Learning Discriminative Representations to Interpret Image Recognition Models

Supplementary Material

Felipe Torres Figueroa

École Centrale de Marseille

QARMA - Laboratoire d'Informatique et de Systèmes (LIS)

Marseille, September 23rd 2024

Table of Contents

1. Opti-CAM: Optimizing saliency maps for interpretability
2. CA-Stream: Attention-based pooling for interpretable image recognition
3. A learning paradigm for interpretable gradients

Table of Contents

1. Opti-CAM: Optimizing saliency maps for interpretability
2. CA-Stream: Attention-based pooling for interpretable image recognition
3. A learning paradigm for interpretable gradients

Evaluating Interpretability

Interpretable Image Recognition

An explanation should demonstrate similar predictive properties to its query:

Input image. (I)

$P_i^c = 0.8756$

Explanation Map. (E^c)

$O_i^c = 0.7442$

$$\text{AD}(\%) := \frac{1}{N}\sum_{i=1}^{N}\frac{[y_i^c - o_i^c]_+}{y_i^c}\cdot 100 \qquad (1)$$

$$\text{AI}(\%) := \frac{1}{N}\sum_{i}^{N}\mathbb{1}_{y_i^c < o_i^c}\cdot 100 \qquad (2)$$

$$\text{AG}(\%) := \frac{1}{N}\sum_{i=1}^{N}\frac{[o_i^c - y_i^c]_+}{1 - y_i^c}\cdot 100 \qquad (3)$$

Interpretability
Causality Analysis

Saliency guided perturbations revelal the importance of salient regions.

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*. $n \leftarrow 0$
$x' \leftarrow \text{Blur}(x)$
$p_n^c \leftarrow f(x)$
while $x \neq x'$ **do**
  According to s, set the next n pixels in x' to corresponding pixels in x
  $n \leftarrow n + 1$
  $p_n^c \leftarrow f(x')$
$ins \leftarrow \text{AreaUnderCurve}(p_n^c \text{vs.} i/n, \forall i = 0, ...n)$
return *ins*

Interpretability
Causality Analysis

Saliency guided perturbations revelal the importance of salient regions.

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
  $n \leftarrow n + 1$
  $p_n^c \leftarrow f(x)$
$del \leftarrow \text{AreaUnderCurve}(p_n^c \text{vs.} i/n, \forall i = 0, ...n)$
return *del*

Interpretability
Weakly Supervised Object Localization

$$\text{OM} := 1 - \left(\max_{B \in \mathbb{B}} \text{IoU}(B, B_p) \right) \mathbb{1}_{c_p = c}, \quad (4)$$

$$\text{LE} := 1 - \max_{B \in \mathbb{B}} \text{IoU}(B, B_p). \quad (5)$$

$$P := \frac{\sum_{\mathbf{p} \in U} S^c_{\mathbf{p}}}{\sum_{\mathbf{p}} S^c_{\mathbf{p}}} \quad (6)$$

