A Learning Paradigm for Interpretable Gradients

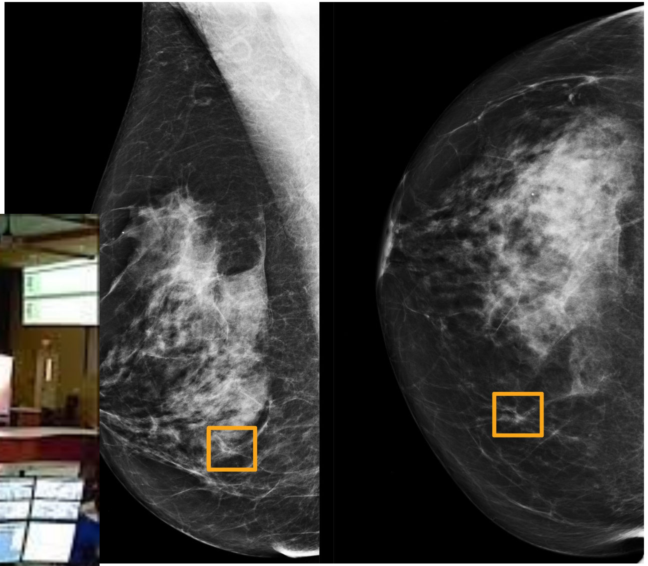
Felipe Torres Figueroa[1], Hanwei Zhang [1], Ronan Sicre[1], Yannis Avrithis [2] and Stephane Ayache[1]

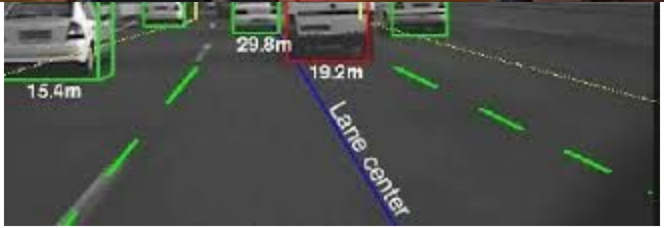
[1] Centrale Marseille, Aix-Marseille Université, CNRS, LIS, Marseille France
[2] Institute of Advanced Research on Artificial Intelligence (IARAI)

Table of Contents

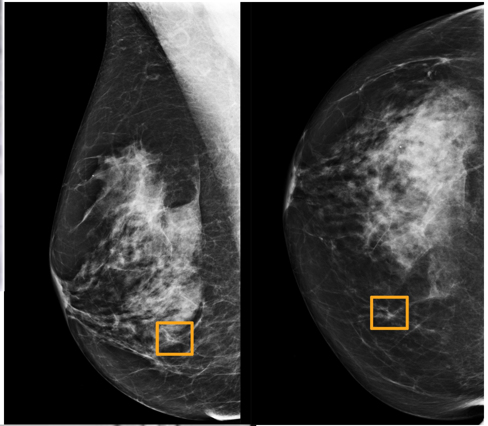
Table of Contents

Motivation: Image Recognition Models Today





Motivation: Explainable AI, What? Why?



What?

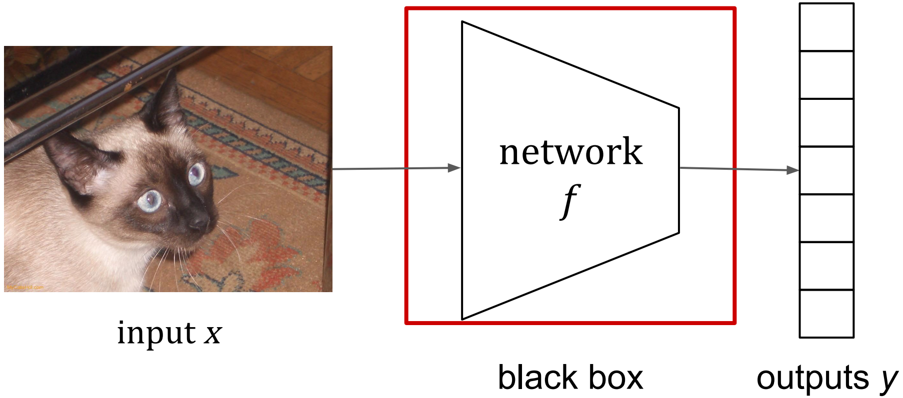
- A way to understand our models?
- A process to uncover the structure of data?
- An approach to improve models?

