	23.8 ± 9.0	22.8 ± 9.0
ScoreCAM	19.9 ± 10.3	18.5 ± 10.6
Opti-CAM	32.7 ± 7.9	32.0 ± 7.8
AblationCAM	21.0 ± 9.9	20.8 ± 9.6
EigenCAM	51.7 ± 19.7	55.8 ± 21.6
HiResCAM	0.0 ± 0.0	0.0 ± 0.0

Conclusion

- **Proceed with caution** when using CAM-based methods
- **Hope:** possible sanity check for saliency maps using
  - our masked CNN
  - datasets STACK-MIX and STACK-GEN
- **Future work:**
  - Extend size of datasets
  - Theory and experiments on other models (ResNet, ...)
  - Check other saliency map methods

Visit our poster (ID 98)!

code and datasets: