Learning Discriminative Representations to Interpret Image Recognition Models

Thèse de Doctorat

Felipe Torres Figueroa

École Centrale de Marseille

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

Marseille, September 23rd 2024

Reviewers: Frédéric Jurie - Univ. Caen
Giorgos Tolias - CTU
Examiners: Frédéric Precioso - Univ Côte d'Azur
Diane Larlus - NAVER Labs.

Supervisors: Stéphane Ayache - Aix Marseille Univ.
Ronan Sicre - Centrale Marseille
Yannis Avrithis - IARAI

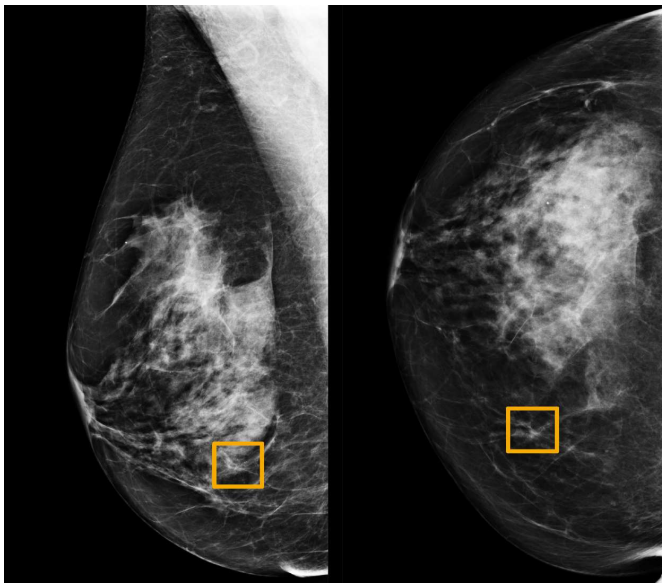
Table of Contents

Table of Contents

Motivation

Introductory Examples

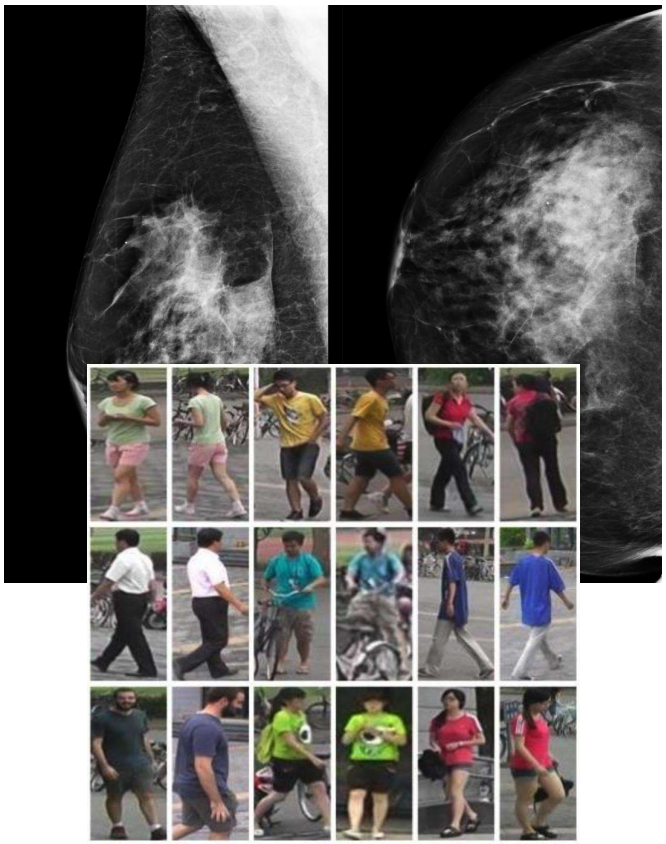
Consider these situations:



Motivation

Introductory Examples

Consider these situations:



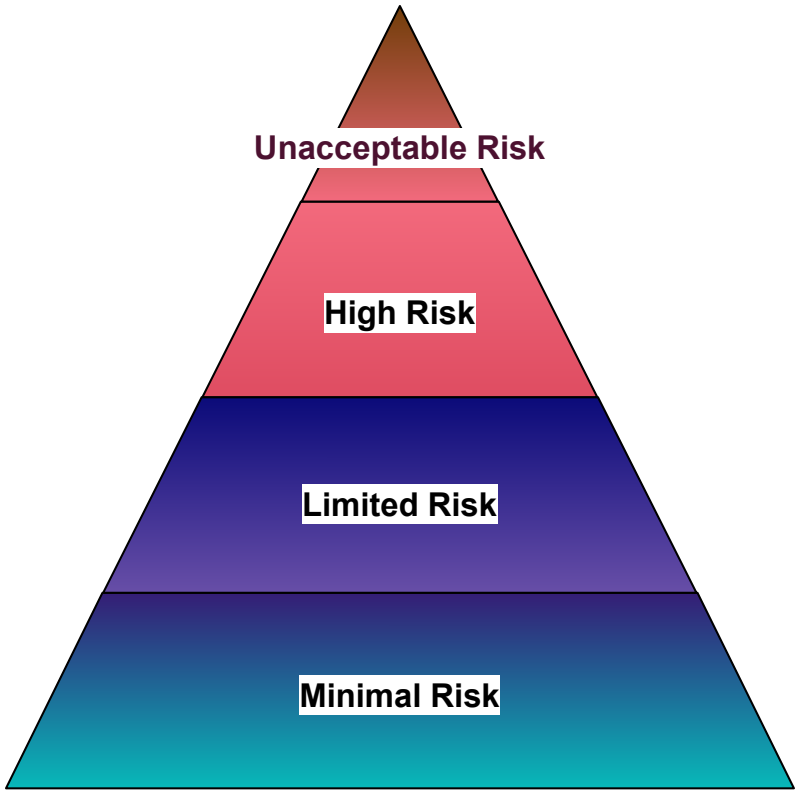
Motivation

Introductory Examples

Consider these situations:

Motivation

AI accepted in society?

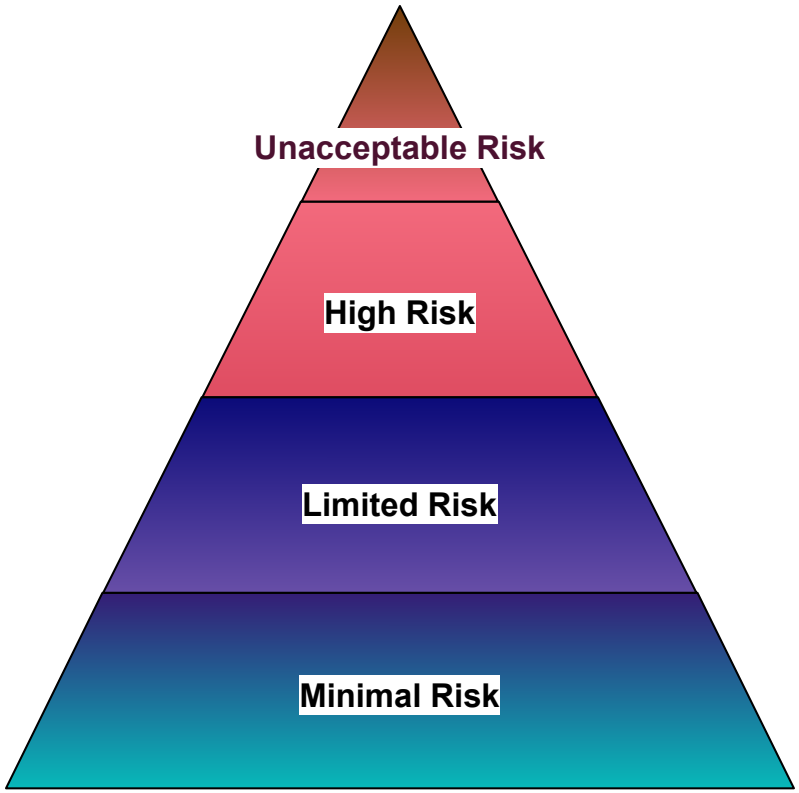


European Act for regulation of AI. [1]

- How do we **know how** a system works?
- How do we **know how** safe a system is?
- If a system fails, **who** is accountable?

Motivation

AI accepted in society?



European Act for regulation of AI. [1]

- How do we **know how** a system works?
- How do we **know how** safe a system is?
- If a system fails, **who** is accountable?

We must **explain** the behaviour of these models and **interpret** their predictions.

Image Recognition Models

An ongoing revolution since 2012

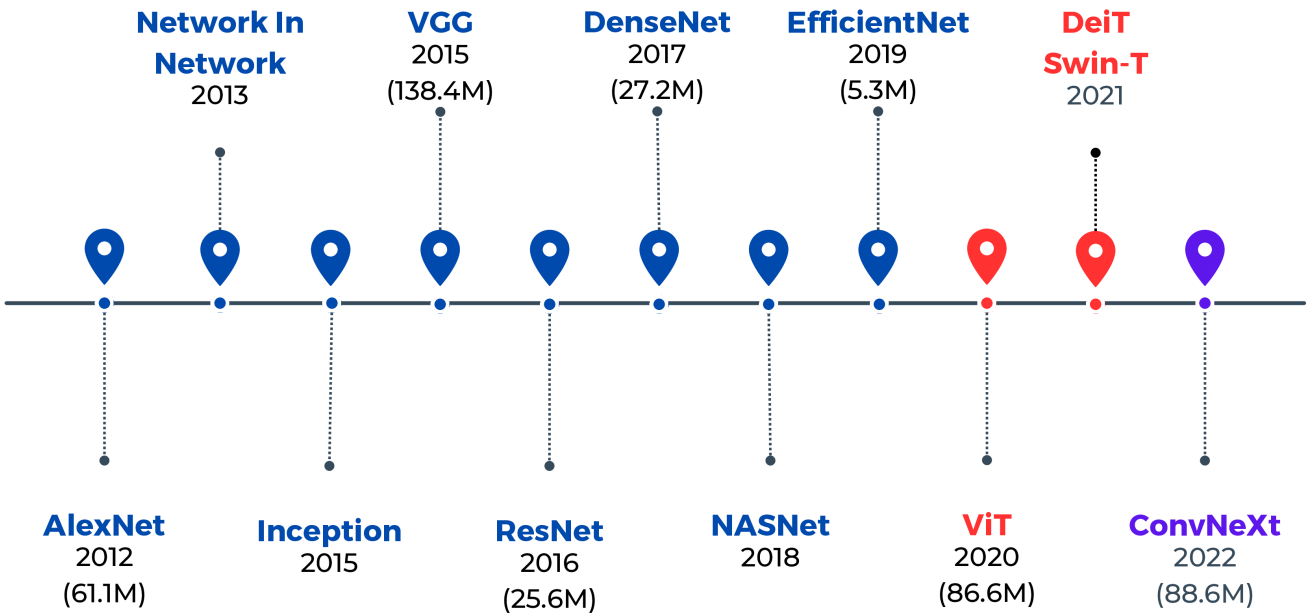
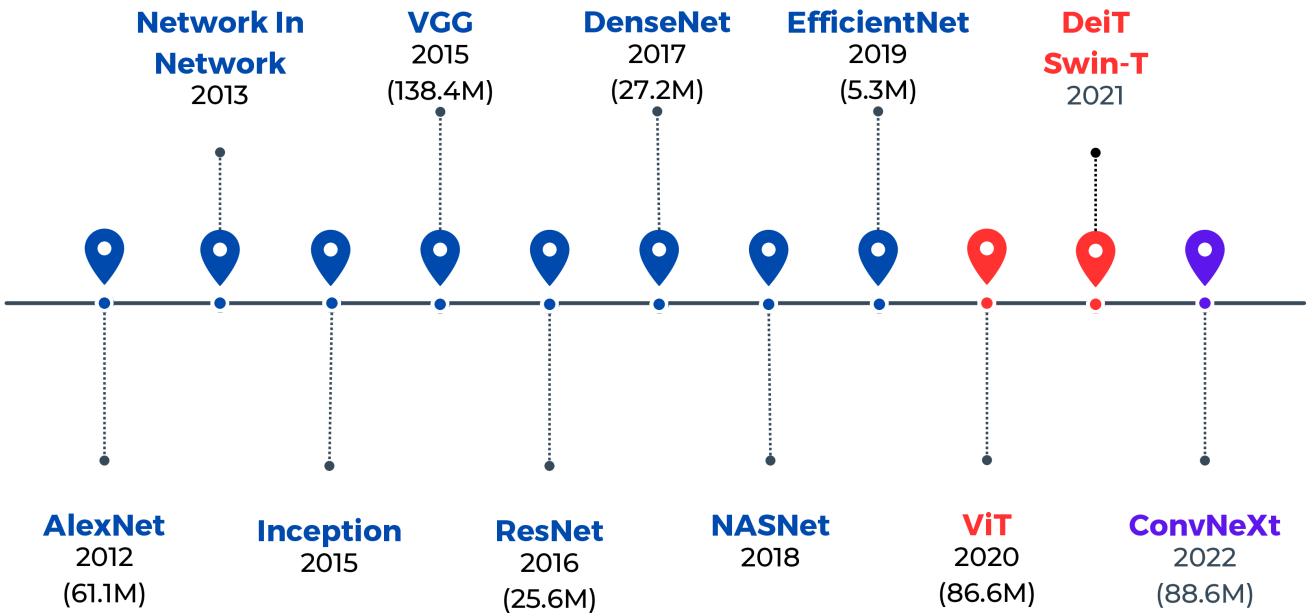


Image Recognition Models

An ongoing revolution since 2012

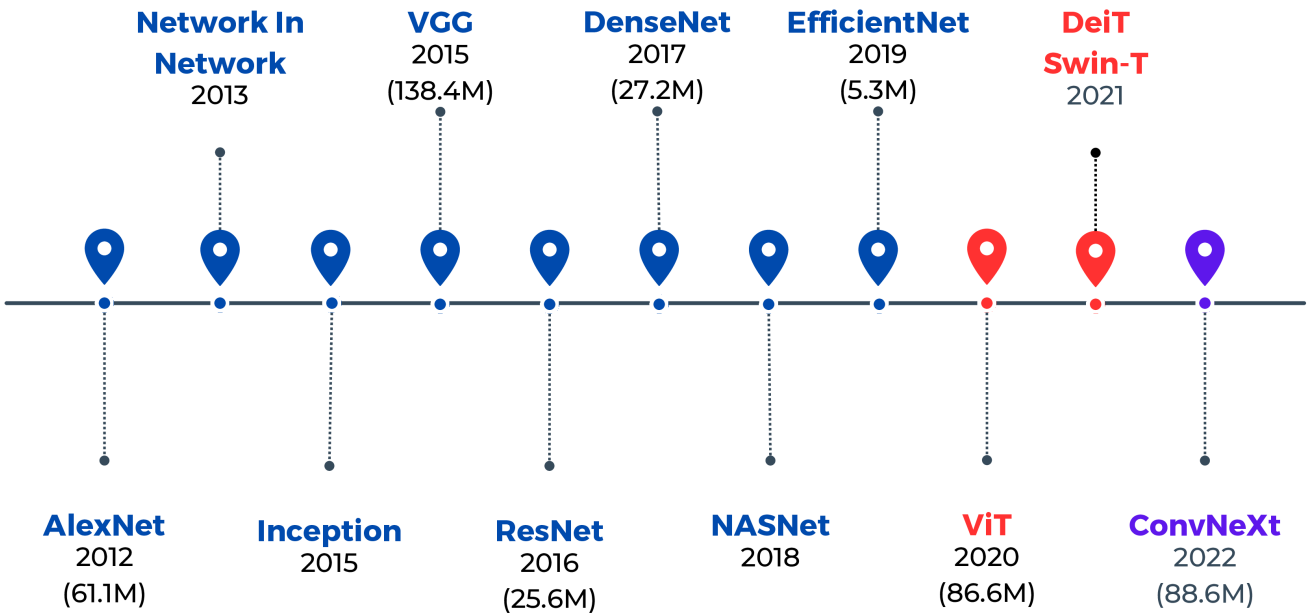


Contributing factors:

- Widespread of Internet.
- Big Dataset collection.
- Popularization of the GPU.

Image Recognition Models

An ongoing revolution since 2012



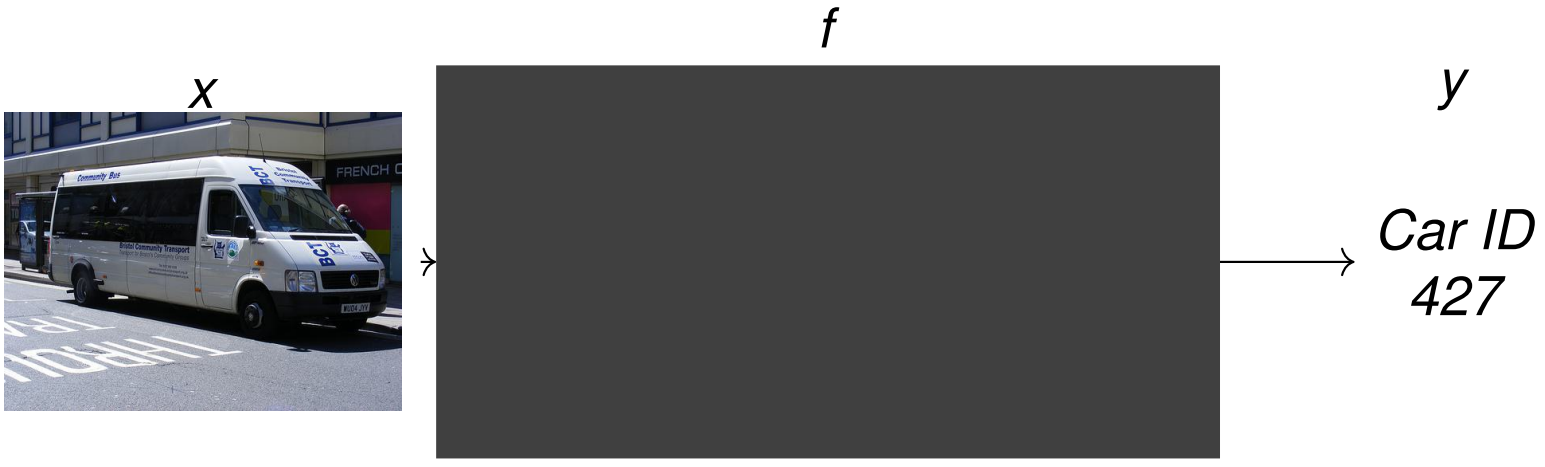
Contributing factors:

- Widespread of Internet.
- Big Dataset collection.
- Popularization of the GPU.