Why?

- Science vs Real Word.
- Accountability & Responsibility.
- Right of an explanation

Lipton, 2016

Motivation: Explainable AI, How?



Explanations → **Interpretations**

Transparency

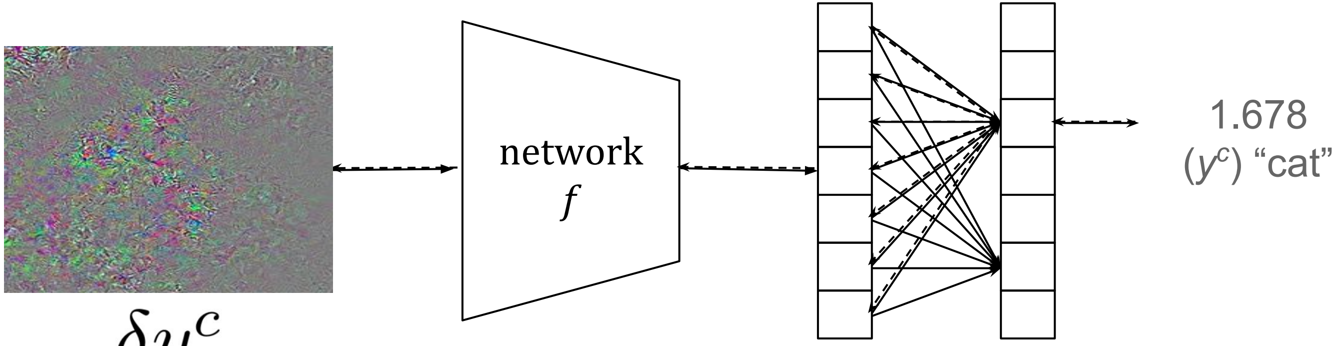
Is a model:
- Decomposable?
- Described in few words?
- Simplificable?

Post-Hoc

Can a model provide?
- Textual explanations?
- Visualizations?
- Explanations by Example?

