Visual interpretability: saliency maps and interpretable classification

Ronan Sicre

LIS, Marseille - QARMA team

Overview

Introduction

Saliency maps for image classification interpretability
Opti-CAM: Optimizing saliency maps for interpretability
Hanwei Zhang, Felipe Torres, Ronan Sicre, Yannis Avrithis, Stephane Ayache

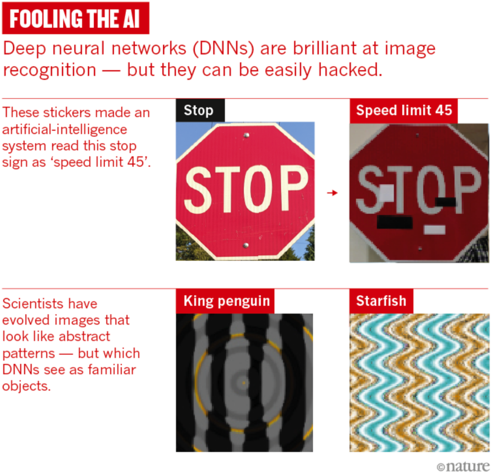
Interpretable image classification with parts
DP-Net: Learning Discriminative Parts for Image Recognition (ICIP 2023)
Ronan Sicre; Hanwei Zhang; Julien Dejasmin; Chiheb Daaloul; Stephane Ayache; Thierry Artières

Interpretability is important for high stakes decisions

Model understanding is absolutely critical in several domains -- particularly those involving *high stakes decisions*!

Building trust for users - Responsibility - Robustness

Interpretability is important for trustworthy DNNs



- Robustness and improvements
- Trust and understanding
- Security, legal necessity and responsibility

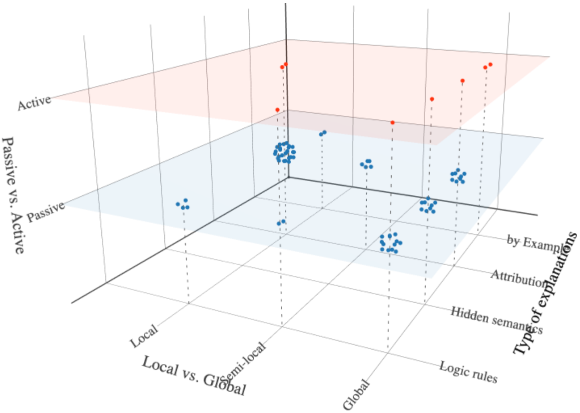
Dimensions of interpretability methods

The mythos of model interpretability... 2018
Transparency vs post-hoc interpretability

A survey on NN interpretability 2020

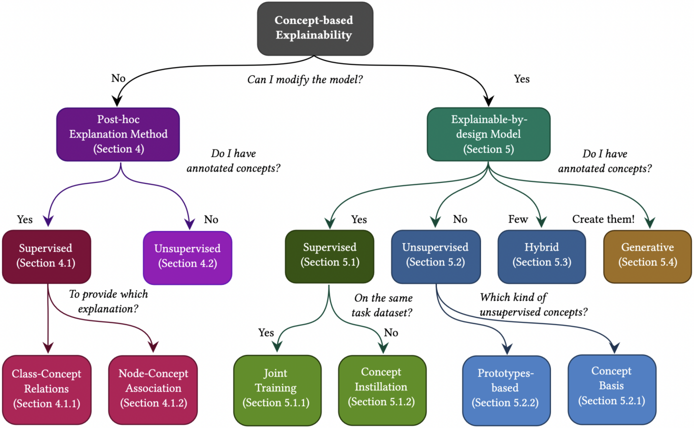
Dimension 1 — Passive vs. Active Approaches	
Passive	Post hoc explain trained neural networks
Active	Actively change the network architecture or training process for better interpretability
Dimension 2 — Type of Explanations (in the order of increasing explanatory power)	
To explain a prediction/class by	
Examples	Provide example(s) which may be considered similar or as prototype(s)
Attribution	Assign credit (or blame) to the input features (e.g. feature importance, saliency masks)
Hidden semantics	Make sense of certain hidden neurons/layers
Rules	Extract logic rules (e.g. decision trees, rule sets and other rule formats)
Dimension 3 — Local vs. Global Interpretability (in terms of the input space)	
Local	Explain network's *predictions on individual samples* (e.g. a saliency mask for an input image)
Semi-local	In between, for example, explain a group of similar inputs together
Global	Explain the network *as a whole* (e.g. a set of rules/a decision tree)