$$R := \frac{\sum_{\mathbf{p} \in U} S^c_{\mathbf{p}}}{|U|}. \quad (7)$$

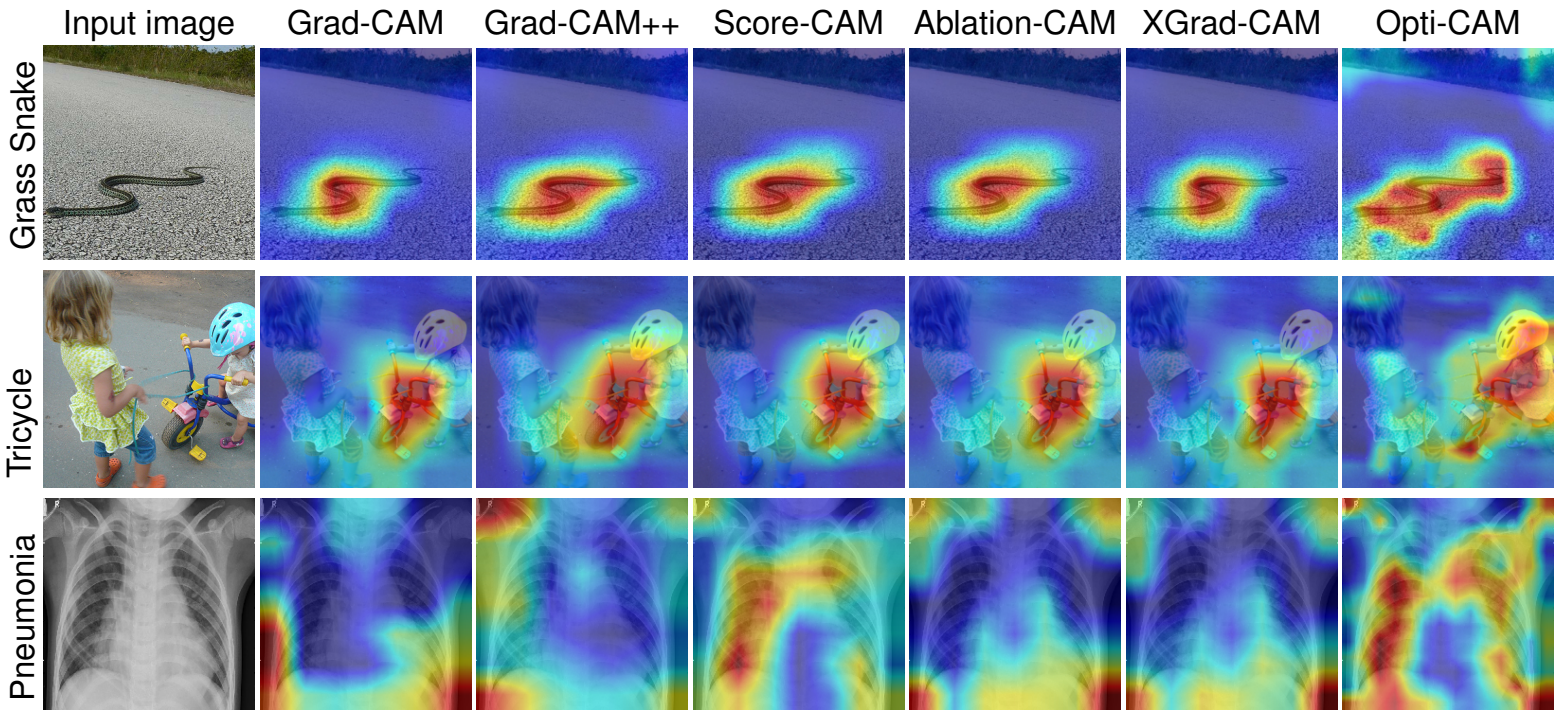
$$\text{BoxAcc}(\eta, \delta) := \max_{B \in \mathbb{B}} \mathbb{1}_{\text{IoU}(B^\eta_p, B) \geq \delta}. \quad (8)$$

$$\text{SP} := \mathbb{1}_{\mathbf{p}^* \in U}. \quad (9)$$

$$\text{SM} := \log \max \left(0.05, \frac{|B_p|}{hw} \right) - \log p^c, \quad (10)$$