Two paradigms:
Convolutional Neural Networks (CNNs).
Transformers.

Interpretability
Motivation

We are interested in understanding models, behaving like a black box:



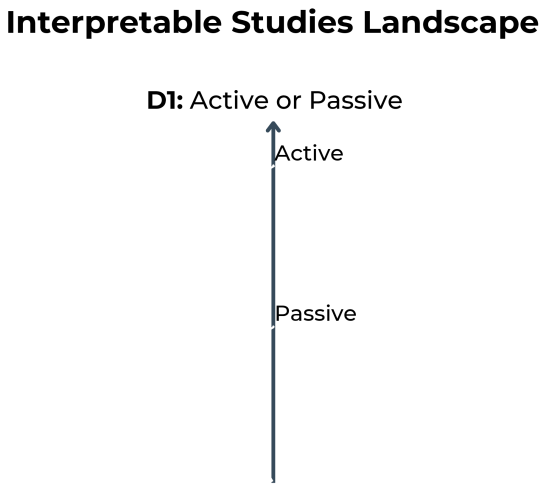
We want to *know why* $f(x) \rightarrow y$

Interpretability

A paradigmn on interpretable studies

Interpretability approaches have been categorized by two studies [2, 3]:

1. **Passive or Active?**
   To add or not modifications to the model.

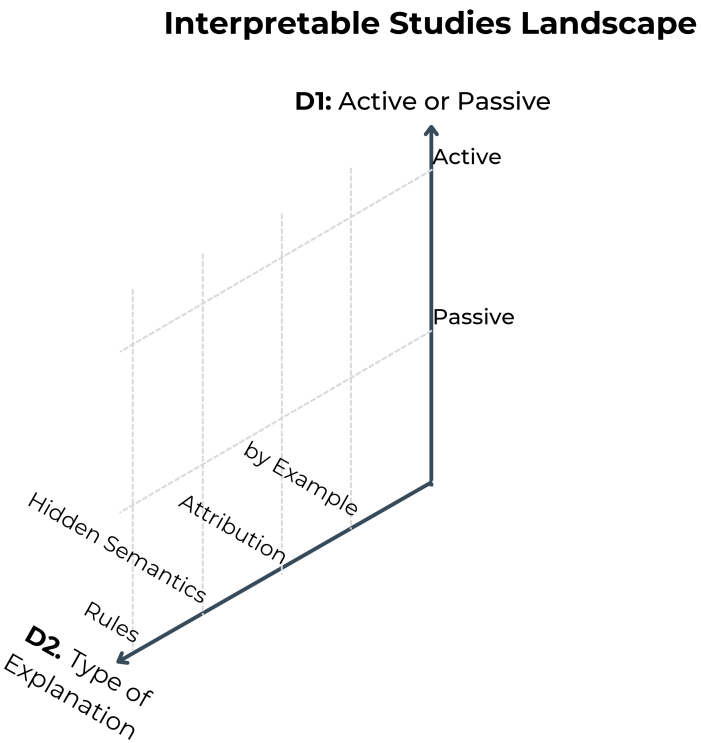


Interpretability

A paradigmn on iterpretable studies

Interpretability approaches have been categorized by two studies [2, 3]:

1. **Passive or Active?**
   To add or not modifications to the model.
2. **Type of Explanation**
   Is it an example? An Attribution? Hidden Semantics? Or a Rule?

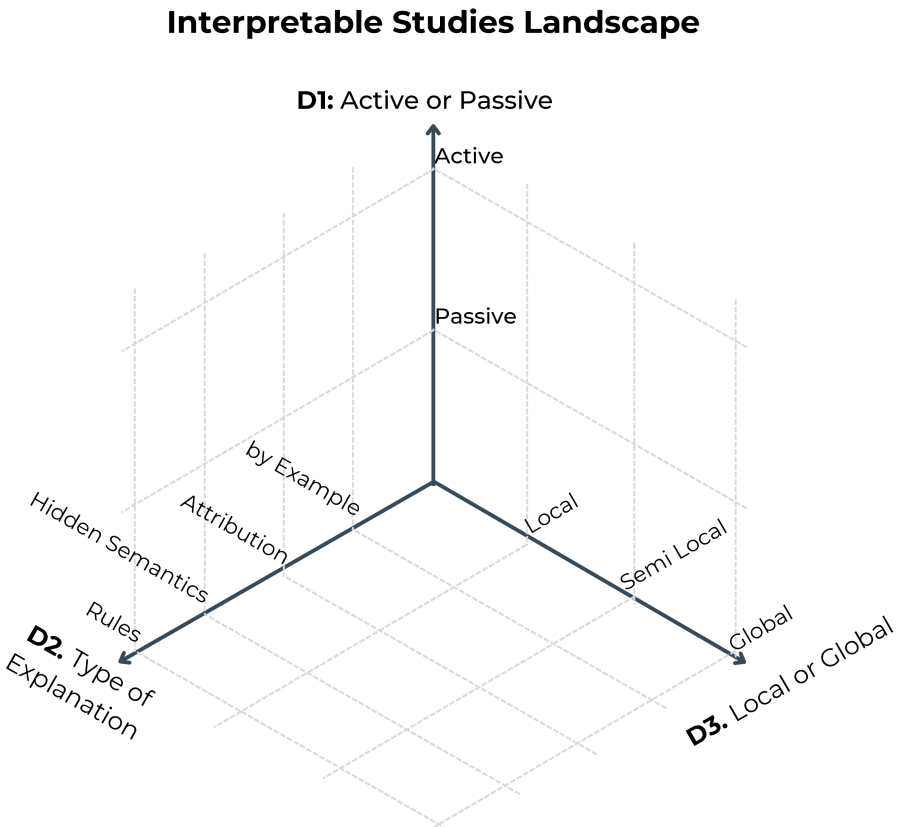


Interpretability

A paradigmn on iterpretable studies

Interpretability approaches have been categorized by two studies [2, 3]:

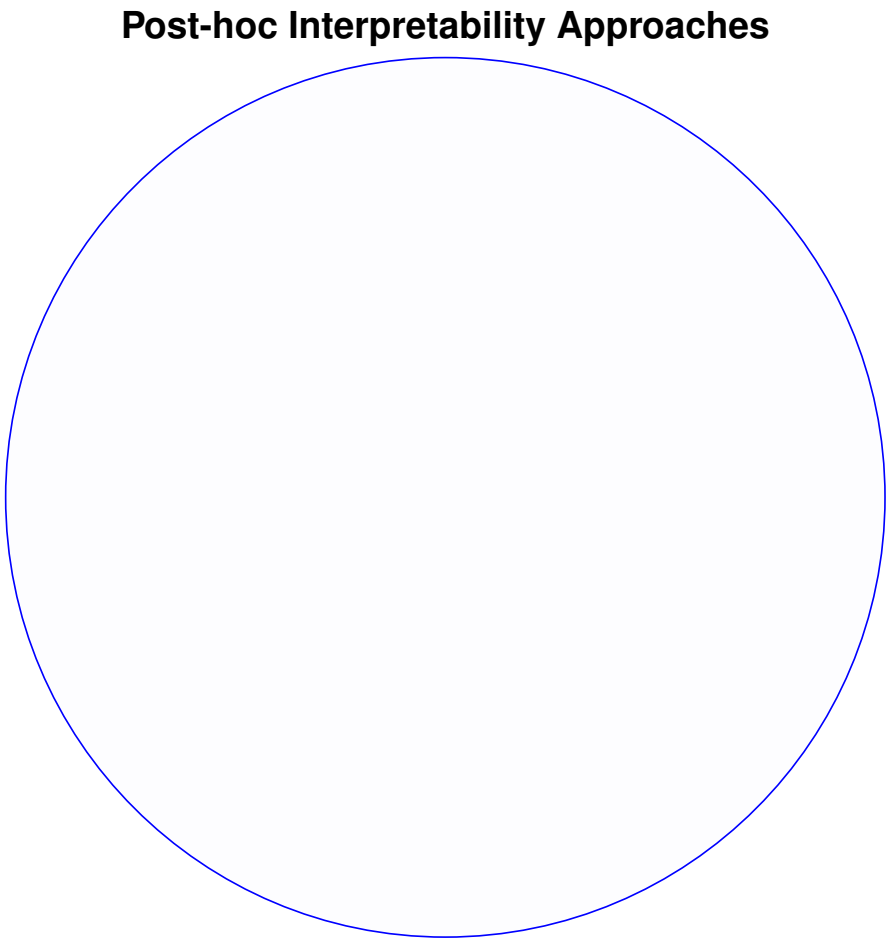
1. **Passive or Active?**[1]
   To add or not modifications to the model.
2. **Type of Explanation**
   Is it an example? An Attribution? Hidden Semantics? Or a Rule?
3. **Local or Global?**
   Is it a brief explanation? Does it cover more data/ of the model? Does it explain the model/data entirely?



[1] Lipton defines Interpretability Similarly. Transparency or Post-hoc Interpretations