Table of Contents

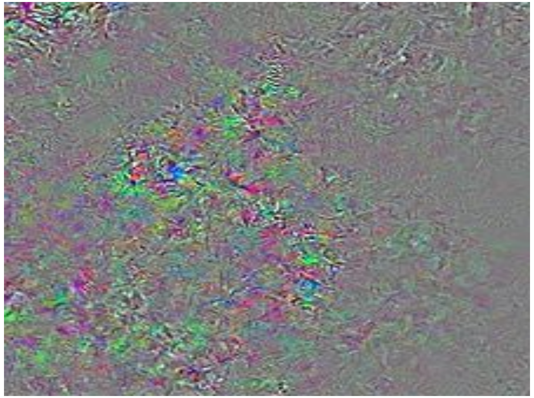
Preliminaries: Backpropagation



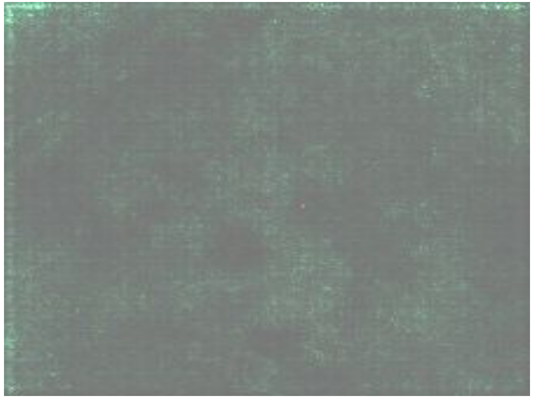
$$\frac{\delta y^c}{\delta x}$$

$$\frac{\delta y^c}{\delta f}$$

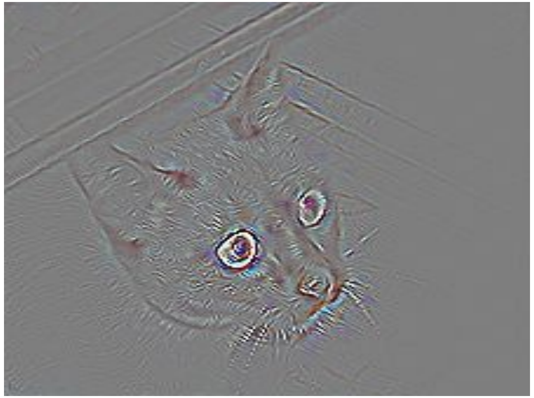
Preliminaries: Gradient



Gradient

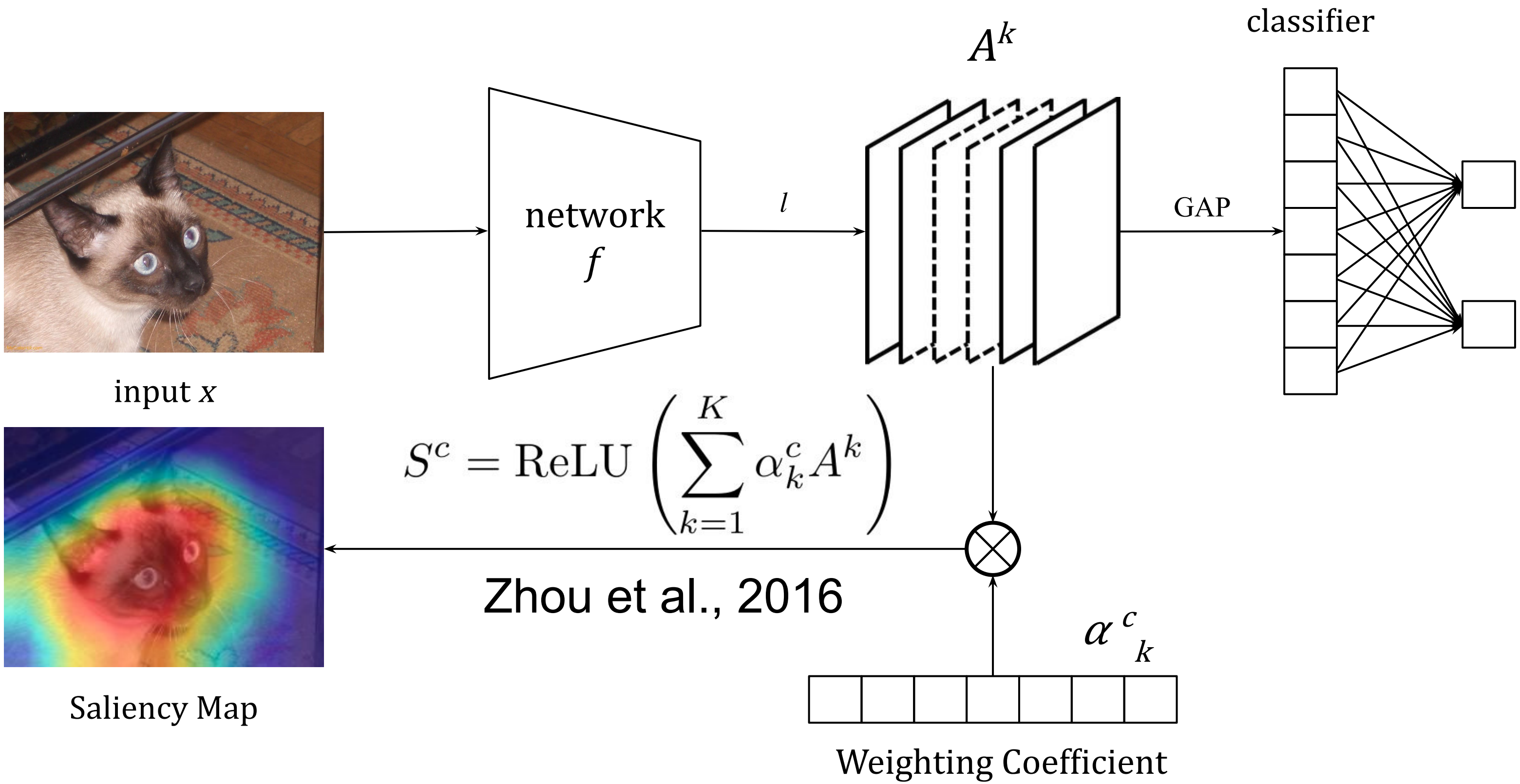


Smooth Gradient
Smilkov et al., 2017



Guided Gradient
Springenberg et al., 2014

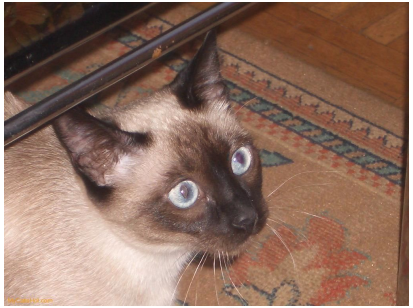
Preliminaries: Class Activation Maps (CAM)