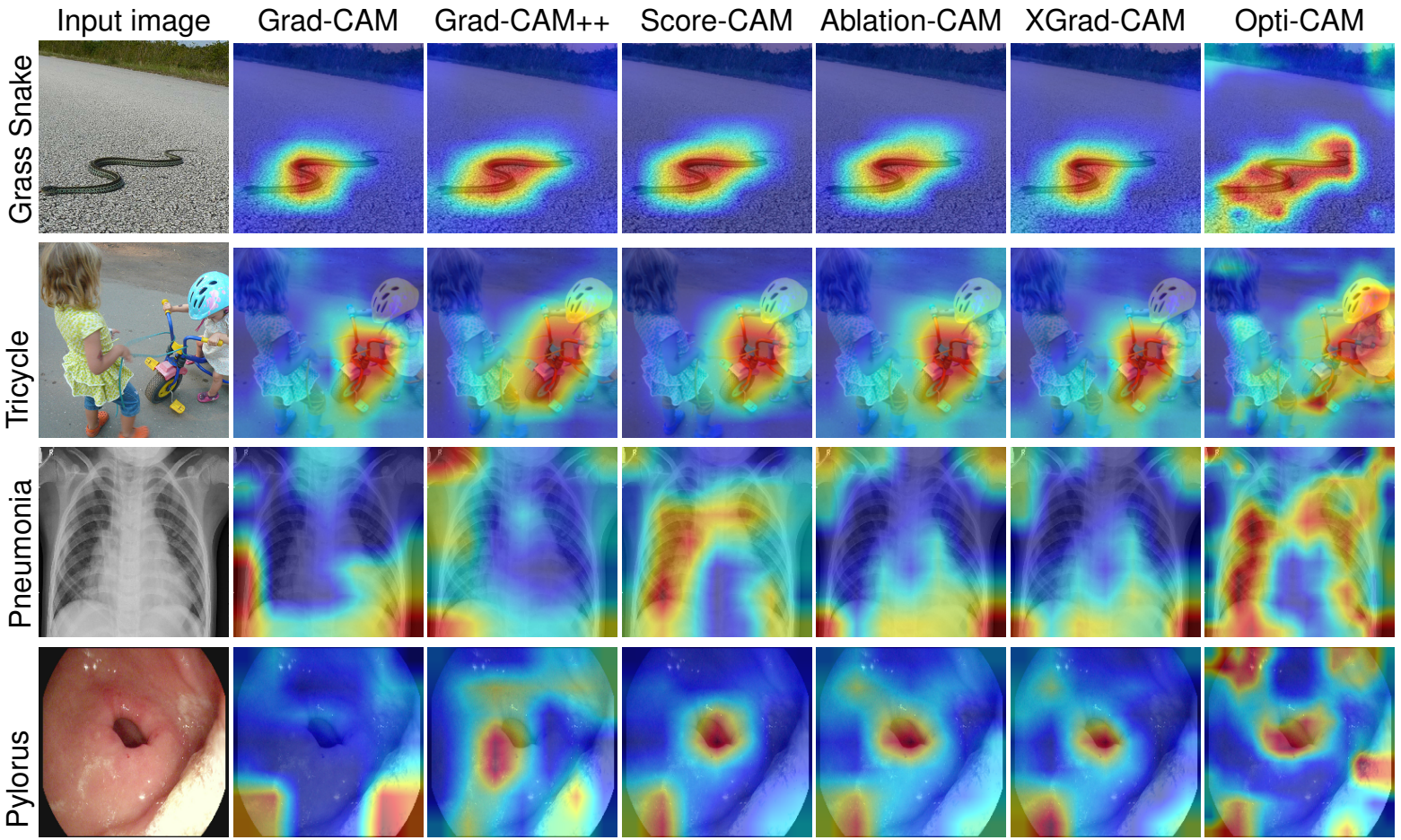
Results

Qualitative Evaluation



Results

Qualitative Evaluation



Results

Quantitative Experiments

Classification Metrics on Transformers:

Method	ViT-B				DeiT-B			
	AD↓	AG↑	AI↑	T(s)	AD↓	AG↑	AI↑	T(s)
Fake-CAM	0.3	0.4	48.3	0.00	0.6	0.3	44.6	0.00
Grad-CAM	69.4	2.5	12.4	0.14	33.5	1.7	12.5	0.11
Grad-CAM	86.3	1.5	1.0	0.15	50.7	0.9	7.2	0.13
Score-CAM	32.0	6.2	33.0	23.69	53.6	2.2	12.2	22.47
XGrad-CAM	88.1	0.4	4.3	0.13	80.5	0.3	4.1	0.12
Layer-CAM	82.0	0.2	2.9	0.24	88.9	0.4	2.6	0.24
ExPerturbation	28.8	6.2	24.4	133.52	60.9	2.0	8.5	129.12
RawAtt	92.6	0.2	2.8	0.02	95.3	0.0	1.8	0.02
Rollout	42.1	5.6	20.9	0.02	55.2	0.8	7.9	0.02
TIBAV	81.7	0.8	5.8	0.16	62.3	0.7	7.1	0.16

Results

Quantitative Experiments

Classification Metrics on Transformers:

Method	ViT-B				DeiT-B			
	AD↓	AG↑	AI↑	T(s)	AD↓	AG↑	AI↑	T(s)
Fake-CAM	0.3	0.4	48.3	0.00	0.6	0.3	44.6	0.00
Grad-CAM	69.4	2.5	12.4	0.14	33.5	1.7	12.5	0.11
Grad-CAM	86.3	1.5	1.0	0.15	50.7	0.9	7.2	0.13
Score-CAM	32.0	6.2	33.0	23.69	53.6	2.2	12.2	22.47
XGrad-CAM	88.1	0.4	4.3	0.13	80.5	0.3	4.1	0.12
Layer-CAM	82.0	0.2	2.9	0.24	88.9	0.4	2.6	0.24
ExPerturbation	28.8	6.2	24.4	133.52	60.9	2.0	8.5	129.12
RawAtt	92.6	0.2	2.8	0.02	95.3	0.0	1.8	0.02
Rollout	42.1	5.6	20.9	0.02	55.2	0.8	7.9	0.02
TIBAV	81.7	0.8	5.8	0.16	62.3	0.7	7.1	0.16
Opti-CAM	**0.6**	**18.0**	**90.1**	16.05	**0.9**	**26.0**	**83.5**	15.17

Results

Quantitative Evaluation

Localization Experiments:

METHOD	ViT-B							DeiT-B						
	OM↓	LE↓	F1↑	BA↑	SP↑	EP↑	SM↓	OM↓	LE↓	F1↑	BA↑	SP↑	EP↑	SM↓
Fake-CAM	62.8	54.0	57.7	47.9	99.8	28.6	0.87	61.4	54.0	57.7	47.9	99.8	28.7	0.83
Grad-CAM	79.6	74.3	29.4	45.0	58.1	31.0	3.27	65.5	60.3	44.3	47.2	62.8	30.2	1.20
Grad-CAM++	84.2	80.6	14.8	23.8	51.4	27.3	4.15	70.6	67.2	34.3	43.6	57.7	30.3	2.14
Score-CAM	77.6	71.6	46.0	54.3	**66.1**	33.1	3.14	79.9	76.2	31.9	43.8	**63.4**	32.2	3.14
XGrad-CAM	82.0	76.9	19.6	41.3	52.8	28.5	3.31	82.0	78.4	19.5	44.1	53.4	28.8	3.03
Layer-CAM	70.7	63.9	20.6	50.5	60.7	32.6	1.44	80.2	77.3	17.6	50.8	62.7	35.1	3.15
ExPerturbation	71.5	64.9	35.9	44.6	62.3	**35.3**	1.34	69.9	64.3	36.2	44.2	63.1	**35.5**	1.16
RawAtt	72.4	64.8	18.5	50.4	55.4	31.6	1.68	73.5	68.2	5.9	**48.1**	46.5	27.3	1.91
Rollout	67.6	58.8	36.9	50.7	57.8	30.0	1.16	63.9	57.0	27.8	47.9	36.5	27.2	0.94
TIBAV	70.1	63.1	26.6	**58.8**	**66.1**	35.0	1.23	68.2	62.2	28.1	59.6	64.1	33.5	1.08
Opti-CAM (ours)	**64.4**	**54.6**	**54.5**	48.0	58.2	28.7	**0.98**	**62.3**	**55.1**	**53.9**	48.0	55.1	28.8	**0.84**