Dimensions of interpretability methods



Concept-based XAI

Concept-based Explainable Artificial Intelligence: A Survey 2023



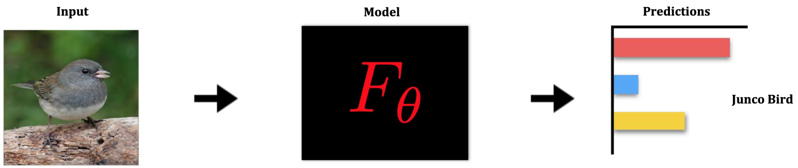
Post-hoc / Passive interpretability

LIME and SHAP: most common model agnostic approach

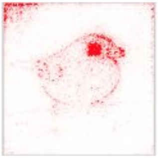
Image classification: methods specific to saliency maps

Ribeiro et al. "" Why should i trust you?" Explaining the predictions of any classifier." 2016.
Lundberg et al. "A unified approach to interpreting model predictions." 2017.

Saliency Map Overview



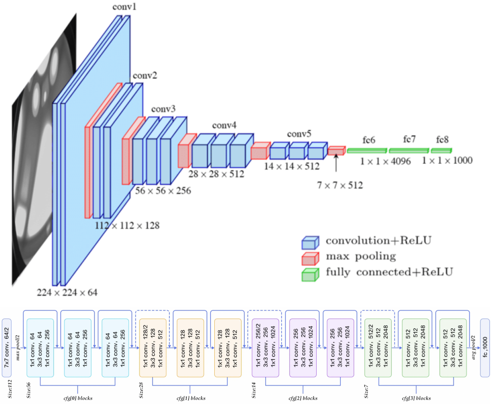
What parts of the input are most relevant for the model's prediction: 'Junco Bird'?



- **Feature Attribution**
- **'Saliency Map'**
- **Heatmap**

CNNs for image classification

CNN architecture of a VGG16 and a ResNet



https://vitalflux.com/different-types-of-cnn-architectures-explained-examples/
https://miro.medium.com/v2/resize:fit:2800/0*pkrso8DZa0m6IAcJ.png

Class activation maps (CAM)

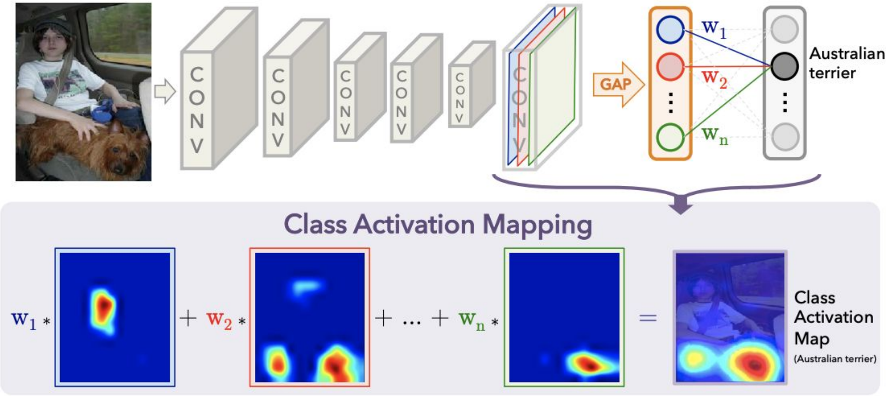


Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

Class activation maps (CAM)

CAM-based saliency maps

linear combination of feature maps $A_\ell^k = f_\ell^k(\mathbf{x})$.
For layer ℓ and class c, the saliency is

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

where w_k^c are the weights and h an activation function.

Grad-CAM

Grad-CAM

$$S_\ell^c(\mathbf{x}) := h\left(\sum_k w_k^c A_\ell^k\right), \tag{2}$$

$h = \mathrm{relu}$ and weights

$$w_k^c := \mathrm{GAP}\left(\frac{\partial y_c}{\partial A_\ell^k}\right), \tag{3}$$

where GAP is global average pooling and y_c is the logit.