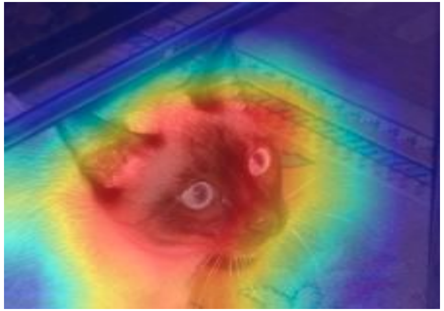
$$S^c = \mathrm{ReLU}\left(\sum_{k=1}^{K} \alpha_k^c A^k\right)$$

Zhou et al., 2016

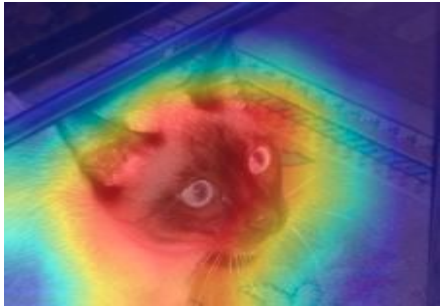
Preliminaries: The many flavours of CAM



$$S^c = \mathrm{ReLU}\left(\sum_{k=1}^{K} \alpha_k^c A^k\right)$$

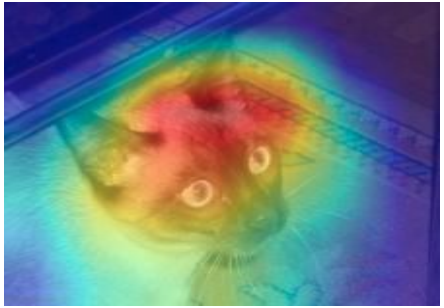


Grad-CAM
Selvaraju et al., 2017



Grad-CAM++
(Chattopadhay et al., 2018)

…



Score-CAM
(Wang et al., 2020)

Table of Contents

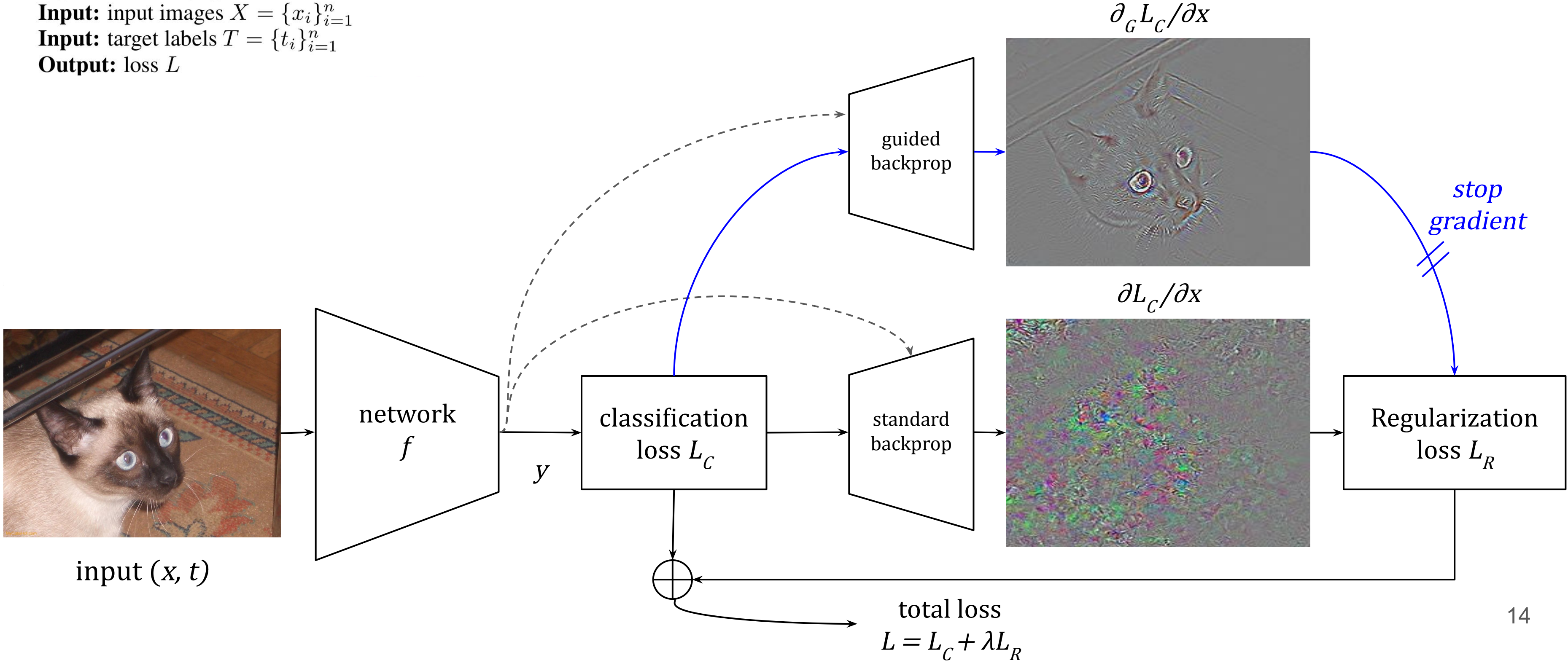
Gradient Denoising: Contributions and Inspiration