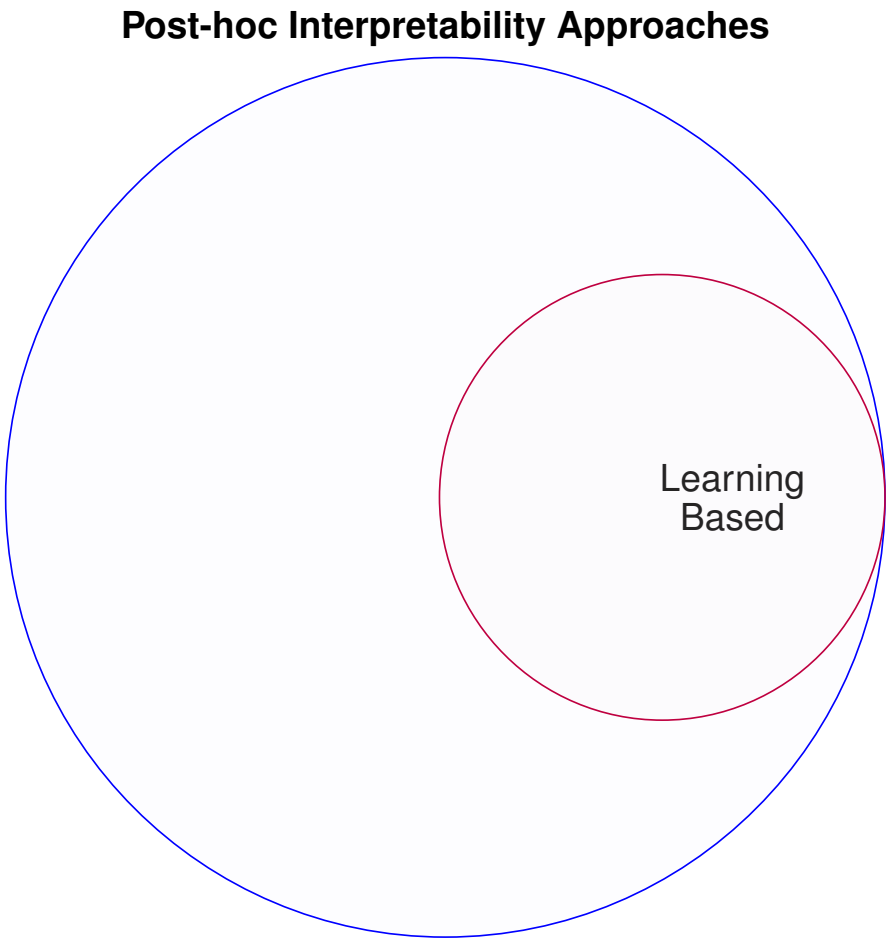
Interpretability

Post-hoc Interpretability



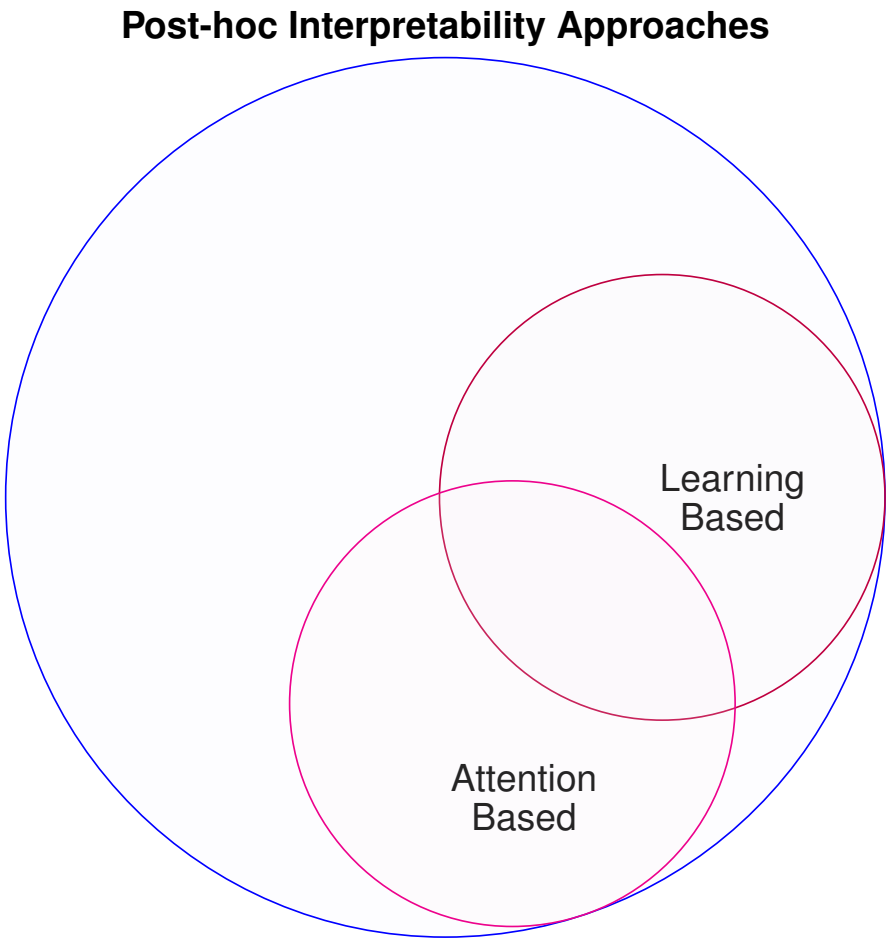
Interpretability

Post-hoc Interpretability



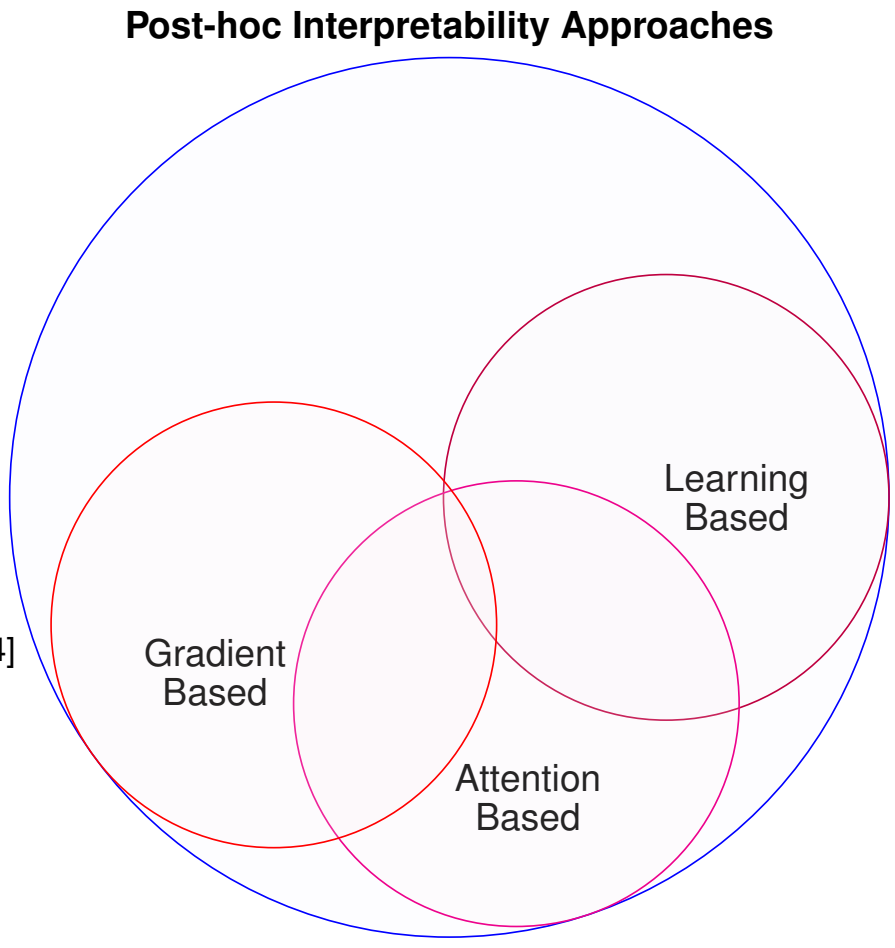
Interpretability

Post-hoc Interpretability



Interpretability

Post-hoc Interpretability



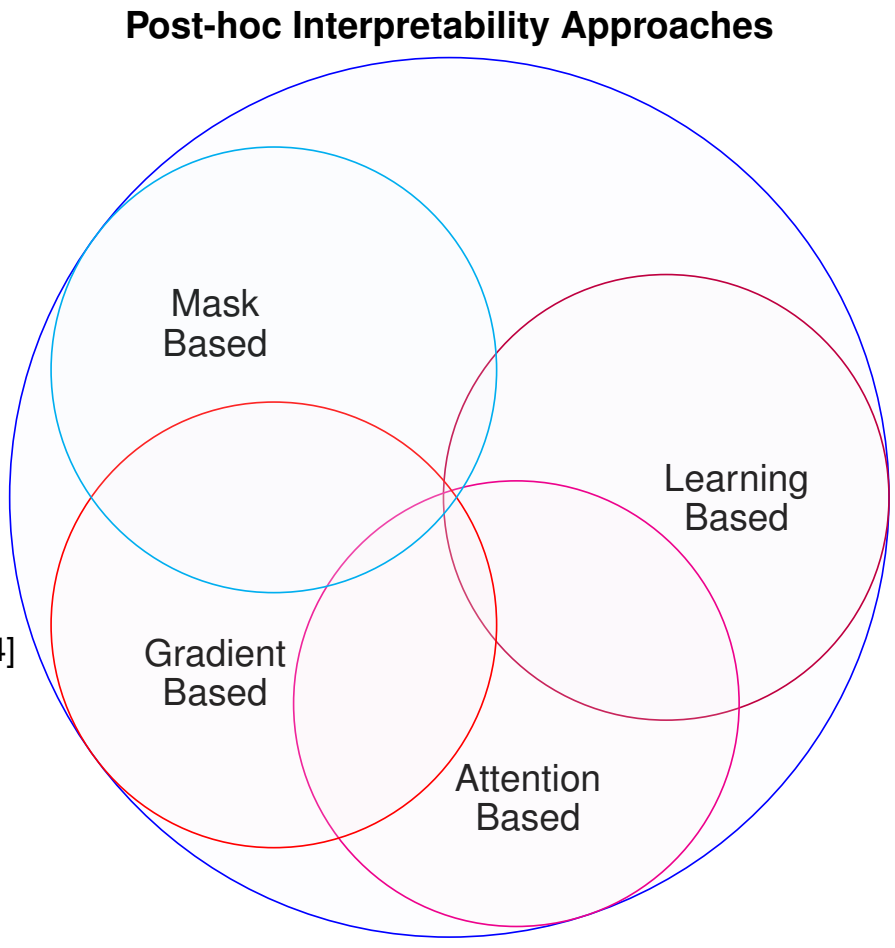
Pure Gradient
Standard backpropagation[4]
Guided backpropagation[5]
Smoothgrad[6]

Interpretability

Post-hoc Interpretability

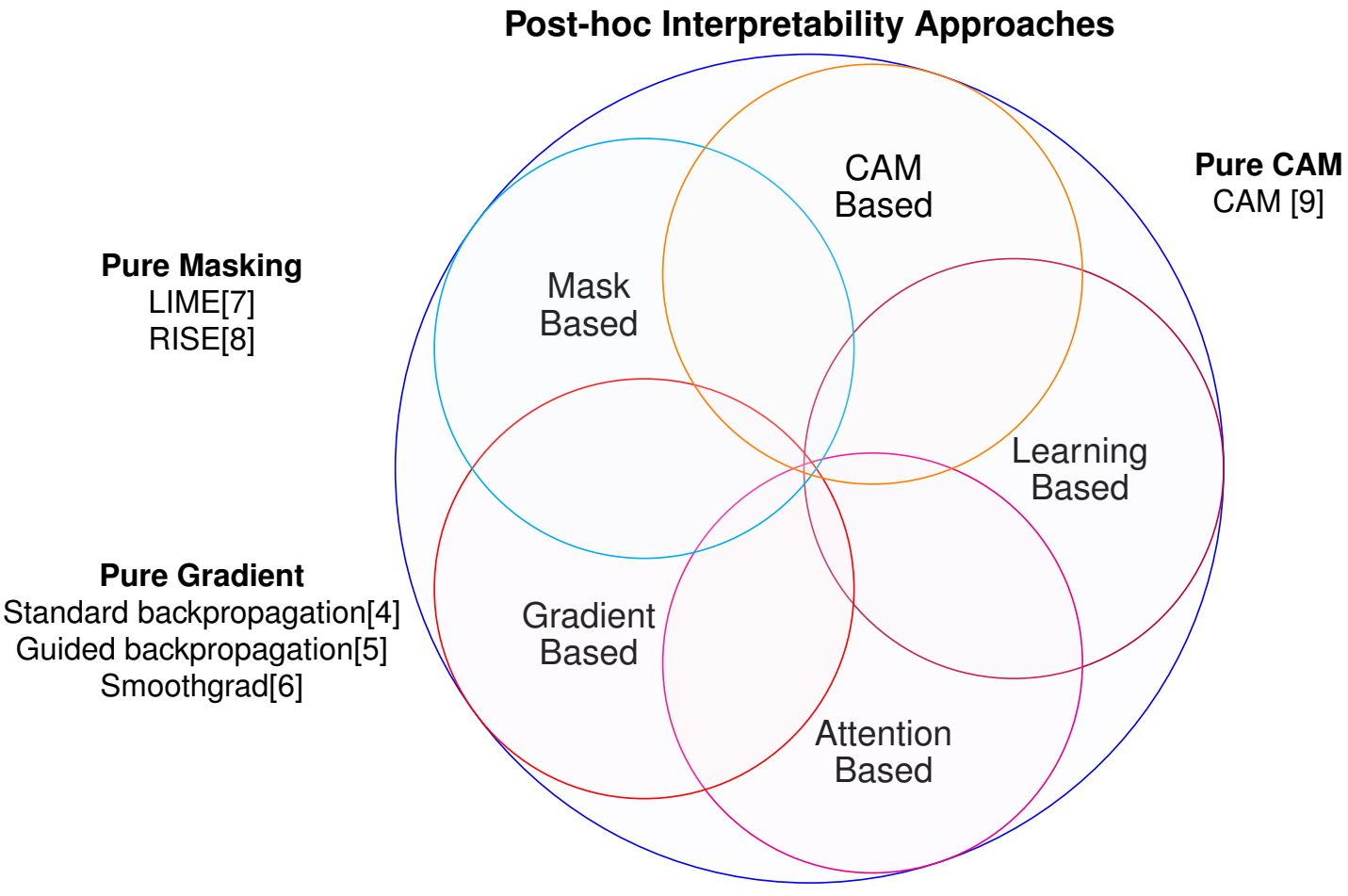
Pure Masking
LIME[7]
RISE[8]

Pure Gradient
Standard backpropagation[4]
Guided backpropagation[5]
Smoothgrad[6]



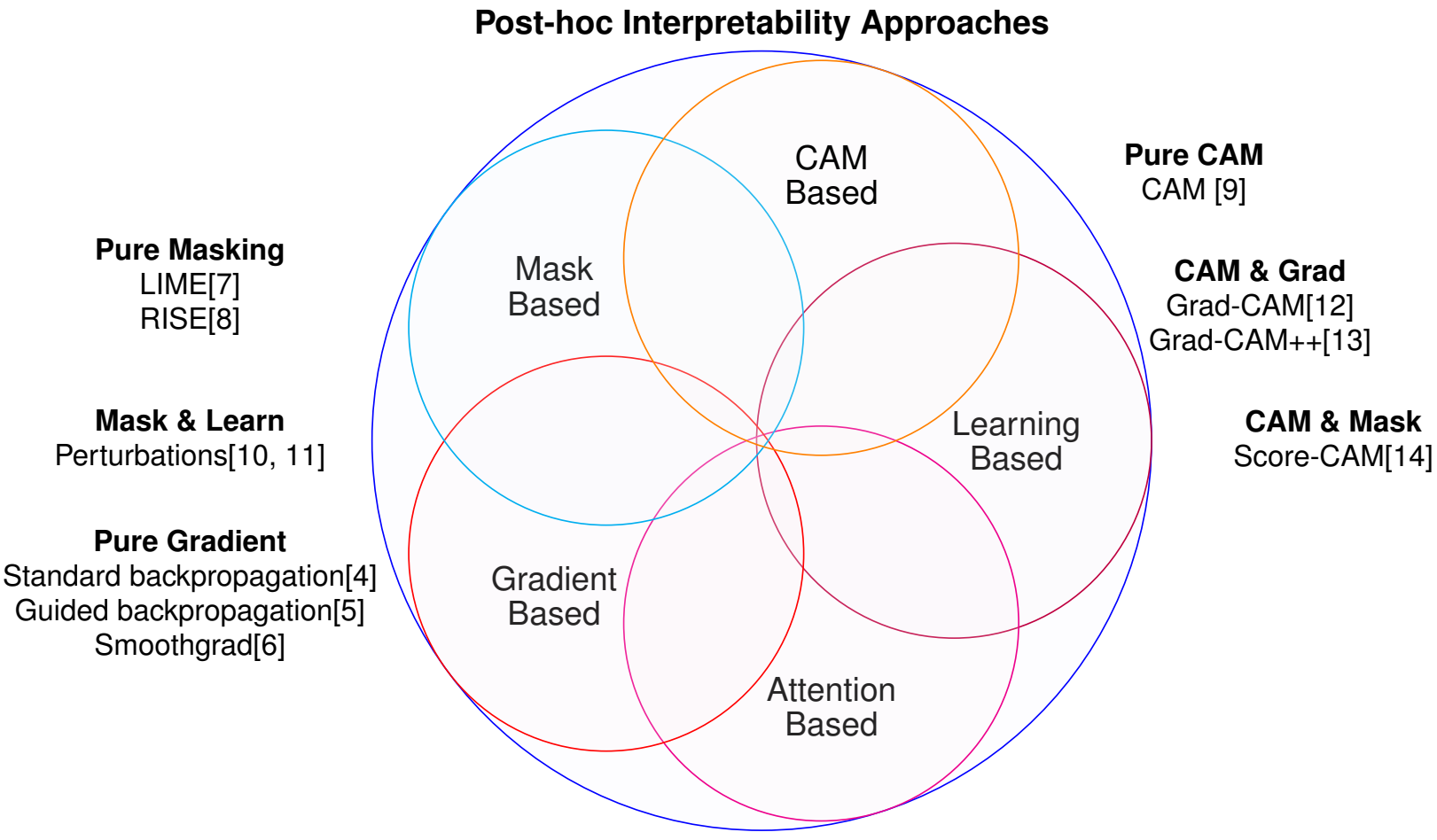
Interpretability

Post-hoc Interpretability



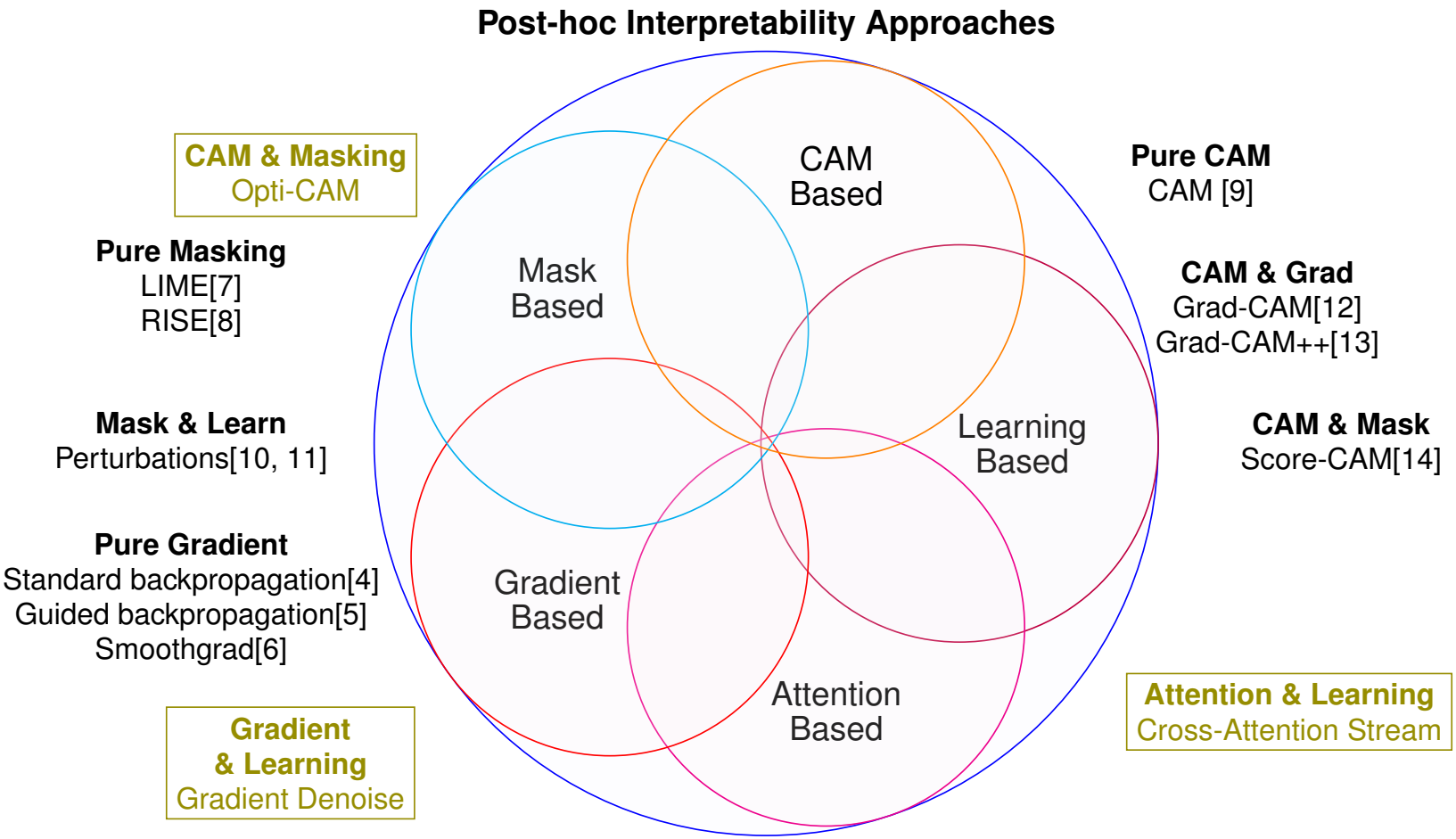
Interpretability

Post-hoc Interpretability



Interpretability

Post-hoc Interpretability



Contributions

Overview

We are interested in understanding models, behaving like a black box:

Contributions

Overview

We are interested in understanding models, behaving like a black box:

How to *know why* $f(x) \rightarrow y$

Contributions
Overview

We are interested in understanding models, behaving like a black box:

How to *know why* $f(x) \rightarrow y$

Saliency Maps

Contributions

Overview

We are interested in understanding models, behaving like a black box:

How to *know why* $f(x) \rightarrow y$

Saliency Maps

Gradients

Contributions

Overview

We are interested in understanding models, behaving like a black box:

How to *know why* $f(x) \rightarrow y$

Saliency Maps

Gradients

Other Representations

Contributions

Overview

We are interested in understanding models, behaving like a black box:

How to *know why* $f(x) \rightarrow y$

Saliency Maps

Gradients

Other Representations

We demonstrate improvements in these approaches

Table of Contents

Preliminaries

Premise

Is it possible to optimize a saliency map probability with the highest probability of prediction?

$p(x)_{774} = 0.745$

$p(x)^{CAM}_{774} = 0.728$

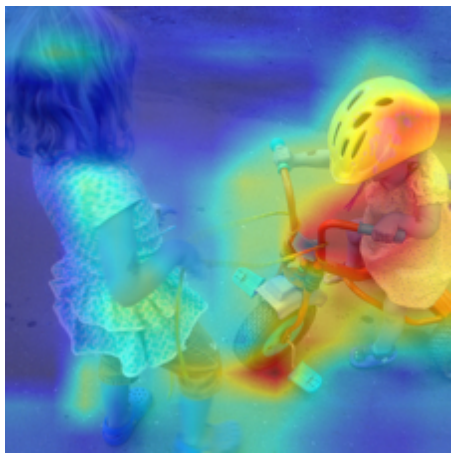
Preliminaries

Premise

Is it possible to optimize a saliency map probability with the highest probability of prediction?