Table of Contents

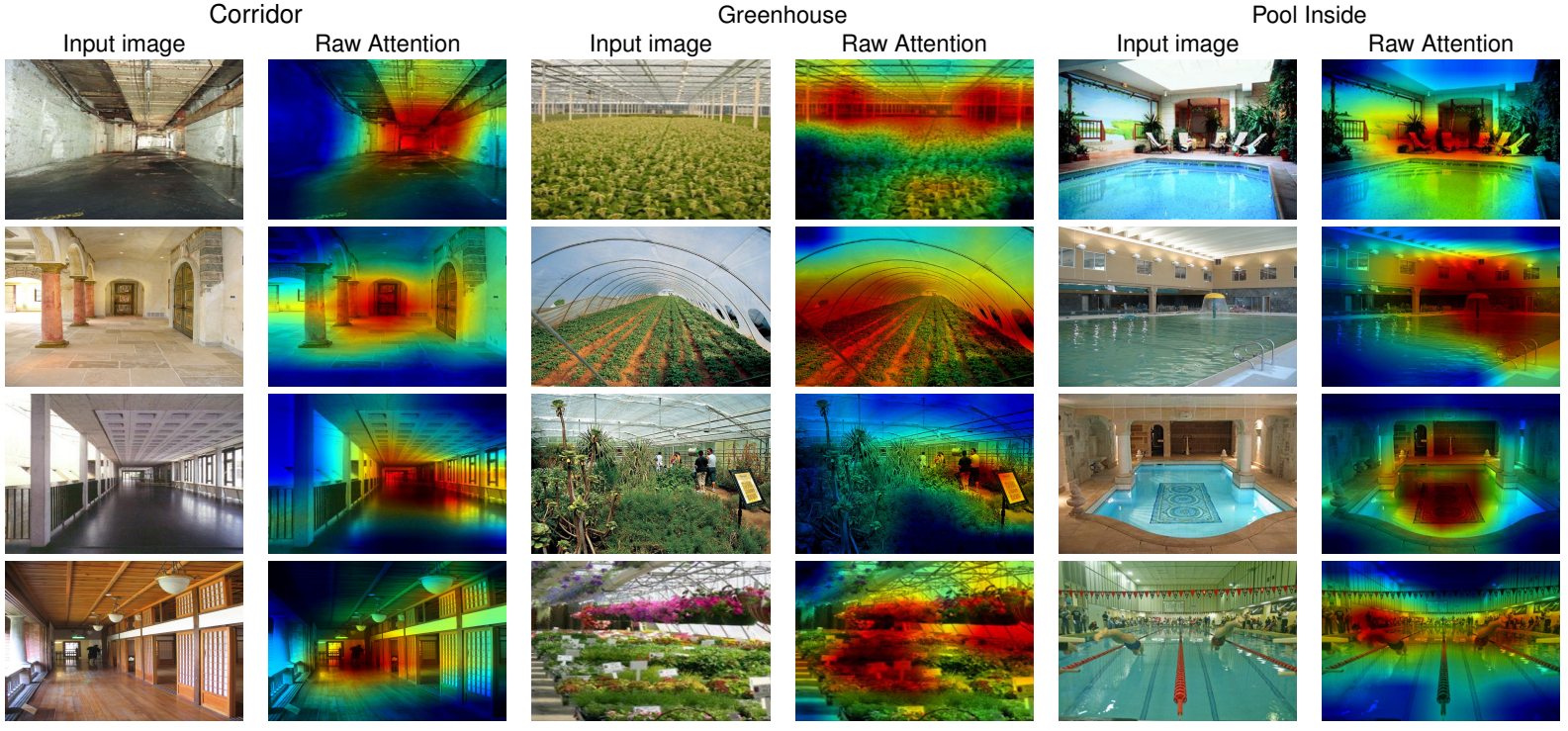
1 Opti-CAM: Optimizing saliency maps for interpretability

2 CA-Stream: Attention-based pooling for interpretable image recognition

3 A learning paradigm for interpretable gradients

Results

Qualitative Experiments



Results

Quantitative Experiments

Network	Method	Pool	AD↓	AG↑	AI↑	I↑	D↓
ResNet-18	Grad-CAM	GAP	17.64	12.73	41.21	63.13	**10.66**
		CA	**16.99**	**17.22**	**44.95**	**65.94**	10.68
	Grad-CAM++	GAP	19.05	11.16	37.99	62.80	**10.75**
		CA	**19.02**	**14.76**	**40.82**	**65.53**	10.82
	Score-CAM	GAP	13.64	12.98	44.53	62.56	**11.37**
		CA	**11.53**	**18.12**	**50.32**	**65.33**	11.51
ConvNeXt-S	Grad-CAM	GAP	42.99	1.69	12.60	48.42	**30.12**
		CA	**22.09**	**14.91**	**32.65**	**84.82**	43.02
	Grad-CAM++	GAP	56.42	1.32	10.35	48.28	**33.41**
		CA	**51.87**	**9.40**	**20.55**	**84.28**	52.58
	Score-CAM	GAP	74.79	1.29	10.10	47.40	**38.21**
		CA	**64.21**	**8.81**	**18.96**	**82.92**	57.46

Results

Quantitative Experiments

PASCAL VOC 2012 - RESNET-50						
	POOLING					MAP↑
	GAP					78.32
	CA					78.35
INTERPRETABILITY METRICS						
METHOD	POOLING	AD↓	AG↑	AI↑	I↑	D↓
Grad-CAM	GAP	**12.61**	9.68	27.88	**89.10**	59.39
	CA	12.77	**15.46**	**34.53**	88.53	**59.16**
Grad-CAM++	GAP	**12.25**	9.68	27.62	**89.34**	54.23
	CA	12.28	**16.76**	**34.87**	89.02	**53.34**
Score-CAM	GAP	14.8	6.76	36.41	71.10	**39.95**
	CA	**10.96**	**21.35**	**43.82**	**89.21**	51.44