- Network responses to inputs can be observed with gradients.
- Guided gradients used to denoise standard gradient.
- Improval of post-hoc interpretations with a transparency inspired approach.

Gradient Denoising : Main algorithm

Algorithm 1: Interpretable gradient loss

Input: network f, parameters θ
Input: input images $X = \{x_i\}_{i=1}^n$
Input: target labels $T = \{t_i\}_{i=1}^n$
Output: loss L



Gradient Denoising : Regularization

- L1 Loss

$$L_{\mathrm{MAE}}(\delta, \delta') = \frac{1}{m}||\delta - \delta'||_1$$

- L2 Loss

$$L_{\mathrm{MSE}}(\delta, \delta') = \frac{1}{m}||\delta - \delta'||_2^2$$

- Cosine Similarity

$$L_{\mathrm{COS}}(\delta, \delta') = \frac{\langle \delta, \delta' \rangle}{||\delta||_2 ||\delta'||_2}$$

- Histogram Intersection

$$L_{\mathrm{HI}}(\delta, \delta') = -\frac{\sum_{i=0}^{m} \min(|\delta_i|, |\delta_i'|)}{||\delta||_1 ||\delta'||_1}$$

Table of Contents

Experiments: Set UP

Training Set UP

- CIFAR-100
- Models:
  - ResNet-18
  - MobileNet-V2
- 200 epochs
- 128 images per batch
- SGD Optimization
- Initial Lr: 10^{-1}, reduced on epochs 60, 120 and 160.

(following https://github.com/weiaicunzai/pytorch-cifar100)

Evaluation Set UP

- Saliency map guided:
  - Generation of CAM activations
  - Evaluation via interpretable recognition
  - Causality Evaluation.

Experiments: Image Recognition

Table 1: *Accuracy* of standard *vs*. our training using ResNet-18 and MobileNet-V2 on CIFAR-100. Using cosine error function for our training.

MODEL	ERROR	λ	ACC
RESNET-18	Baseline	–	73.42
	Ours	7.5×10^{-3}	72.86
MOBILENET-V2	Baseline	–	59.43
	Ours	1×10^{-3}	62.36