$p(x)_{774} = 0.745$

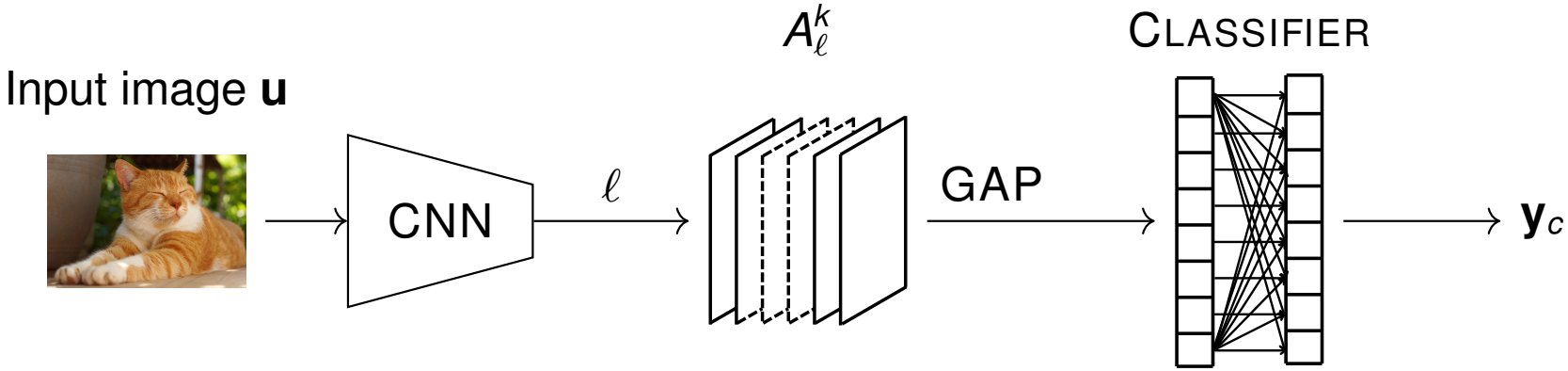
$p(x)^{CAM}_{774} = 0.728$



$p(x)^{Optimal}_{774} = 0.853$

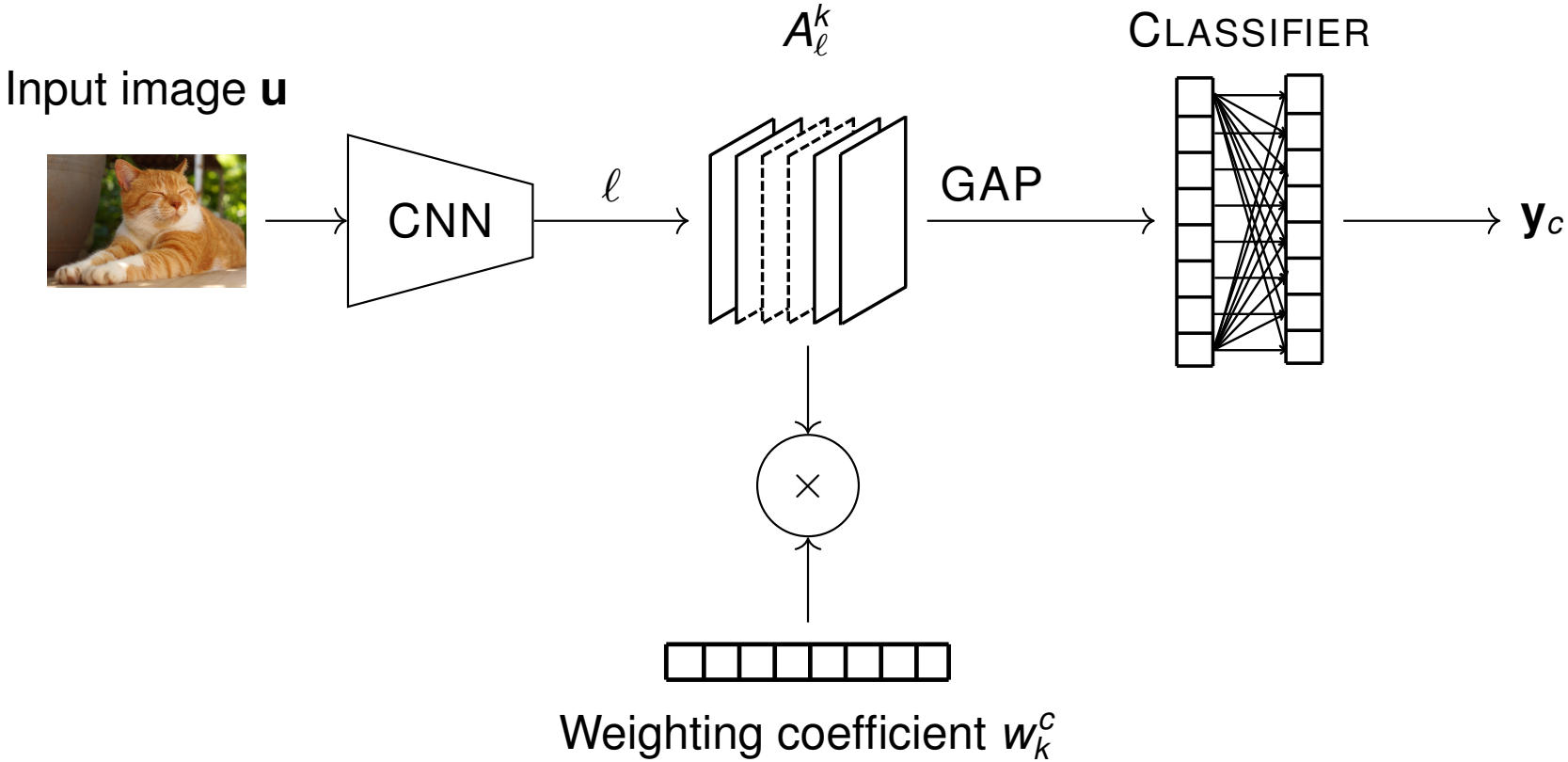
Background

CAM Methods

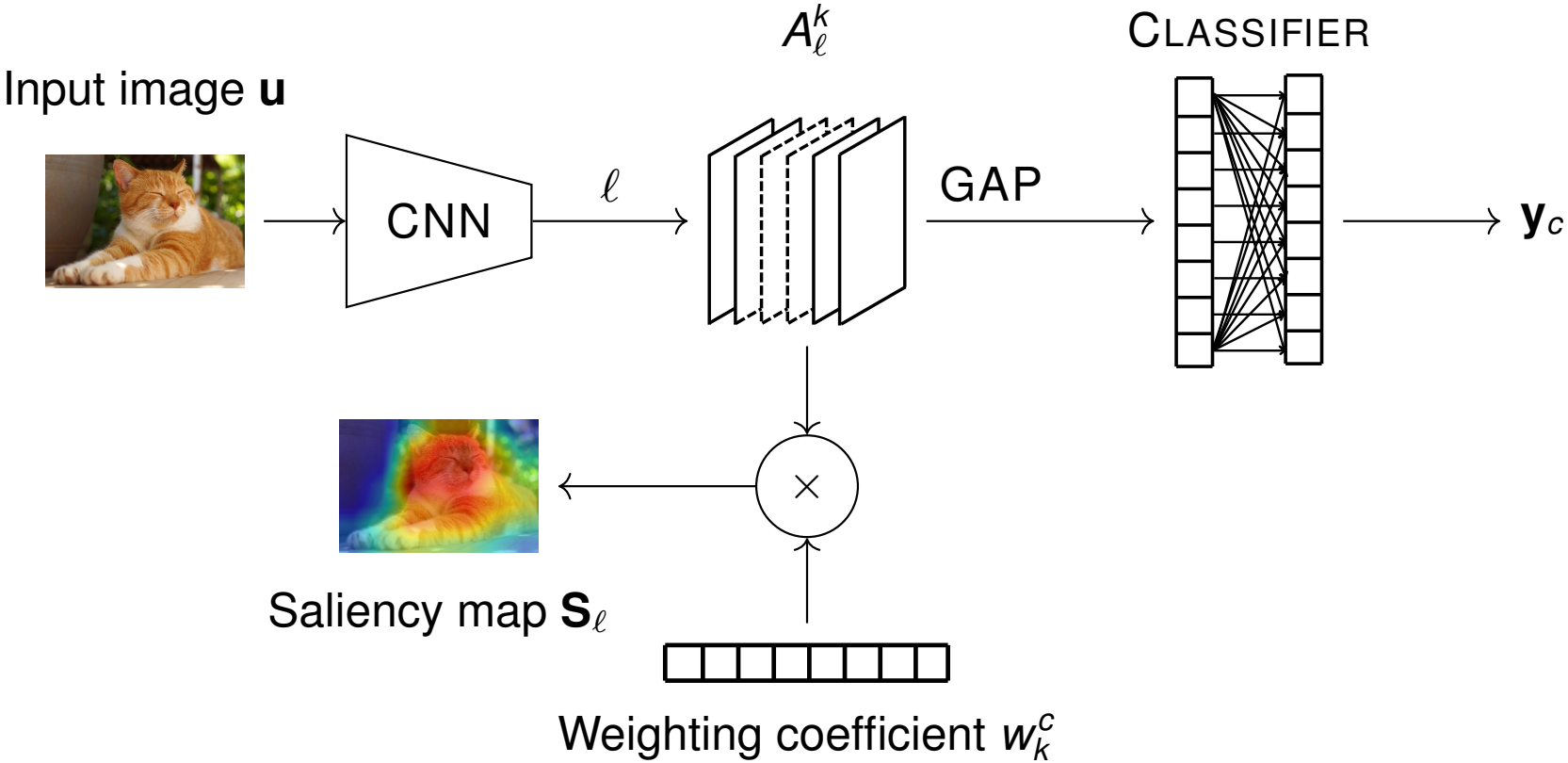


Background

CAM Methods



Background
CAM Methods



Background
CAM Formalized

For an activation function (h) and a feature map (A_ℓ^k):

$$S_\ell^c(\mathbf{u}) := h\left(\sum_k w_k^c A_\ell^k\right) \tag{1}$$

Background
CAM Formalized

For an activation function (h) and a feature map (A_ℓ^k):

$$S_\ell^c(\mathbf{u}) := h\left(\sum_k w_k^c A_\ell^k\right) \tag{1}$$

The weighting coefficient ($\mathbf{w_k^c}$) defines the CAM variant.

Background
CAM Formalized

For an activation function (h) and a feature map (A_ℓ^k):

$$S_\ell^c(\mathbf{u}) := h\left(\sum_k w_k^c A_\ell^k\right) \tag{1}$$

The weighting coefficient ($\mathbf{w_k^c}$) defines the CAM variant.
Milestone CAM models:

CAM[9]:
w_k^c comes from the weights for one class. Only on the last convolutional layers followed by GAP and simple classifiers.

Background
CAM Formalized

For an activation function (h) and a feature map (A_ℓ^k):

$$S_\ell^c(\mathbf{u}) := h\left(\sum_k w_k^c A_\ell^k\right) \tag{1}$$

The weighting coefficient ($\mathbf{w_k^c}$) defines the CAM variant.
Milestone CAM models:

CAM[9]:
w_k^c comes from the weights for one class. Only on the last convolutional layers followed by GAP and simple classifiers.

Grad-CAM/Grad-CAM++ [12, 13]:
w_k^c is computed using gradients flowing from prediction to input space. It can be used at any point of the model.

Background
CAM Formalized

For an activation function (h) and a feature map (A_ℓ^k):

$$S_\ell^c(\mathbf{u}) := h\left(\sum_k w_k^c A_\ell^k\right) \tag{1}$$

The weighting coefficient ($\mathbf{w_k^c}$) defines the CAM variant.
Milestone CAM models:

CAM[9]:
w_k^c comes from the weights for one class. Only on the last convolutional layers followed by GAP and simple classifiers.

Grad-CAM/Grad-CAM++ [12, 13]:
w_k^c is computed using gradients flowing from prediction to input space. It can be used at any point of the model.

Score-CAM[14]:
w_k^c is computed from the contributions of individual feature maps masking the input image.

Background
CAM Formalized

For an activation function (h) and a feature map (A_ℓ^k):

$$S_\ell^c(\mathbf{u}) := h\left(\sum_k w_k^c A_\ell^k\right) \tag{1}$$

The weighting coefficient ($\mathbf{w_k^c}$) defines the CAM variant.
Milestone CAM models:

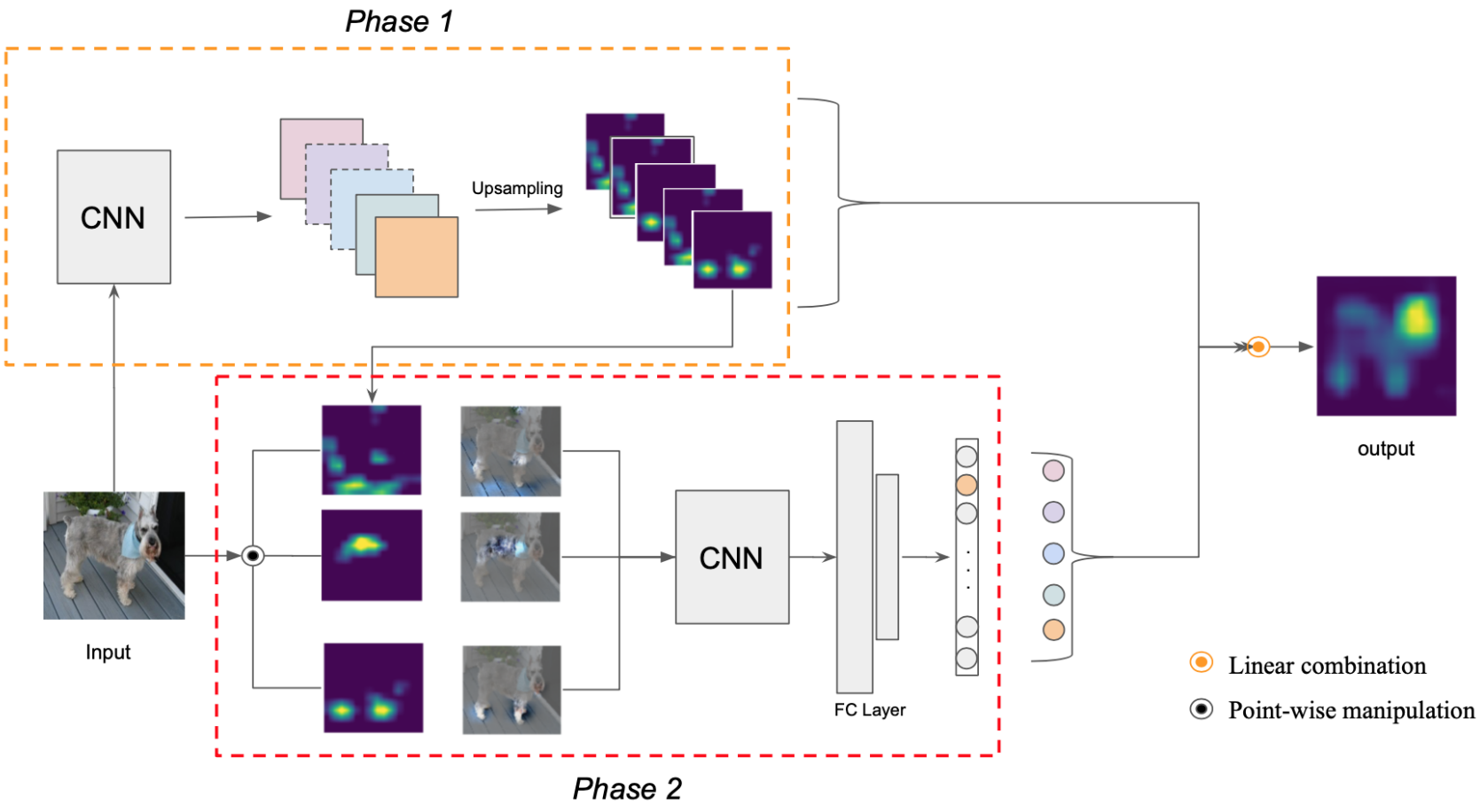
CAM[9]:
w_k^c comes from the weights for one class. Only on the last convolutional layers followed by GAP and simple classifiers.

Grad-CAM/Grad-CAM++ [12, 13]:
w_k^c is computed using gradients flowing from prediction to input space. It can be used at any point of the model.

Score-CAM[14]:
w_k^c is computed from the contributions of individual feature maps masking the input image.