Results

Quantitative Experiments

CUB-200-2011 - ResNet-50						
	Pooling					Acc↑
	GAP					76.96
	CA					75.90
Interpretability Metrics						
Method	Pooling	AD↓	AG↑	AI↑	I↑	D↓
Grad-CAM	GAP	10.87	10.29	45.81	65.71	**6.17**
	CA	**10.44**	**17.61**	**53.54**	**74.60**	6.56
Grad-CAM++	GAP	11.35	9.68	44.32	65.64	**5.92**
	CA	**11.01**	**16.50**	**51.63**	**74.64**	6.21
Score-CAM	GAP	9.05	10.62	48.90	65.58	5.94
	CA	**6.37**	**19.50**	**60.41**	**74.22**	**2.14**

Results

Quantitative Experiments

Accuracy and Parameters			
Placement	CLS dim	#Param	Acc↑
$S_0 - S_4$	64	6.96M	**74.70**
$S_1 - S_4$	256	6.95M	74.67
$S_2 - S_4$	512	6.82M	74.67
$S_3 - S_4$	1024	6.29M	74.67
$S_4 - S_4$	2048	4.20M	74.63

Results

Quantitative Experiments

Interpretability Metrics						
Method	Placement	AD↓	AG↑	AI↑	I↑	D↓
Grad-CAM	$S_0 - S_4$	**12.54**	**22.67**	48.56	75.53	13.50
	$S_1 - S_4$	12.69	22.65	48.31	75.53	13.41
	$S_2 - S_4$	**12.54**	21.67	**48.58**	75.54	13.50
	$S_3 - S_4$	12.69	22.28	47.89	**75.55**	13.40
	$S_4 - S_4$	12.77	20.65	47.14	74.32	**13.37**
Grad-CAM++	$S_0 - S_4$	13.99	19.29	44.60	75.21	13.78
	$S_1 - S_4$	13.99	19.29	44.62	75.21	13.78
	$S_2 - S_4$	13.71	**19.90**	**45.43**	75.34	13.50
	$S_3 - S_4$	13.69	19.61	45.04	**75.36**	13.50
	$S_4 - S_4$	**13.67**	18.36	44.40	74.19	**13.30**
Score-CAM	$S_0 - S_4$	**7.09**	23.65	54.20	74.91	14.68
	$S_1 - S_4$	**7.09**	23.65	54.20	74.92	14.68
	$S_2 - S_4$	**7.09**	**23.66**	**54.21**	74.91	14.68
	$S_3 - S_4$	7.74	23.03	52.92	**74.97**	14.65
	$S_4 - S_4$	7.52	19.45	50.45	74.19	**14.46**

Results

Quantitative Experiments

Accuracy and Parameters						
	Representation			#Param		Acc↑
	Class agnostic			32.53M		74.70
	Class specific			32.59M		74.68
Interpretability Metrics						
Method	Representation	AD↓	AG↑	AI↑	I↑	D↓
Grad-CAM	Class agnostic	12.54	22.67	48.56	75.53	13.50
	Class specific	12.53	22.66	48.58	75.54	13.50
Grad-CAM++	Class agnostic	13.99	19.29	44.60	75.21	13.78
	Class specific	13.99	19.28	44.62	75.20	13.78
Score-CAM	Class agnostic	7.09	23.65	54.20	74.91	14.68
	Class specific	7.08	23.64	54.15	74.99	14.53