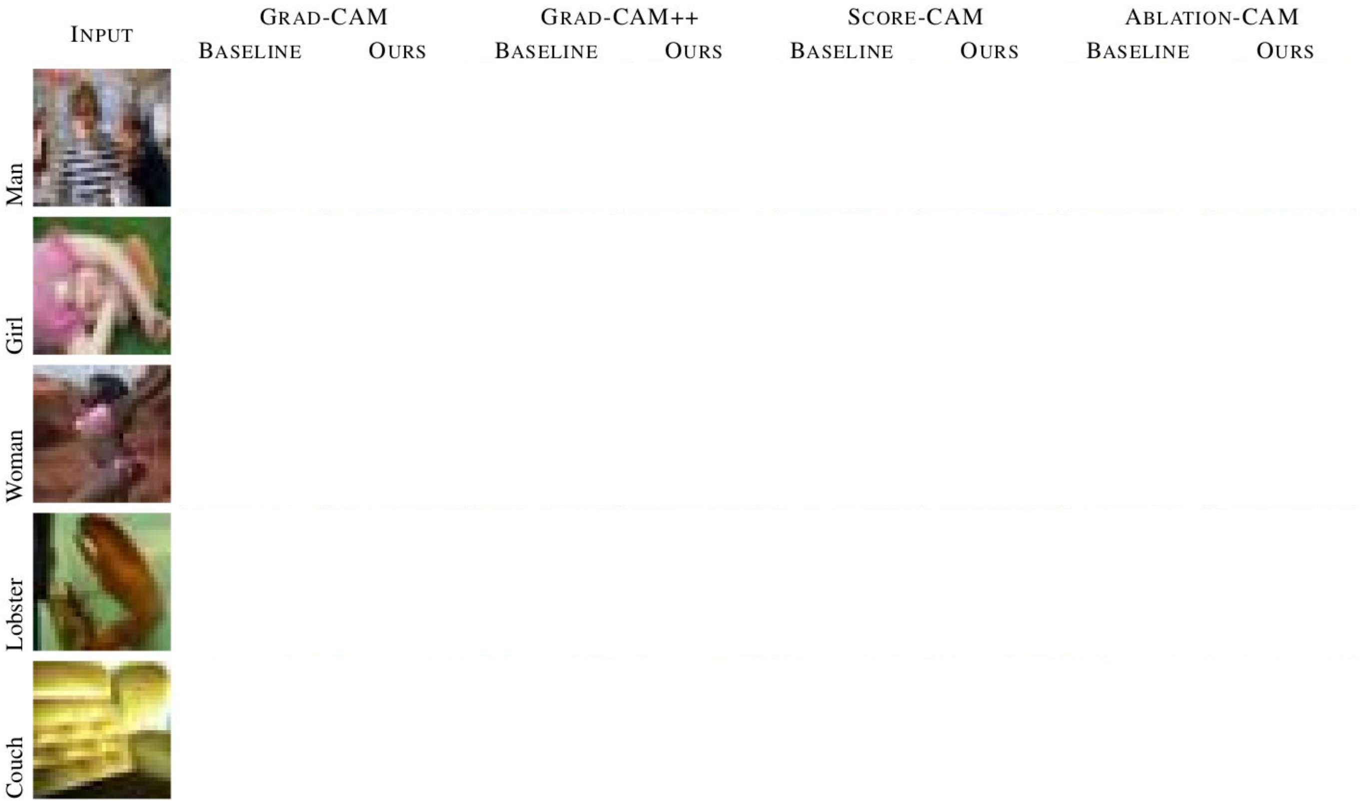
Experiments: Interpretable Image Recognition

Table 2: *Interpretability metrics* of standard *vs*. our training using ResNet-18 and MobileNet-V2 on CIFAR-100. Using cosine error function for our training.

ResNet-18						
Method	Error	AD↓	AG↑	AI↑	Ins↑	Del↓
Grad-CAM	Baseline	30.16	15.23	29.99	58.47	17.47
	Ours	28.09	16.19	31.53	58.76	17.57
Grad-CAM++	Baseline	31.40	14.17	28.47	58.61	17.05
	Ours	29.78	15.07	29.60	58.90	17.22
Score-CAM	Baseline	26.49	18.62	33.84	58.42	18.31
	Ours	24.82	19.49	35.51	59.11	18.34
Ablation-CAM	Baseline	31.96	14.02	28.33	58.36	17.14
	Ours	29.90	15.03	29.61	58.70	17.37
Axiom-CAM	Baseline	30.16	15.23	29.98	58.47	17.47
	Ours	28.09	16.20	31.53	58.76	17.57

MobileNet-V2						
Method	Error	AD↓	AG↑	AI↑	Ins↑	Del↓
Grad-CAM	Baseline	44.64	6.57	25.62	44.64	14.34
	Ours	40.89	7.31	27.08	45.57	15.20
Grad-CAM++	Baseline	45.98	6.12	24.10	44.72	14.76
	Ours	40.76	6.85	26.46	45.51	14.92
Score-CAM	Baseline	40.55	7.85	28.57	45.62	14.52
	Ours	36.34	9.09	30.50	46.35	14.72
Ablation-CAM	Baseline	45.15	6.38	25.32	44.62	15.03
	Ours	41.13	7.03	26.10	45.38	15.12
Axiom-CAM	Baseline	44.65	6.57	25.62	44.64	15.27
	Ours	40.89	7.31	27.08	45.57	15.20