Score-CAM

Score-CAM

$$S_\ell^c(\mathbf{x}) := h\left(\sum_k w_k^c A_\ell^k\right), \tag{4}$$

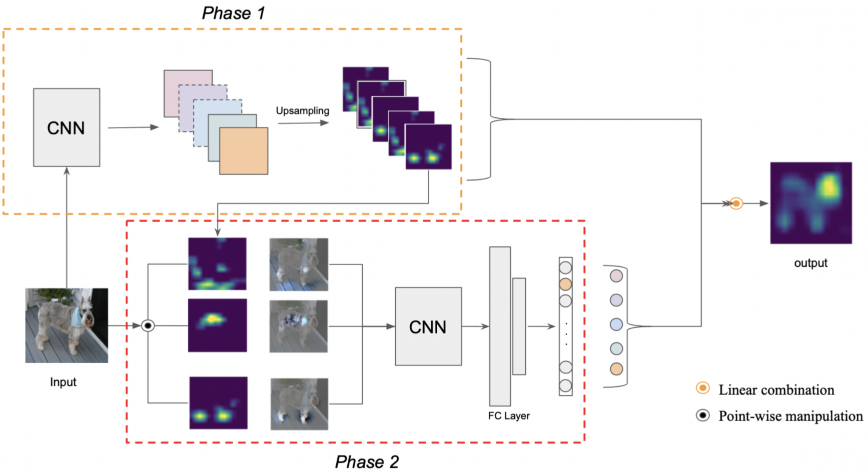
$h = \mathrm{relu}$ and weights $w_k^c := \mathrm{softmax}(\mathbf{u}^c)_k$,
where $\mathbf{u}^c$ is the increase in confidence for class c of the input image $\mathbf{x}$ masked by the saliency map:

$$u_k^c := f(\mathbf{x} \odot n(\mathrm{up}(A_\ell^k)))_c - f(\mathbf{x})_c, \tag{5}$$

$\odot$ is Hadamard product, up upsampling, n normalization.

Cons: requires as many forward as features.

ScoreCAM


Phase 1
CNN
Upsampling
output
Input
CNN
FC Layer
Linear combination
Point-wise manipulation
Phase 2

Masking-based methods

Masking-based methods: extremal perturbations

Optimization in the input space of a masking objective
Optimization per image like adversarial examples.

$$S^c(\mathbf{x}) := \arg\max_{\mathbf{m}\in\mathcal{M}} f(\mathbf{x} \odot n(\mathrm{up}(\mathbf{m})))_c + \lambda R(\mathbf{m}). \qquad (6)$$

A mask $\mathbf{m}$ is directly optimized without relying on feature maps.

Cons: the optimization is complex and requires regularization.

Fong et al: Understanding deep networks via extremal perturbations and smooth masks (2019)

Opti-CAM

Optimization of activation weights (CAM) of masking objective.
Optimization per image like adversarial examples.

$$S_\ell^c(\mathbf{x}) := h\left(\sum_k w_k^c A_\ell^k\right), \tag{7}$$

$w_k := \operatorname{softmax}(\mathbf{u})_k$, where $\mathbf{u}$ is the variable

$$S_\ell(\mathbf{x}; \mathbf{u}) := \sum_k \operatorname{softmax}(\mathbf{u})_k A_\ell^k. \tag{8}$$

Opti-CAM

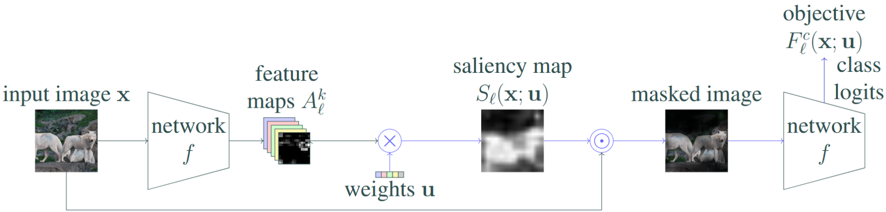
We find the vector $\mathbf{u}^*$ that maximizes the model prediction for class c,
when the input image $\mathbf{x}$ is masked by saliency map $S_\ell(\mathbf{x};\mathbf{u}^*)$:

$$\mathbf{u}^* := \arg\max_{\mathbf{u}} F_\ell^c(\mathbf{x};\mathbf{u}), \text{ where } F_\ell^c(\mathbf{x};\mathbf{u}) := f(\mathbf{x} \odot n(\mathrm{up}(S_\ell(\mathbf{x};\mathbf{u})))). \tag{9}$$