The approach found in Score-CAM is interesting. Diving deeper:

Background
Score-CAM



Original from [14]

Background
Score-CAM Formalized

Following equation (1):

Background
Score-CAM Formalized

Following equation (1):

- $h := \mathbf{ReLU}$
- $\mathbf{w_k^c} := softmax(\mathbf{u}^c)_k$

Background
Score-CAM Formalized

Following equation (1):

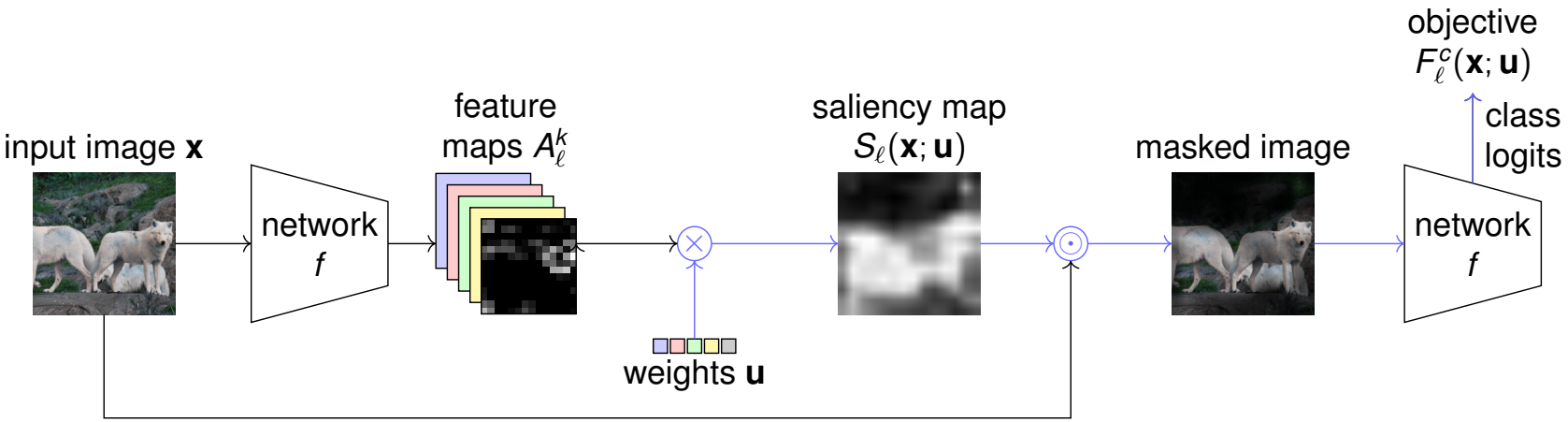
- $h := \mathbf{ReLU}$
- $\mathbf{w_k^c} := softmax(\mathbf{u}^c)_k$

With feature map contribution ($\mathbf{u^c}_k$), model (f), normalization (n) and upsampling (**up**):

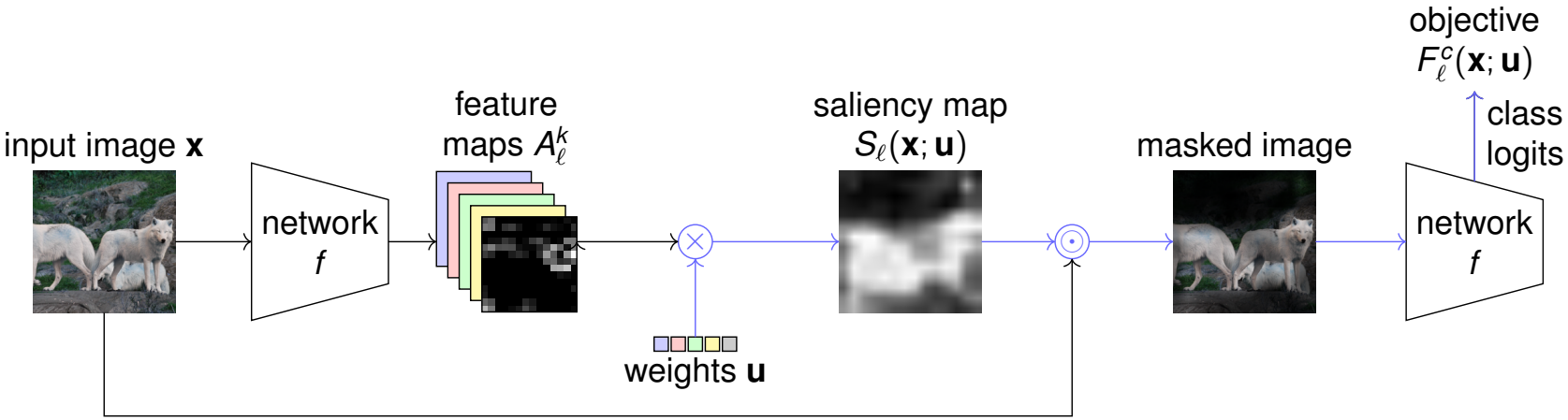
$$u_k^c := f(\mathbf{x} \odot n(\text{up}(A_\ell^k)))_c - f(\mathbf{x})_c \tag{2}$$

Opti-CAM

Approach



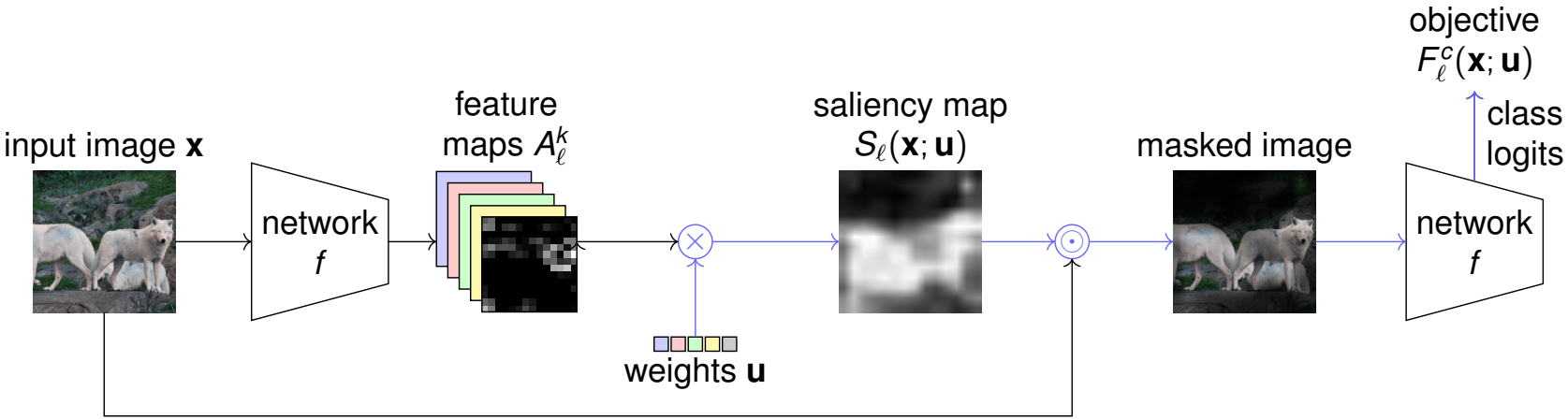
Opti-CAM
Approach



Consider S_ℓ, a function of $\mathbf{u}$ and x:

$$S_\ell(\mathbf{x};\mathbf{u}) := \sum_k \text{softmax}(\mathbf{u})_k A_\ell^k \tag{3}$$

Opti-CAM
Approach



Consider S_ℓ, a function of $\mathbf{u}$ and x:

$$S_\ell(\mathbf{x};\mathbf{u}) := \sum_k \text{softmax}(\mathbf{u})_k A_\ell^k \tag{3}$$

$\mathbf{u}^*$ is the vector that maximizes class c probability, when masking the the input with S_ℓ.

$$\mathbf{u}^* := \underset{\mathbf{u}}{\text{argmax}} F_\ell^c(\mathbf{x};\mathbf{u}) \tag{4}$$

And the objective $F_\ell^c(\mathbf{x};\mathbf{u}) := f(\mathbf{x} \odot n(\text{up}(S_\ell(\mathbf{x};\mathbf{u}))))$

Evaluating Interpretability

Three paradigms

Evaluating Interpretability

Three paradigms

- **Interpretable Image Recognition[13]**
  Average Drop (AD), Average Increase (AI), Average Gain (AG).

Evaluating Interpretability

Three paradigms

- **Interpretable Image Recognition[13]**
  Average Drop (AD), Average Increase (AI), Average Gain (AG).
- **Causality Analysis[8]**
  Insertion & Deletion.

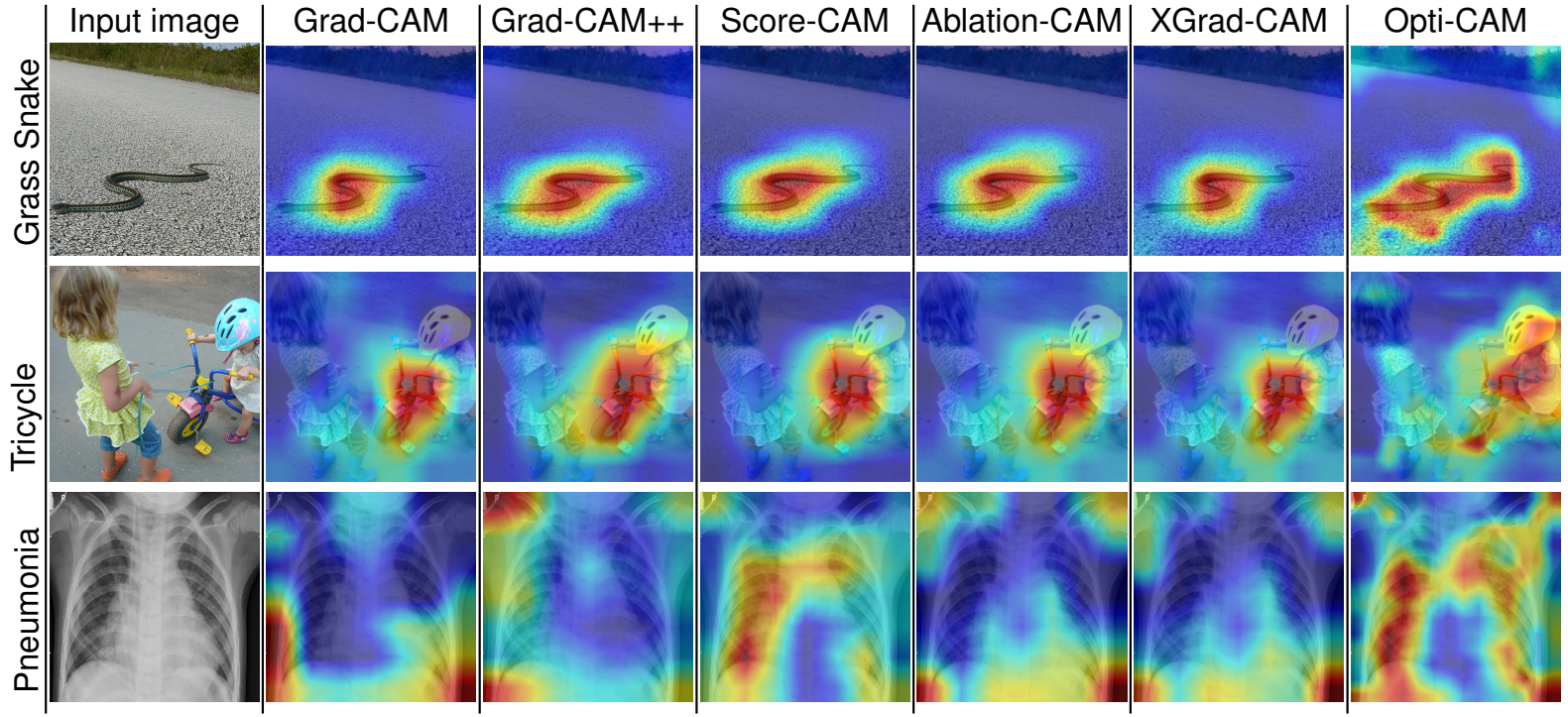
Evaluating Interpretability

Three paradigms

- **Interpretable Image Recognition[13]**
  Average Drop (AD), Average Increase (AI), Average Gain (AG).
- **Causality Analysis[8]**
  Insertion & Deletion.
- **Weakly Supervised Localization Approach**
  Official Metric (OM), Localization Error (LE), Pixel-wise F_1 score (F1) Box Accuracy (BA), Standard Pointing Game (SP), Energy Pointing Game (EP).

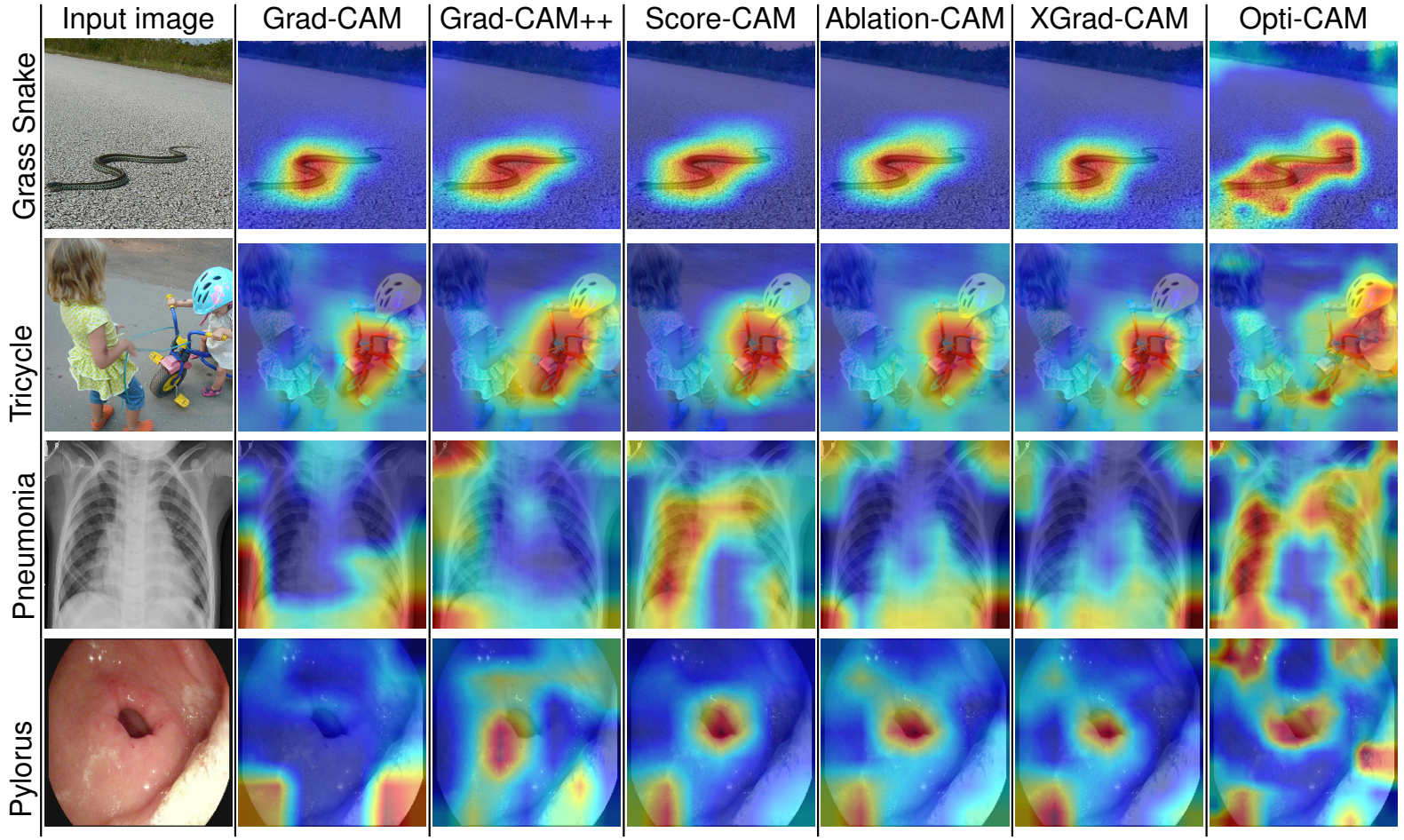
Results

Qualitative Evaluation



Results

Qualitative Evaluation



Results

Quantitative Evaluation

Classification Experiments:

Method	ResNet50						VGG16					
	AD↓	AG↑	AI↑	*I* ↑	*D* ↓	T(s)	AD↓	AG↑	AI↑	*I* ↑	*D* ↓	T(s)
Fake-CAM	0.8	1.6	46.0	50.7	28.1	0.00	0.5	0.6	42.6	46.1	26.9	0.00
Grad-CAM	12.2	17.6	44.4	66.3	14.7	0.03	14.2	14.7	40.6	**64.1**	11.6	0.02
Grad-CAM++	12.9	16.0	42.1	66.0	14.7	0.03	17.1	10.2	33.4	62.9	12.2	0.02
Score-CAM	8.6	26.6	56.7	65.7	16.3	15.22	13.5	15.6	41.7	62.5	12.1	3.11
Ablation-CAM	12.5	16.4	42.8	65.9	14.6	18.26	15.5	12.6	36.9	63.8	11.4	2.98
XGrad-CAM	12.2	17.6	44.4	66.3	14.7	0.03	13.8	14.8	41.2	**64.1**	11.7	0.02
Layer-CAM	15.6	15.0	38.8	67.0	**14.2**	0.08	48.9	3.1	13.5	58.3	**6.4**	0.07
ExPerturbation	38.1	9.5	22.5	**70.7**	15.0	152.96	43.0	7.1	20.5	61.1	15.0	83.20
Opti-CAM	**1.5**	**68.8**	**92.8**	62.0	19.7	4.15	**1.3**	**71.2**	**92.7**	59.2	11.0	3.94

Results
Quantitative Evaluation

Classification Experiments:

Method	ResNet50						VGG16					
	AD↓	AG↑	AI↑	*I*↑	*D*↓	T(s)	AD↓	AG↑	AI↑	*I*↑	*D*↓	T(s)
Fake-CAM	0.8	1.6	46.0	50.7	28.1	0.00	0.5	0.6	42.6	46.1	26.9	0.00
Grad-CAM	12.2	17.6	44.4	66.3	14.7	0.03	14.2	14.7	40.6	**64.1**	11.6	0.02
Grad-CAM++	12.9	16.0	42.1	66.0	14.7	0.03	17.1	10.2	33.4	62.9	12.2	0.02
Score-CAM	8.6	26.6	56.7	65.7	16.3	15.22	13.5	15.6	41.7	62.5	12.1	3.11
Ablation-CAM	12.5	16.4	42.8	65.9	14.6	18.26	15.5	12.6	36.9	63.8	11.4	2.98
XGrad-CAM	12.2	17.6	44.4	66.3	14.7	0.03	13.8	14.8	41.2	**64.1**	11.7	0.02
Layer-CAM	15.6	15.0	38.8	67.0	**14.2**	0.08	48.9	3.1	13.5	58.3	**6.4**	0.07
ExPerturbation	38.1	9.5	22.5	**70.7**	15.0	152.96	43.0	7.1	20.5	61.1	15.0	83.20
Opti-CAM	**1.5**	**68.8**	**92.8**	62.0	19.7	4.15	**1.3**	**71.2**	**92.7**	59.2	11.0	3.94

The optimization objective is well aligned with AD, AI, AG.
Insertion and Deletion are better suited for sparse saliency maps.