Experiments: Qualitative Results



Experiments: Ablation Experiments

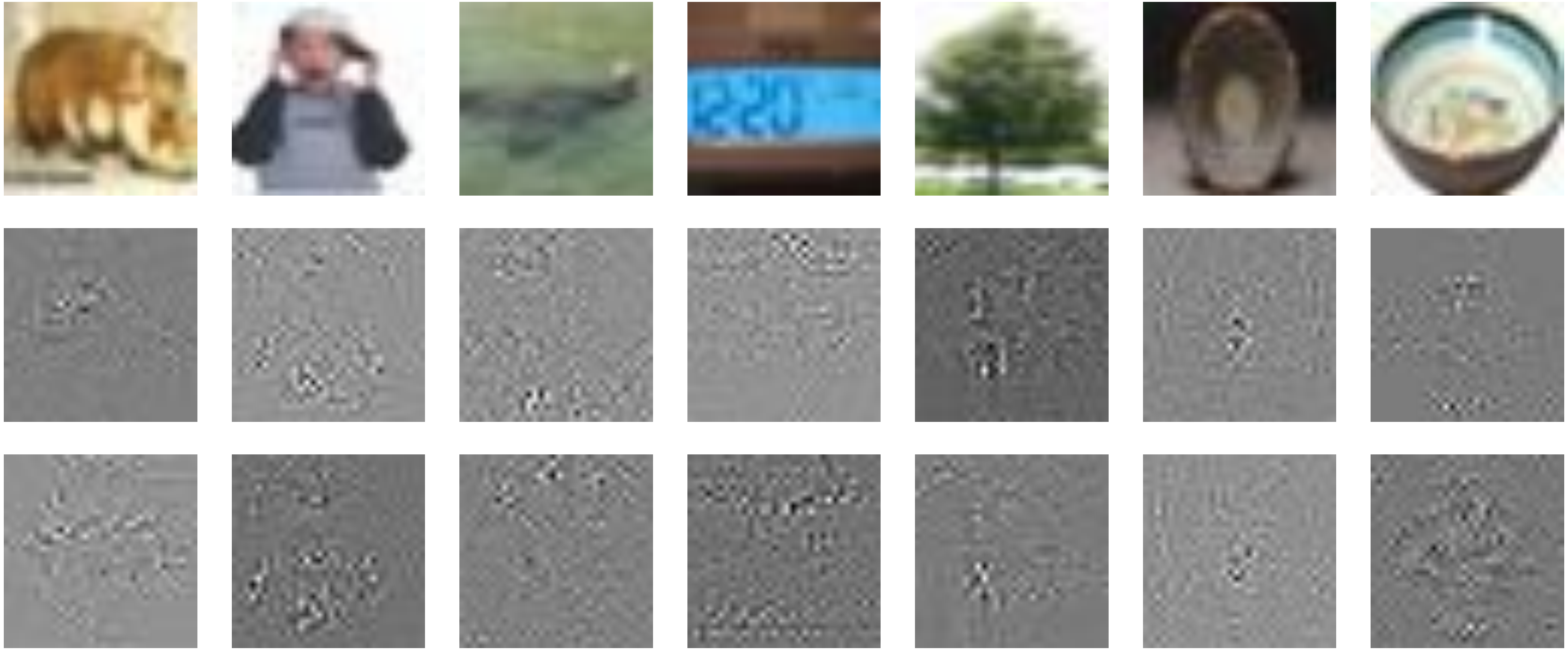
Table 3: Effect of *error function* on our approach, using ResNet-18 and Grad-CAM attributions on CIFAR-100.

Error function	Acc	AD↓	AG↑	AI↑	Ins↑	Del↓
Baseline	73.42	30.16	15.23	29.99	58.47	17.47
Cosine	72.86	**28.09**	**16.19**	**31.53**	58.76	17.57
Histogram	73.88	30.39	14.78	29.38	58.52	**17.35**
MAE	73.41	30.33	15.06	29.61	58.13	17.95
MSE	73.86	29.64	15.19	30.11	**59.05**	18.02

Table 4: Effect of *regularization coefficient* λ (9) on our approach, using ResNet-18 and Grad-CAM attributions on CIFAR-100. Using cosine error function for our training.