Table of Contents

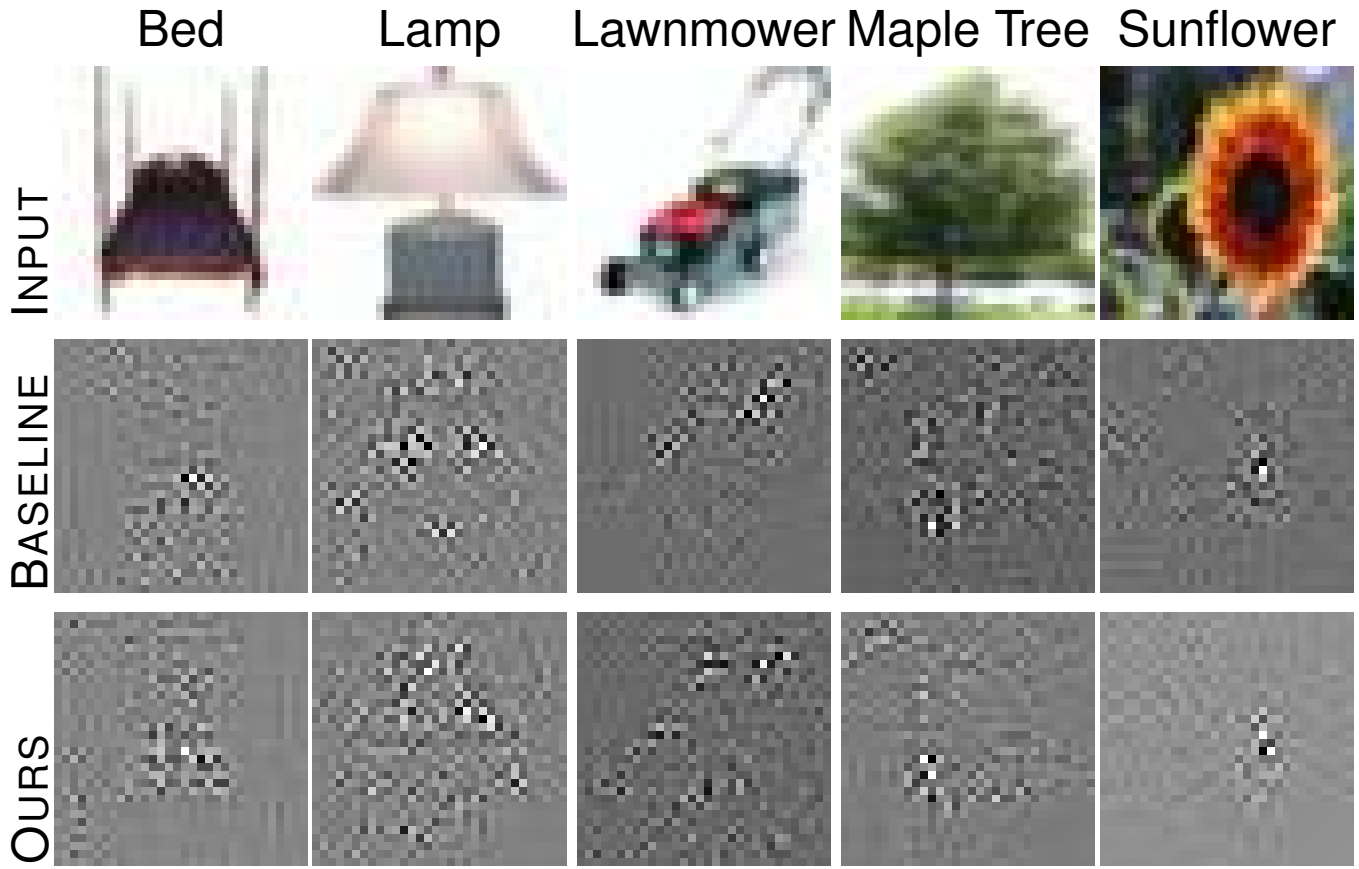
1 Opti-CAM: Optimizing saliency maps for interpretability