Results

Quantitative Evaluation

Localization Experiments:

METHOD	RESNET50							VGG16						
	OM↓	LE↓	F1↑	BA↑	SP↑	EP↑	SM↓	OM↓	LE↓	F1↑	BA↑	SP↑	EP↑	SM↓
Fake-CAM	63.6	54.0	57.7	47.9	99.8	28.5	0.98	64.7	54.0	57.7	47.9	99.8	28.5	1.07
Grad-CAM	72.9	65.8	49.8	**56.2**	69.8	33.3	1.30	71.1	62.3	42.0	54.2	64.8	32.0	1.39
Grad-CAM++	73.1	66.1	**50.4**	**56.2**	69.9	33.1	1.29	70.8	61.9	44.3	55.2	66.2	32.3	1.38
Score-CAM	**72.2**	64.9	49.6	54.5	68.7	32.4	**1.25**	71.2	62.5	**45.3**	**58.5**	**68.2**	33.4	1.40
Ablation-CAM	72.8	65.7	50.2	56.1	69.9	33.1	1.26	71.3	62.6	43.2	56.2	65.7	32.7	1.39
XGrad-CAM	72.9	65.8	49.8	**56.2**	69.8	33.3	1.30	70.8	62.0	41.9	53.5	64.4	31.6	1.41
Layer-CAM	73.1	66.0	50.1	55.5	**70.0**	33.0	1.29	70.5	61.5	28.0	54.7	65.0	32.4	1.45
ExPerturbation	73.6	66.6	37.5	44.2	64.8	**38.2**	1.59	74.1	66.4	37.8	43.3	62.7	**36.1**	1.74
Opti-CAM	**72.2**	**64.8**	47.3	49.2	59.4	30.5	1.34	**69.1**	**59.9**	44.1	51.2	61.4	30.7	**1.34**

Results

Quantitative Evaluation

Localization Experiments:

METHOD	RESNET50							VGG16						
	OM↓	LE↓	F1↑	BA↑	SP↑	EP↑	SM↓	OM↓	LE↓	F1↑	BA↑	SP↑	EP↑	SM↓
Fake-CAM	63.6	54.0	57.7	47.9	99.8	28.5	0.98	64.7	54.0	57.7	47.9	99.8	28.5	1.07
Grad-CAM	72.9	65.8	49.8	**56.2**	69.8	33.3	1.30	71.1	62.3	42.0	54.2	64.8	32.0	1.39
Grad-CAM++	73.1	66.1	**50.4**	**56.2**	69.9	33.1	1.29	70.8	61.9	44.3	55.2	66.2	32.3	1.38
Score-CAM	**72.2**	64.9	49.6	54.5	68.7	32.4	**1.25**	71.2	62.5	**45.3**	**58.5**	**68.2**	33.4	1.40
Ablation-CAM	72.8	65.7	50.2	56.1	69.9	33.1	1.26	71.3	62.6	43.2	56.2	65.7	32.7	1.39
XGrad-CAM	72.9	65.8	49.8	**56.2**	69.8	33.3	1.30	70.8	62.0	41.9	53.5	64.4	31.6	1.41
Layer-CAM	73.1	66.0	50.1	55.5	**70.0**	33.0	1.29	70.5	61.5	28.0	54.7	65.0	32.4	1.45
ExPerturbation	73.6	66.6	37.5	44.2	64.8	**38.2**	1.59	74.1	66.4	37.8	43.3	62.7	**36.1**	1.74
Opti-CAM	**72.2**	**64.8**	47.3	49.2	59.4	30.5	1.34	**69.1**	**59.9**	44.1	51.2	61.4	30.7	**1.34**

Localization is hard. It acts more aligned to human interpretations than model ones.

Conclusions

- Classifier specific saliency maps. The probability of prediction of a class is the highest.
- No regularization needed unlike [10, 11]. Optimization only over feature dimensions.
- Localization is difficult for Opti-CAM. It is mostly human based.
- Average Gain complements the evaluation procedure.
- Published at Computer Vision and Image Understanding (CVIU) 2024.

Table of Contents

Preliminaries

Premise

Is it possible to incorporate transformers ideas onto CNNs, enhancing the predictive power of CAM explanations?

$p = 0.7623$

$p = 0.7145$

Preliminaries

Premise

Is it possible to incorporate transformers ideas onto CNNs, enhancing the predictive power of CAM explanations?

$p = 0.7623$

$p = 0.7145$

$p = 0.7478$

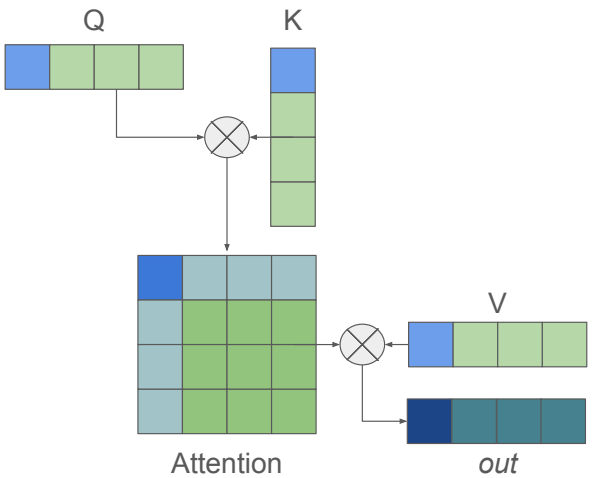
Background
Self Attention

Self-attention is the building block of Transformers:

Background
Self Attention

Self-attention is the building block of Transformers:

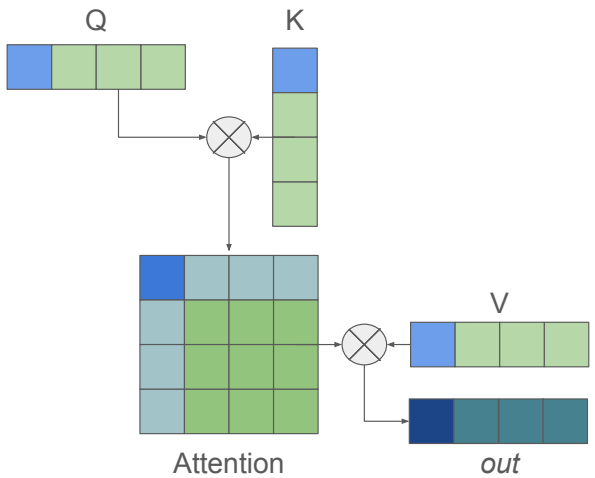
$$\text{Self Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_\ell}}\right)V \tag{5}$$



Background
Self Attention

Self-attention is the building block of Transformers:

$$\text{Self Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_\ell}}\right) V \tag{5}$$



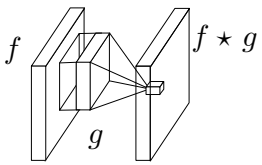
An abstract class representation ([CLS]) updated with information from the embedding. [15]

Background
Convolutions and Self Attention

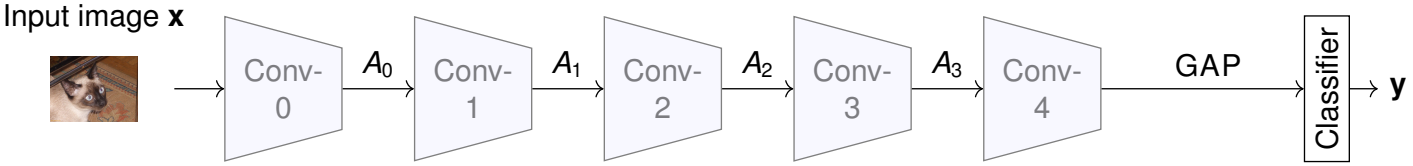
Different building blocks rely on different pooling mechanisms:

Background
Convolutions and Self Attention

Different building blocks rely on different pooling mechanisms:

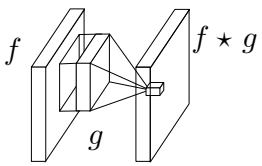


Original from [16]

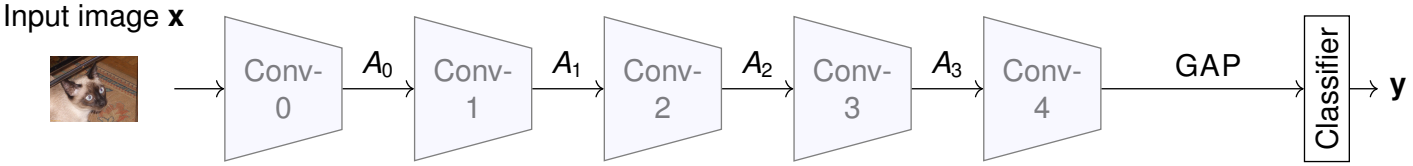


Background
Convolutions and Self Attention

Different building blocks rely on different pooling mechanisms:



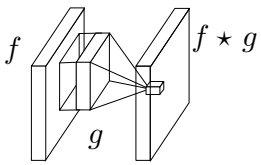
Original from [16]



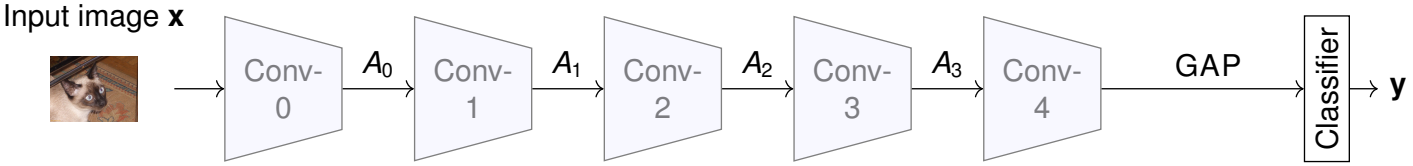
GAP considers an uniform pooling.

Background
Convolutions and Self Attention

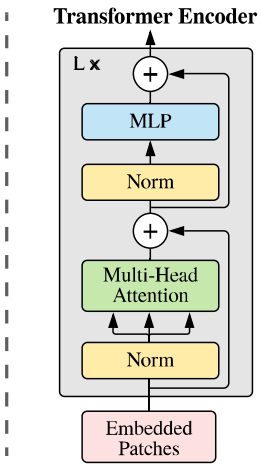
Different building blocks rely on different pooling mechanisms:



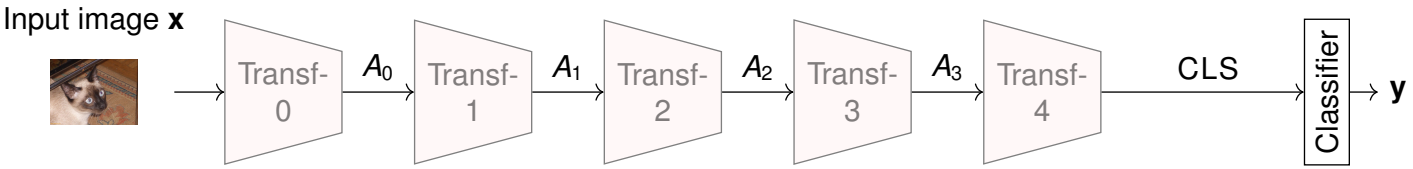
Original from [16]



GAP considers an uniform pooling.

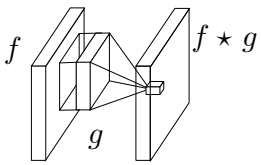


Original from [17]

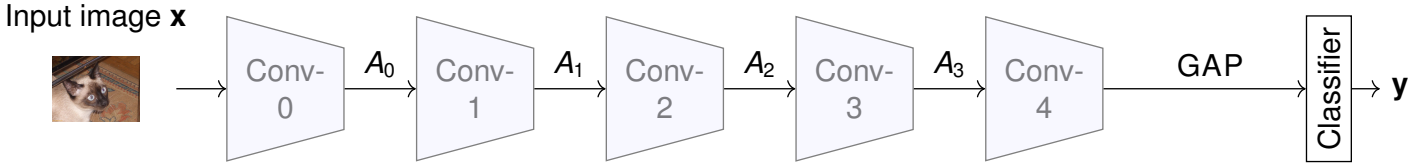


Background
Convolutions and Self Attention

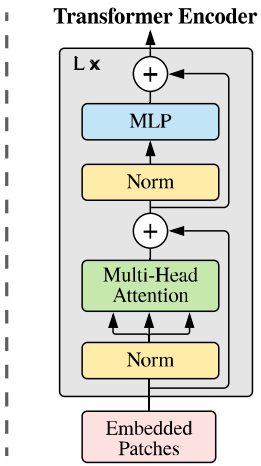
Different building blocks rely on different pooling mechanisms:



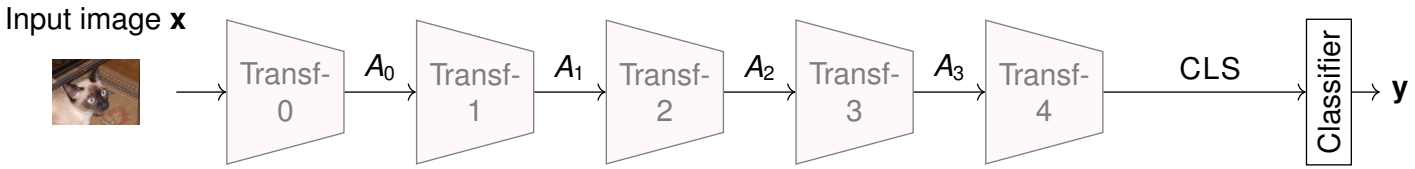
Original from [16]



GAP considers an uniform pooling.



Original from [17]



CLS is weighted with attention.

Motivation

Including CLS in CNNs

Cross Attention: Assuming an arbitrary embedding $\mathbf{q}$:

$$\mathrm{CA}_\ell(\mathbf{q}_\ell, A_\ell) := A^\top \mathbf{q} = A_\ell^\top h_\ell(A_\ell \mathbf{q}_\ell), \in \mathbb{R}_\ell^d \tag{6}$$

Motivation
Including CLS in CNNs

Cross Attention: Assuming an arbitrary embedding $\mathbf{q}$:

$$\mathrm{CA}_\ell(\mathbf{q}_\ell, A_\ell) := A^\top \mathbf{q} = A_\ell^\top h_\ell(A_\ell \mathbf{q}_\ell), \in \mathbb{R}_\ell^d \tag{6}$$

For a feature map A_ℓ, a pooled representation can be expressed via attention vector $\mathbf{a} = h_\ell(A_\ell \mathbf{q}_\ell)$:

$$\mathrm{CA}_\ell(\mathbf{q}_\ell, A_\ell) := A_\ell^\top \mathbf{a} \tag{7}$$

Motivation

Including CLS in CNNs

Cross Attention: Assuming an arbitrary embedding $\mathbf{q}$:

$$\mathrm{CA}_\ell(\mathbf{q}_\ell, A_\ell) := A^\top \mathbf{q} = A_\ell^\top h_\ell(A_\ell \mathbf{q}_\ell), \in \mathbb{R}_\ell^d \tag{6}$$

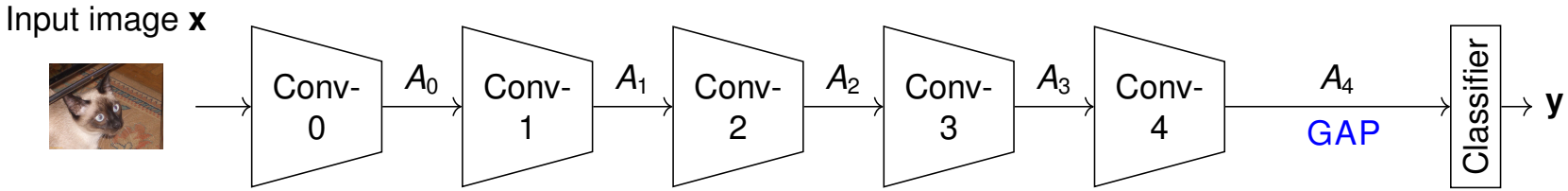
For a feature map A_ℓ, a pooled representation can be expressed via attention vector $\mathbf{a} = h_\ell(A_\ell \mathbf{q}_\ell)$:

$$\mathrm{CA}_\ell(\mathbf{q}_\ell, A_\ell) := A_\ell^\top \mathbf{a} \tag{7}$$

A pooling similar to GAP, masked in a fashion similar to CAM.