λ	Acc	AD↓	AG↑	AI↑	Ins↑	Del↓
0	73.42	30.16	15.23	29.99	58.47	17.47
1×10^{-3}	**73.71**	29.52	15.17	30.03	59.23	**17.45**
2.5×10^{-3}	72.99	30.53	15.82	30.56	59.04	17.96
5×10^{-3}	72.46	30.10	16.06	30.67	57.47	17.80
7.5×10^{-3}	72.86	**28.09**	**16.20**	**31.53**	58.76	17.57
1×10^{-2}	73.28	28.97	15.75	31.16	58.99	17.50
1×10^{-1}	73.00	28.93	16.13	31.55	**59.66**	17.95
1	73.30	28.44	16.02	31.31	58.64	17.48
10	73.04	29.28	15.23	30.47	58.74	17.47

Experiments: Gradient Visualization



Thank you,

Any Questions?

Code will be available soon

Stay in contact!

https://ftorres11.github.io

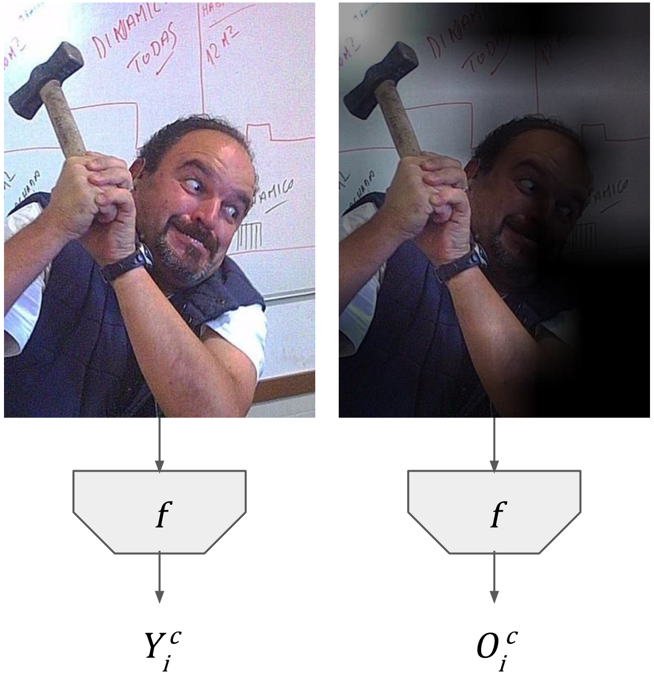
felipe.torres@lis-lab.fr

References

- Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In WACV.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In CVPR.
- Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3):31–57.
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). MobileNetv2: Inverted residuals and linear bottlenecks. In CVPR.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In ICCV.
- Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825.
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In CVPR.

Measuring Interpretability

c: 587 | hammer



$$\mathrm{AD} = \frac{1}{N}\sum_{n=1}^{N}\frac{\max(0, Y_i^c - O_i^c)}{Y_i^c}$$

$$\mathrm{IC} = \frac{1}{N}\sum_{n=1}^{N}\mathbb{1}(O_i^c > Y_i^c)$$

Chattopadhyay et al., 2017

$$\mathrm{AG} = \frac{1}{N}\sum_{n=1}^{N}\frac{\max(0, O_i^c - Y_i^c)}{Y_i^c}$$

Measuring Interpretability

Algorithm 1: Insertion Algorithm

Input: black-box f, image x, saliency map s^c, number of pixels N removed per step.

Output: insertion score *ins*. $n \leftarrow 0$

$x' \leftarrow \mathrm{Blur}(x)$

$p_n^c \leftarrow f(x)$

while $x \neq x'$ **do**

  According to s, set the next n pixels in x' to corresponding pixels in x

  $n \leftarrow n + 1$

  $p_n^c \leftarrow f(x')$

$ins \leftarrow \mathrm{AreaUnderCurve}(p_n^c \mathrm{vs.} i/n, \forall i = 0, \ldots n)$

return *ins*

Measuring Interpretability

Algorithm 2: Deletion Algorithm

Input: black-box f, image x, saliency map s^c, number of pixels N removed per step.

Output: deletion score del.

$n \leftarrow 0$

$p_n^c \leftarrow f(x)$

while x has non-zero pixels **do**

  According to s, set the next n pixels in x to 0

  $n \leftarrow n+1$

  $p_n^c \leftarrow f(x)$

$del \leftarrow \text{AreaUnderCurve}(p_n^c \text{vs.} i/n, \forall i = 0, \ldots n)$

return del