2 CA-Stream: Attention-based pooling for interpretable image recognition

3 A learning paradigm for interpretable gradients

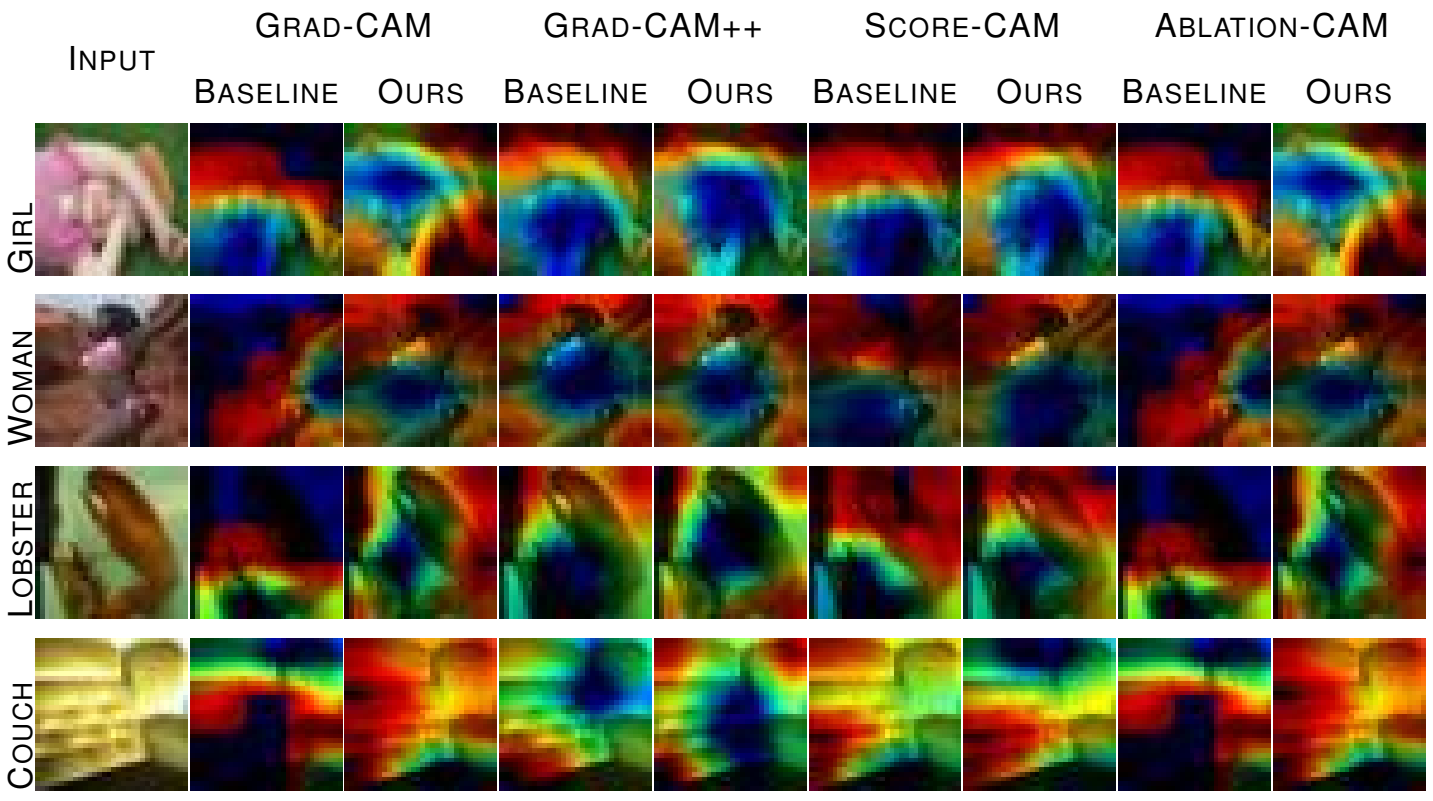
Results
Qualitative Experiments

Denoising effect



Results
Qualitative Experiments

CAM Visualizations



Results
Quantitative Experiments

Recognition Metrics

Recognition Metrics			
Model	Error	λ	Acc↑
ResNet-18	-	-	**73.42**
	Cosine	7.5×10^{-3}	72.86
MobileNet-V2	-	-	59.43
	Cosine	1×10^{-3}	**62.36**

Results
Quantitative Experiments

Recognition Metrics

RECOGNITION METRICS			
MODEL	ERROR	λ	ACC↑
RESNET-18	-	-	**73.42**
	COSINE	7.5×10^{-3}	72.86
MOBILENET-V2	-	-	59.43
	COSINE	1×10^{-3}	**62.36**

Recognition properties are maintained.

Results

Quantitative Experiments

Interpretability metrics

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

Results
Quantitative Experiments

Interpretability metrics

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

Interpretable properties are enhanced. Deletion still poses an issue.

Table of Contents

1. Opti-CAM: Optimizing saliency maps for interpretability
2. CA-Stream: Attention-based pooling for interpretable image recognition
3. A learning paradigm for interpretable gradients

References I