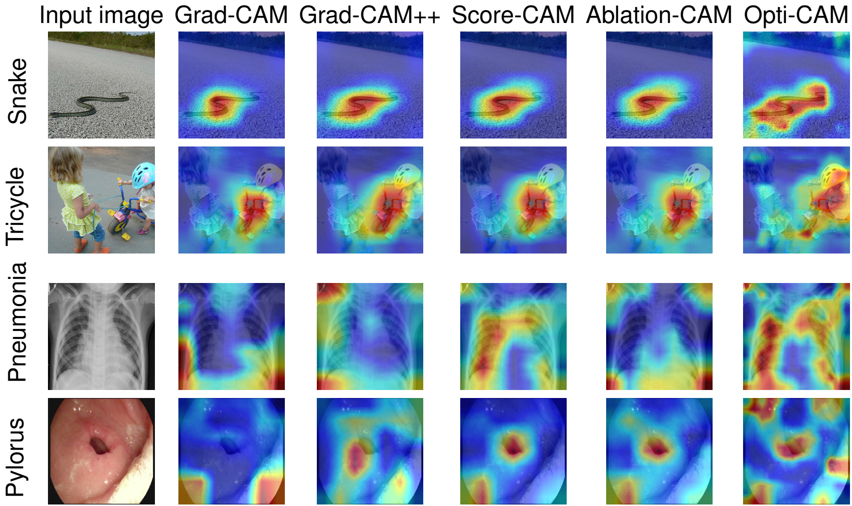
The saliency map $S_\ell(\mathbf{x};\mathbf{u})$ is upscaled and normalized.
Finally we have

$$S_\ell^c(\mathbf{x}) := S_\ell(\mathbf{x};\mathbf{u}^*) = S_\ell(\mathbf{x};\arg\max_{\mathbf{u}} F_\ell^c(\mathbf{x};\mathbf{u})), \tag{10}$$

Opti-CAM


input image $\mathbf{x}$
network f
feature maps A_ℓ^k
weights $\mathbf{u}$
saliency map $S_\ell(\mathbf{x}; \mathbf{u})$
masked image
network f
class logits
objective $F_\ell^c(\mathbf{x}; \mathbf{u})$

Visualizations



Saliency map evaluation

Recent field: No concensus, No good practice.

Faithfulness Evaluation: Average Drop, Average Increase (Increase in confidence), Average Gain.

Causal Metrics: Insertion, Deletion.

Weakly-Supervised Object Localization: Official Metric (OM), Localization Error (LE), Pixel-wise F_1 score (F1), Box Accuracy (BA), Standard Pointing game (SP), Energy Pointing game (EP).

Saliency map evaluation: Faithfulness

Average Drop (AD) how much predictive power is lost when masking .

$$AD(\%) = \sum_{i=1}^{N} \frac{max(0, Y_i^c - O_i^c)}{Y_i^c} \tag{11}$$

Average Gain (AG) how much gain in predictive power for the masked image.

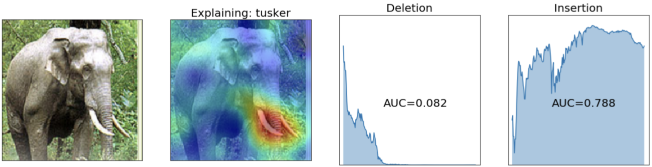
$$AG(\%) = \sum_{i=1}^{N} \frac{max(0, O_i^c - Y_i^c)}{Y_i^c} \tag{12}$$

Average Increase (AI) percentage of images where the masked image has a higher score.

$$AI(\%) = \frac{1}{N} \sum_{i}^{N} \mathbb{1}(Y_i^c < O_i^c) * 100 \tag{13}$$

Saliency map evaluation: Causal metrics

- **Insertion** starts from a blurry image and gradually insert the pixel ranked by saliency, At each iteration the images are passed through the network to compute the prediction ratio.
- **Deletion** gradually removes the most salient pixels. Removed pixels are replaced by black.