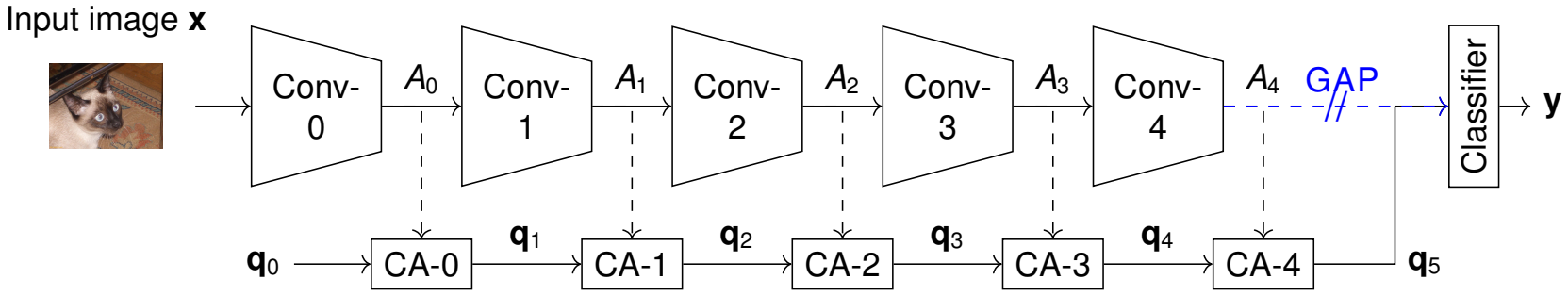
Cross Attention Stream

Given a model f we consider locations where critical operations take place:



Cross Attention Stream

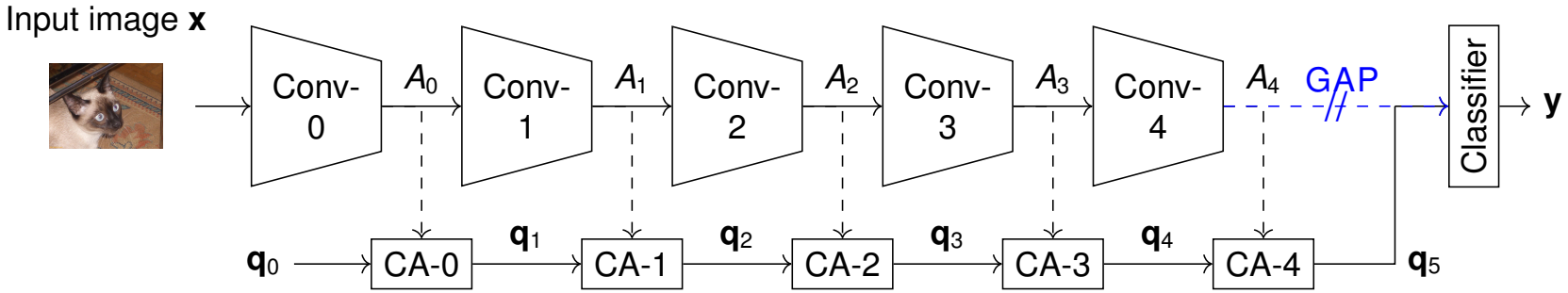
Given a model f we consider locations where critical operations take place:



$$f = g \circ \text{GAP} \circ f_L \circ \dots f_0 \tag{8}$$

Cross Attention Stream

Given a model f we consider locations where critical operations take place:

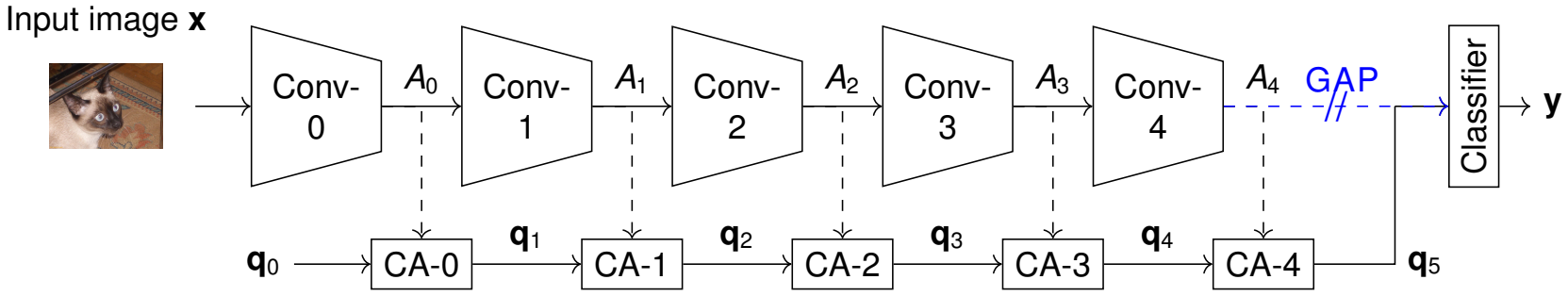


$$f = g \circ \text{GAP} \circ f_L \circ \dots f_0 \tag{8}$$

A [CLS] token embedding is initialized as $\mathbf{q}_0 \in \mathbb{R}^{d_0}$.

Cross Attention Stream

Given a model f we consider locations where critical operations take place:



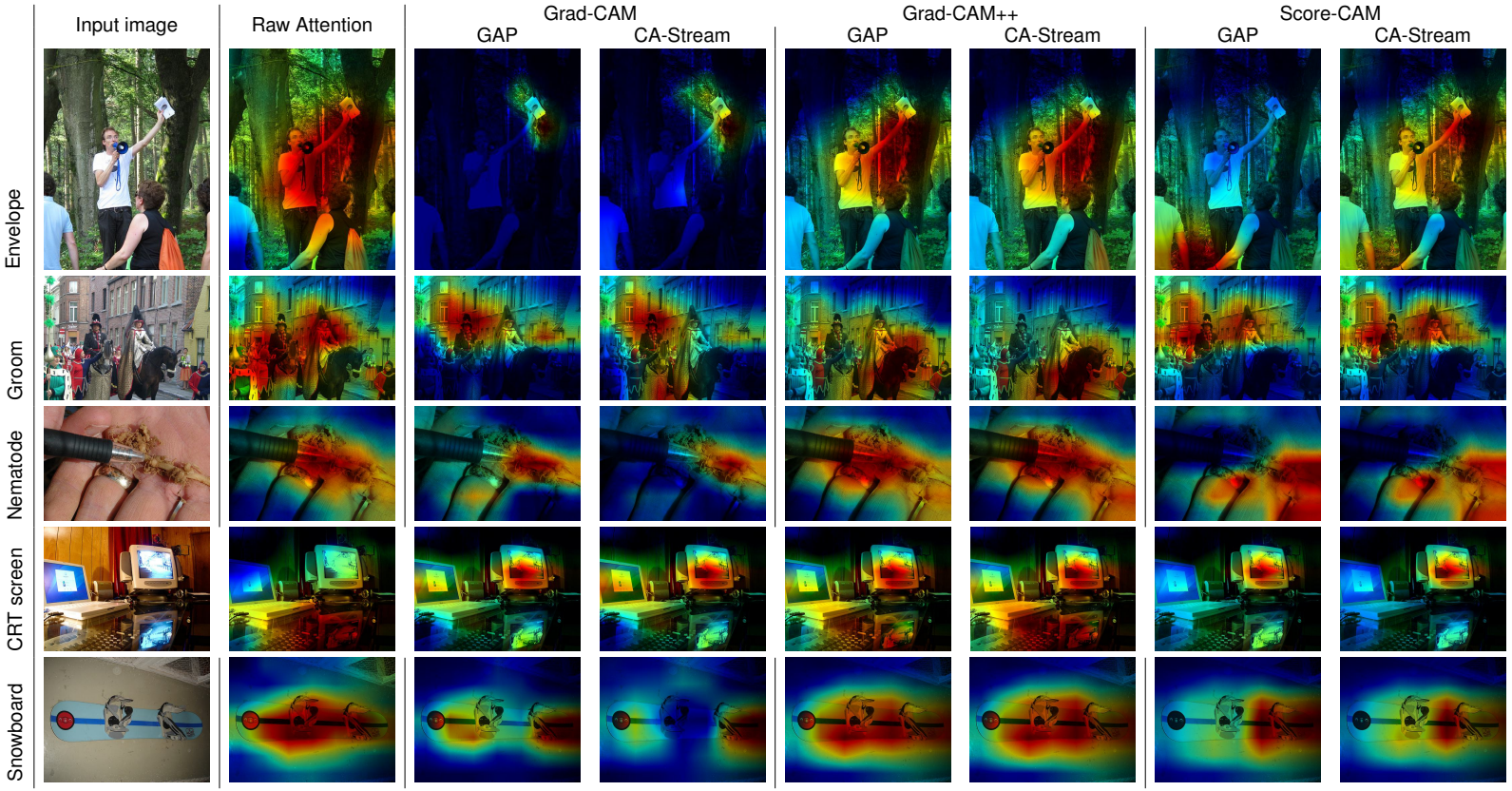
$$f = g \circ \mathrm{GAP} \circ f_L \circ \ldots f_0 \tag{8}$$

A [CLS] token embedding is initialized as $\mathbf{q}_0 \in \mathbb{R}^{d_0}$.
Updated with a secquence of embeddings $\mathbf{q}_\ell \in \mathbb{R}^{d_\ell}$,
interacting along a stream with features on stage ℓ:

$$\mathbf{q}_{\ell+1} = W_\ell \cdot \mathrm{CA}_\ell(\mathbf{q}_\ell, A_\ell) \tag{9}$$

Results

Qualitative Experiments



Results
Quantitative Experiments

Recognition results

Accuracy and Parameters					
Network	Pool	GFLOPs	#Param	Param%	Acc↑
ResNet-18	GAP	3.648	11.69M	3.71	67.28
	CA	3.652	12.13M		**67.54**
ResNet-50	GAP	8.268	25.56M	27.27	74.55
	CA	8.288	32.53M		**74.70**
ConvNeXt-S	GAP	17.395	50.22M	1.95	**83.26**
	CA	17.400	51.20M		83.14
ConvNeXt-B	GAP	30.747	88.59M	1.96	**83.72**
	CA	30.753	90.33M		83.51

CA (Ours) Maintains recognition properties.

Results

Quantitative Experiments

Network	Method	Pool	AD↓	AG↑	AI↑	I↑	D↓
ResNet-50	Grad-CAM	GAP	13.04	17.56	44.47	72.57	**13.24**
		CA	**12.54**	**22.67**	**48.56**	**75.53**	13.50
	Grad-CAM++	GAP	**13.79**	15.87	42.08	72.32	**13.33**
		CA	13.99	**19.29**	**44.60**	**75.21**	13.78
	Score-CAM	GAP	8.83	17.97	48.46	71.99	**14.31**
		CA	**7.09**	**23.65**	**54.20**	**74.91**	14.68
ConvNeXt-B	Grad-CAM	GAP	33.72	2.43	15.25	52.85	**29.57**
		CA	**19.45**	**13.96**	**32.89**	**86.38**	45.29
	Grad-CAM++	GAP	**34.01**	2.37	15.60	52.83	**29.17**
		CA	36.69	**8.00**	**21.95**	**85.39**	53.42
	Score-CAM	GAP	43.55	2.23	15.67	50.96	**39.49**
		CA	**23.51**	**11.04**	**27.35**	**83.41**	60.53

Results

Quantitative Experiments

Network	Method	Pool	AD↓	AG↑	AI↑	I↑	D↓
ResNet-50	Grad-CAM	GAP	13.04	17.56	44.47	72.57	**13.24**
		CA	**12.54**	**22.67**	**48.56**	**75.53**	13.50
	Grad-CAM++	GAP	**13.79**	15.87	42.08	72.32	**13.33**
		CA	13.99	**19.29**	**44.60**	**75.21**	13.78
	Score-CAM	GAP	8.83	17.97	48.46	71.99	**14.31**
		CA	**7.09**	**23.65**	**54.20**	**74.91**	14.68
ConvNeXt-B	Grad-CAM	GAP	33.72	2.43	15.25	52.85	**29.57**
		CA	**19.45**	**13.96**	**32.89**	**86.38**	45.29
	Grad-CAM++	GAP	**34.01**	2.37	15.60	52.83	**29.17**
		CA	36.69	**8.00**	**21.95**	**85.39**	53.42
	Score-CAM	GAP	43.55	2.23	15.67	50.96	**39.49**
		CA	**23.51**	**11.04**	**27.35**	**83.41**	60.53

CA (Ours) obtains stronger representations for interpretability.
Deletion poses issues due OOD data.

Conclusions

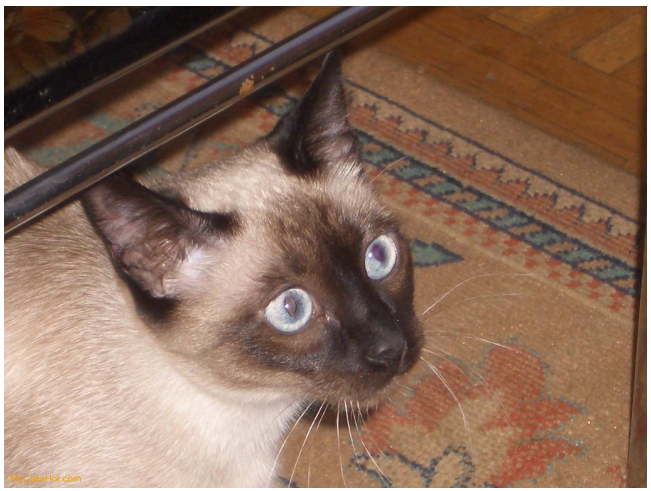
- CAM behaves like Attention. Attention on final layers ought to behave like CAM.
- Pooling matters. Smarter pooling yields improved interpretations, maintaining recognition performance.
- Deletion requires a revisit. Zeros on images still provide a signal.
- Published in the XAI Workshop at CVPR 2024.

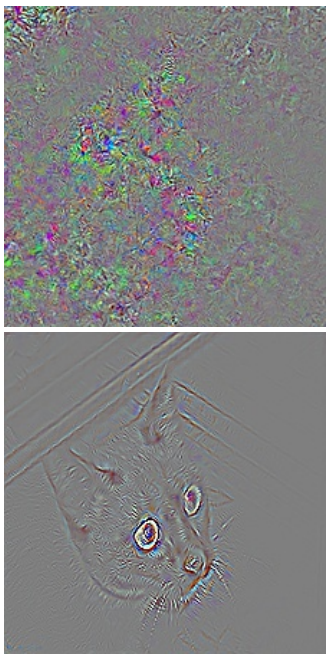
Table of Contents

Preliminaries

Premise

Can smoothing gradients during training, improve interpretability properties?





Background
Leveraging Backpropagation

The response (y) of a model (f), after forwarding an input x can be visualized on the input space, via backpropagation:

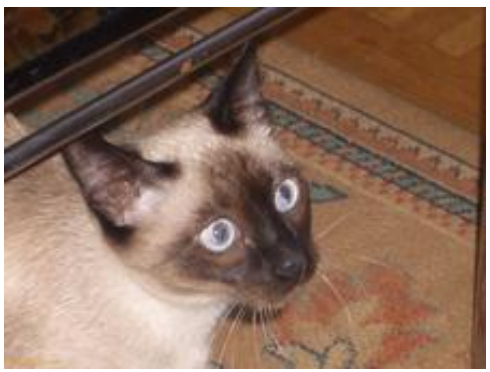
$$\frac{\partial x}{\partial y} = \frac{\partial x}{\partial F_0}\frac{\partial F_0}{\partial F_1} \cdot \cdots \cdot \frac{\partial F_{-1}}{\partial y} \tag{10}$$

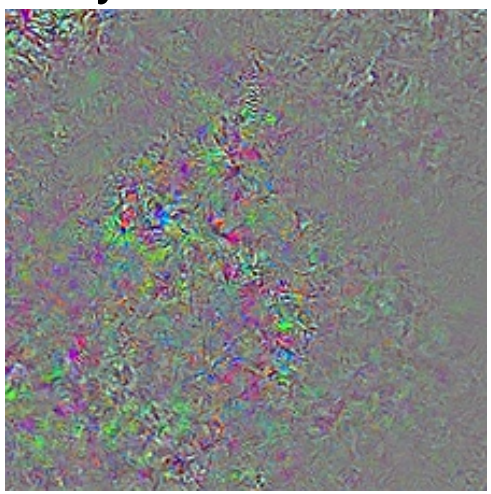
Background
Leveraging Backpropagation

The response (y) of a model (f), after forwarding an input x can be visualized on the input space, via backpropagation:

$$\frac{\partial x}{\partial y} = \frac{\partial x}{\partial F_0}\frac{\partial F_0}{\partial F_1} \cdot \cdots \cdot \frac{\partial F_{-1}}{\partial y} \tag{10}$$

However, due to non-linearities it is noisy:

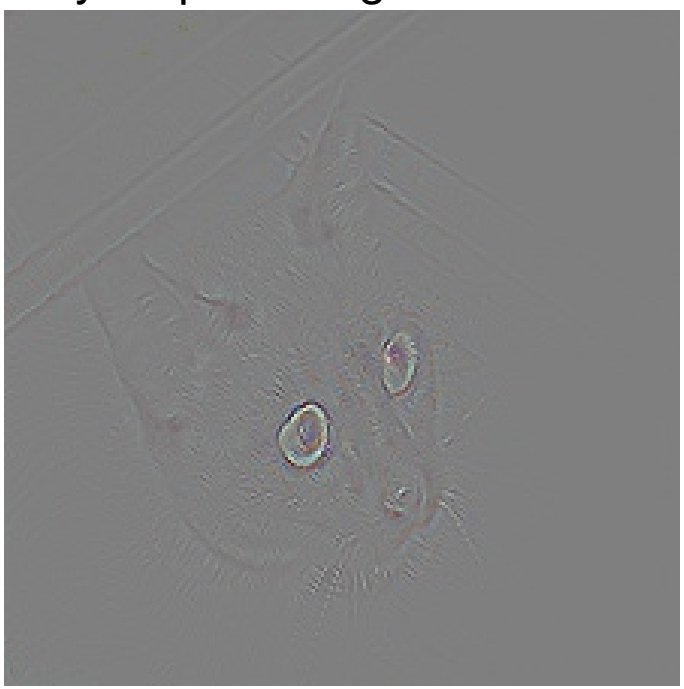




Background
Soft Gradients

Soft gradients can be achieved by two approaches:

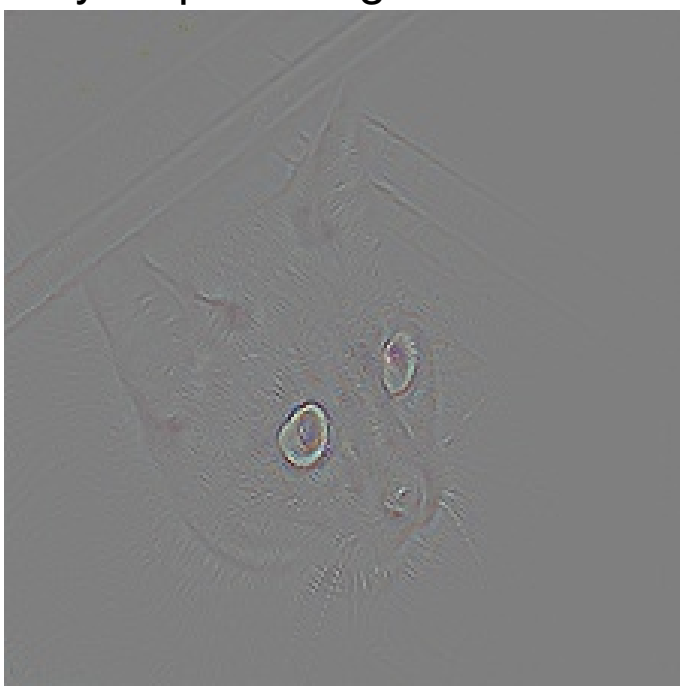
Guided Backpropagation [5]
Achieved by constraining backprop to only let positive gradients flow:



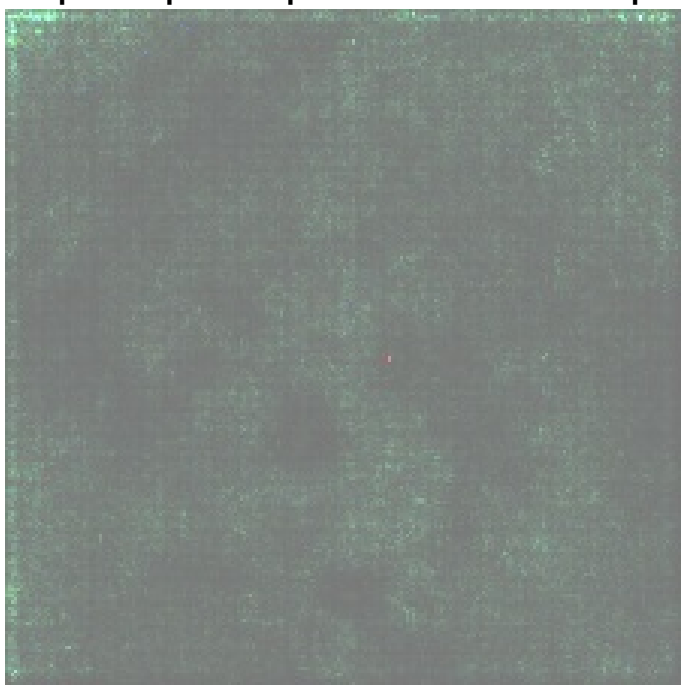
Background
Soft Gradients

Soft gradients can be achieved by two approaches:

Guided Backpropagation [5]
Achieved by constraining backprop to only let positive gradients flow:



Smoothgrad [6]
Achieved by iteratively adding noise to the input space prior to forward pass:



Interpretable Gradients

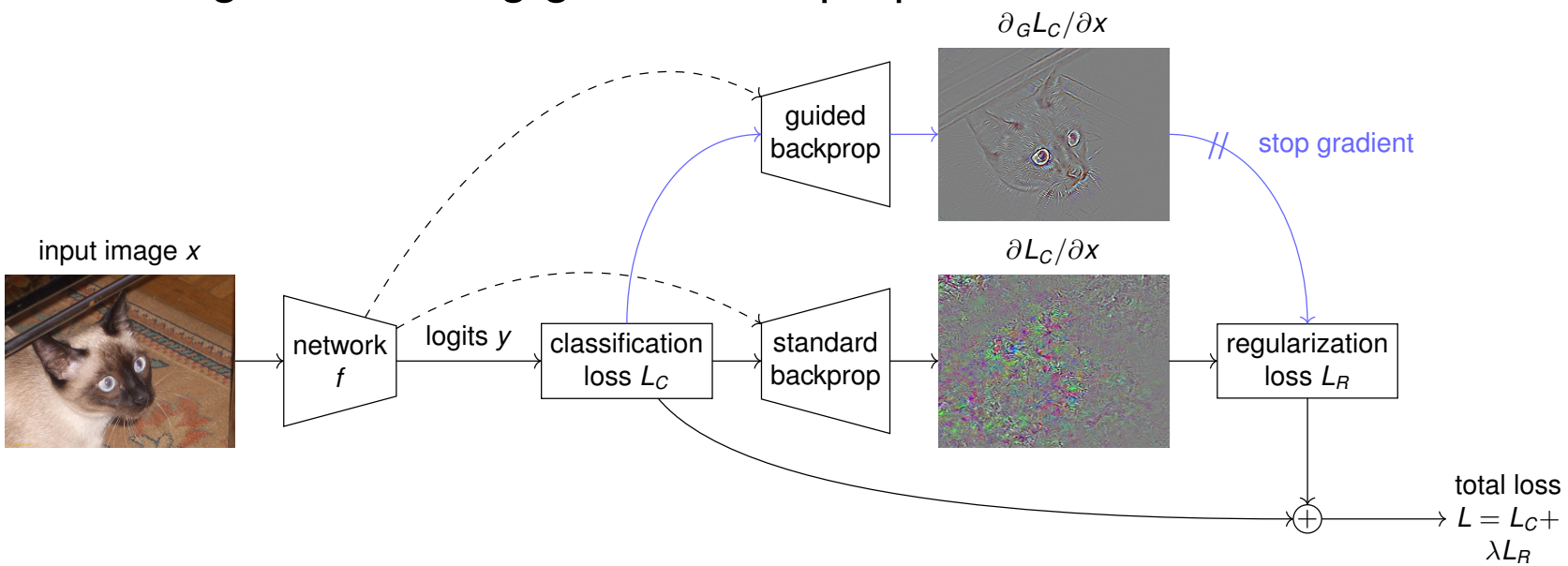
Formalized

Goal: Denoise the gradient during training using guided backprop.
How? Regularize using guided backprop.

Interpretable Gradients

Formalized

Goal: Denoise the gradient during training using guided backprop.
How? Regularize using guided backprop.



Interpretable Gradients

Smooth Gradient as Regularization

Assuming target labels T:

$$L_R(X, \theta, T) = \frac{1}{n} \sum_{i=1}^{n} E(\delta x_i, \delta_G x_i), \tag{11}$$

Interpretable Gradients

Smooth Gradient as Regularization

Assuming target labels T:

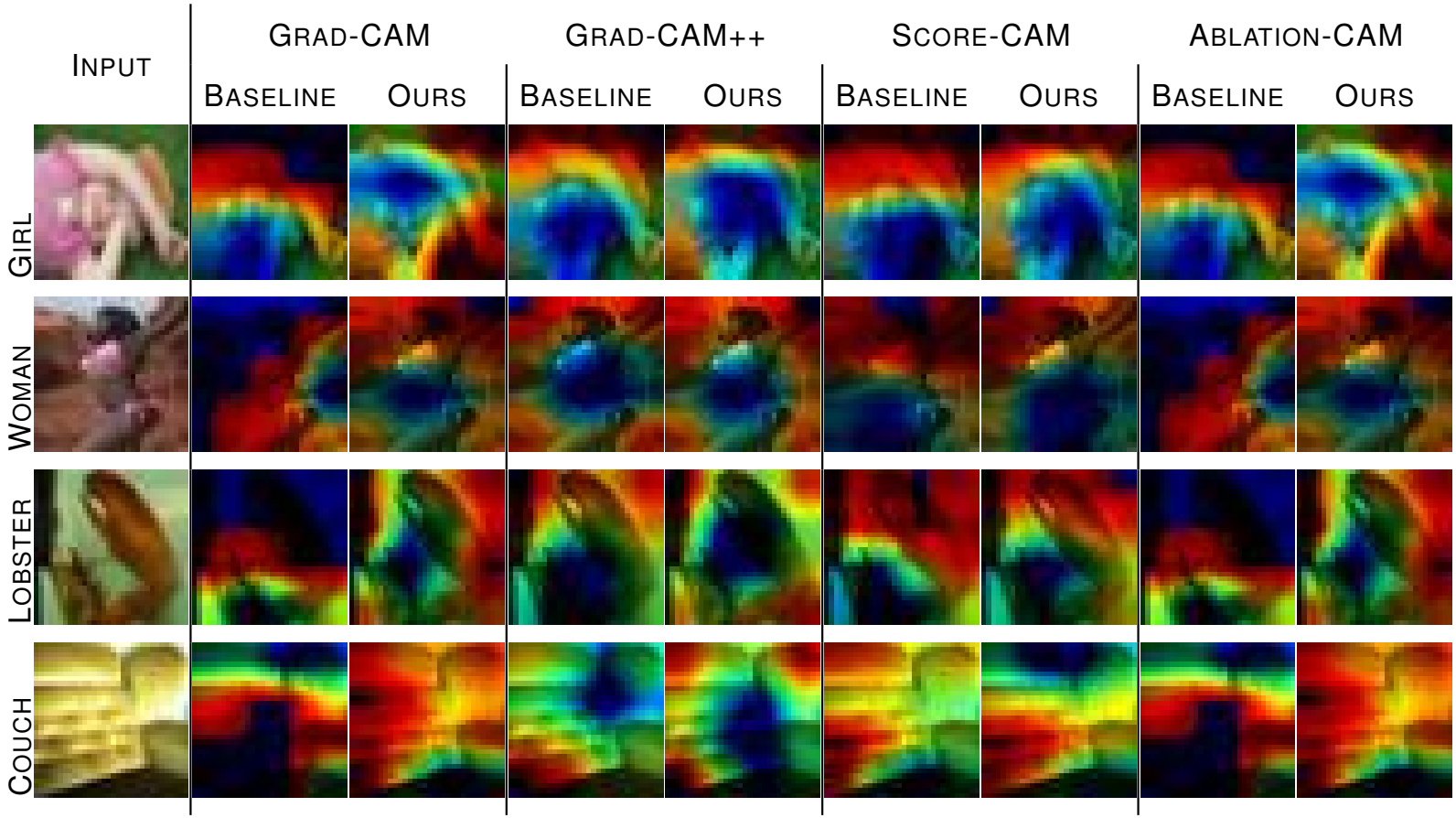
$$L_R(X, \theta, T) = \frac{1}{n} \sum_{i=1}^{n} E(\delta x_i, \delta_G x_i), \tag{11}$$

And the total loss:

$$L(x, \theta, t) = L_C(x, \theta, t) + \lambda L_R(x, \theta, t), \tag{12}$$

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

INTERPRETABLE RECOGNITION METRICS						
RESNET-18						
METHOD	L_R	AD↓	AG↑	AI↑	INS↑	DEL↓
GRAD-CAM	-	30.16	15.23	29.99	58.47	**17.47**
	OURS	**28.09**	**16.19**	**31.53**	**58.76**	17.57
GRAD-CAM++	-	31.40	14.17	28.47	58.61	**17.05**
	OURS	**29.78**	**15.07**	**29.60**	**58.90**	17.22
SCORE-CAM	-	26.49	18.62	33.84	58.42	**18.31**
	OURS	**24.82**	**19.49**	**35.51**	**59.11**	18.34
ABLATION-CAM	-	31.96	14.02	28.33	58.36	**17.14**
	OURS	**29.90**	**15.03**	**29.61**	**58.70**	17.37
AXIOM-CAM	-	30.16	15.23	29.98	58.47	**17.47**
	OURS	**28.09**	**16.20**	**31.53**	**58.76**	17.57

Results
Quantitative Experiments

Interpretability metrics

INTERPRETABLE RECOGNITION METRICS						
RESNET-18						
METHOD	L_R	AD↓	AG↑	AI↑	INS↑	DEL↓
GRAD-CAM	-	30.16	15.23	29.99	58.47	**17.47**
	OURS	**28.09**	**16.19**	**31.53**	**58.76**	17.57
GRAD-CAM++	-	31.40	14.17	28.47	58.61	**17.05**
	OURS	**29.78**	**15.07**	**29.60**	**58.90**	17.22
SCORE-CAM	-	26.49	18.62	33.84	58.42	**18.31**
	OURS	**24.82**	**19.49**	**35.51**	**59.11**	18.34
ABLATION-CAM	-	31.96	14.02	28.33	58.36	**17.14**
	OURS	**29.90**	**15.03**	**29.61**	**58.70**	17.37
AXIOM-CAM	-	30.16	15.23	29.98	58.47	**17.47**
	OURS	**28.09**	**16.20**	**31.53**	**58.76**	17.57

Interpretable properties are enhanced. Deletion still poses an issue.

Conclusions

- Denoising CNN gradients improve interpretability properties, maintining recognition.
- Negative gradients contribute to stability during training.
- Resolution matters: small inputs require shallow layer targeting.
- Published in VISAPP 2024.

Table of Contents

Closing Remarks
Conclusion

General Conclusions:

- Context is important for a prediction. Salient information is found in the background.
- Pooling matters. Small changes lead to enhanced interpretations, with little penalty to predictive power.
- Noisy gradients are a byproduct of training. Negative values act as regularizers.
- Insertion & Deletion are prolematic. OOD data affects the classifier.
- *For me?*

Closing Remarks

Conclusion

General Conclusions:

- Context is important for a prediction. Salient information is found in the background.
- Pooling matters. Small changes lead to enhanced interpretations, with little penalty to predictive power.
- Noisy gradients are a byproduct of training. Negative values act as regularizers.
- Insertion & Deletion are prolematic. OOD data affects the classifier.
- *For me?* Interpretability is a challenging field, I like challenges as hard as they can be.
- *For me?*

Closing Remarks

Conclusion

General Conclusions:

- Context is important for a prediction. Salient information is found in the background.
- Pooling matters. Small changes lead to enhanced interpretations, with little penalty to predictive power.
- Noisy gradients are a byproduct of training. Negative values act as regularizers.
- Insertion & Deletion are prolematic. OOD data affects the classifier.
- *For me?* Interpretability is a challenging field, I like challenges as hard as they can be.
- *For me?* Collaboration is paramount. Discussing and complementing ideas with Hanwei Zhang helped a lot.

Closing Remarks

Perspectives

Short Term

- Less complex approaches are desired. Continuation of gradient denoise is possible.
- A thorough reality check for *fair* comparison is desired across methods. Set the stage for future comparisons.
- Explanations should extend beyond *groundtruth*. The reality check could include such scenarios.
- *For me?*

Closing Remarks

Perspectives

Short Term

- Less complex approaches are desired. Continuation of gradient denoise is possible.
- A thorough reality check for *fair* comparison is desired across methods. Set the stage for future comparisons.
- Explanations should extend beyond *groundtruth*. The reality check could include such scenarios.
- *For me?* Expand my network of collaboration. The more people I share and work with, the better.
- *For me?*

Closing Remarks

Perspectives

Short Term

- Less complex approaches are desired. Continuation of gradient denoise is possible.
- A thorough reality check for *fair* comparison is desired across methods. Set the stage for future comparisons.
- Explanations should extend beyond *groundtruth*. The reality check could include such scenarios.
- *For me?* Expand my network of collaboration. The more people I share and work with, the better.
- *For me?* I want to pursue a career in research, either in academy or industry.

Long Term

Closing Remarks

Perspectives

- Deep models are growing in complexity. Designing approaches to interpret transformers is one option.
- Interpretability must be extended to high impact domains: Medicine, Surveillance, **Autonomous Vehicles**.

Table of Contents

References I


T. Madiega, “Artificial intelligence act,” *European Parliament: European Parliamentary Research Service*, 2021.


Z. C. Lipton, “The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.” *Queue*, vol. 16, no. 3, pp. 31–57, 2018.


Y. Zhang, P. Tiňo, A. Leonardis, and K. Tang, “A survey on neural network interpretability,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 5, pp. 726–742, 2021.


K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” *ICLR Workshop*, 2014.


J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for simplicity: The all convolutional net,” *arXiv preprint arXiv:1412.6806*, 2014.


D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, “Smoothgrad: removing noise by adding noise,” *arXiv preprint arXiv:1706.03825*, 2017.


M. T. Ribeiro, S. Singh, and C. Guestrin, “”why should i trust you?”: Explaining the predictions of any classifier,” in *SIGKDD*, ser. KDD ’16, 2016.


P. Vitali, D. Abir, and S. Kate, “Rise: Randomized input sampling for explanation of black-box models,” *BMVC*, 2018.


B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *CVPR*, 2016.


R. C. Fong and A. Vedaldi, “Interpretable explanations of black boxes by meaningful perturbation,” in *ICCV*, 2017.

References II


R. Fong, M. Patrick, and A. Vedaldi, "Understanding deep networks via extremal perturbations and smooth masks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2950–2958.


R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization," *CoRR*, vol. abs/1610.02391, 2016. [Online]. Available: http://arxiv.org/abs/1610.02391


A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *WACV*, 2018.


H. Wang, M. Du, F. Yang, and Z. Zhang, "Score-cam: Improved visual explanations via score-weighted class activation mapping," *CoRR*, vol. abs/1910.01279, 2019. [Online]. Available: http://arxiv.org/abs/1910.01279


J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.


M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.


A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.