Opti-CAM results

Method	ResNet50			VGG16			ViT-B			ResNet50		VGG16	
	$AD\downarrow$	$AG\uparrow$	$AI\uparrow$	$AD\downarrow$	$AG\uparrow$	$AI\uparrow$	$AD\downarrow$	$AG\uparrow$	$AI\uparrow$	$I\uparrow$	$D\downarrow$	$I\uparrow$	$D\downarrow$
Fake-CAM	0.8	1.6	46.0	0.5	0.6	42.6	0.3	0.4	48.3	50.7	28.1	46.1	26.9
Grad-CAM	12.2	17.6	44.4	14.2	14.7	40.6	69.4	2.5	12.4	66.3	14.7	**64.1**	11.6
Grad-CAM++	12.9	16.0	42.1	17.1	10.2	33.4	86.3	1.5	1.0	66.0	14.7	62.9	12.2
Score-CAM	8.6	26.6	56.7	13.5	15.6	41.7	32.0	6.2	33.0	65.7	16.3	62.5	12.1
XGrad-CAM	12.2	17.6	44.4	13.8	14.8	41.2	88.1	0.4	4.3	66.3	14.7	**64.1**	11.7
Layer-CAM	15.6	15.0	38.8	48.9	3.1	13.5	82.0	0.2	2.9	67.0	**14.2**	58.3	**6.4**
ExPerturb.	38.1	9.5	22.5	43.0	7.1	20.5	28.8	6.2	24.4	**70.7**	15.0	61.1	15.0
Opti-CAM	**1.5**	**68.8**	**92.8**	**1.3**	**71.2**	**92.7**	**0.6**	**18.0**	**90.1**	62.0	19.7	59.2	11.0

AD, AG and AI are aligned with our optimization objective
I, D: OOD data, biased towards sparse saliency maps.

Opti-CAM results

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
ExPerturb	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**

Opit-CAM results

METHOD	AD↓			↑			AI↑		
	S	$B \cap S$	$S \setminus B$	S	$B \cap S$	$S \setminus B$	S	$B \cap S$	$S \setminus B$
$S := B$	67.2	–	–	2.3	–	–	9.2	–	–
$S := I \setminus B$	44.0	–	–	2.8	–	–	16.3	–	–
Fake-CAM	0.5	67.2	44.1	0.7	2.3	2.8	42.0	9.2	18.9
Grad-CAM	15.0	72.6	52.1	15.3	1.8	6.0	40.4	8.4	19.4
G-CAM++	16.5	72.9	53.1	10.6	1.6	4.1	35.2	7.3	17.1
Score-CAM	12.5	71.5	50.5	16.1	2.2	6.3	42.5	8.6	20.8
Abl-CAM	15.1	72.8	52.1	13.5	1.7	5.6	39.9	7.8	19.0
XGrad-CAM	14.3	72.6	51.4	15.1	1.8	6.0	42.1	8.0	20.1
Layer-CAM	49.2	84.2	74.4	2.7	0.4	1.2	12.7	4.4	7.3
ExPerturb.	43.8	81.6	71.0	7.1	1.4	3.2	18.9	5.6	11.1
Opti-CAM	**1.4**	**62.5**	**34.8**	**66.3**	**8.7**	**25.8**	**92.5**	**18.6**	**47.1**

Explanations and localization are two different tasks.

Opti-CAM conclusions

Evaluation: good practice, limitations of the metrics.

Improve saliency map methods for Transformers

Parts and prototypes

Prototype/Part based architectures:

Scene recognition with prototype-agnostic scene layout, 2019
This looks like that: deep learning for interpretable image recognition, 2019
Protopshare: Prototypical parts sharing... 2021
Neural prototype trees for interpretable fine-grained image reco. 2021
Interpretable image classification with differentiable prototypes... 2022
PIP-Net: Patch-Based Intuitive Prototypes for Interpretable... 2023

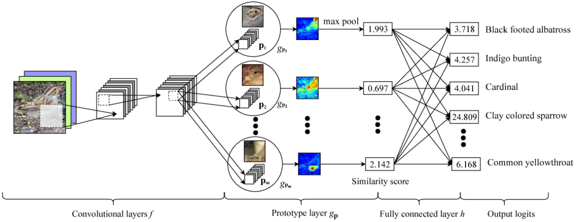
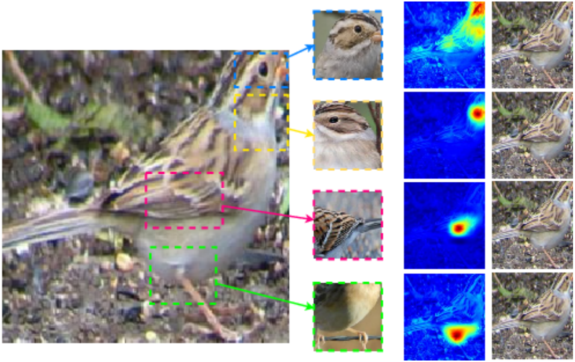


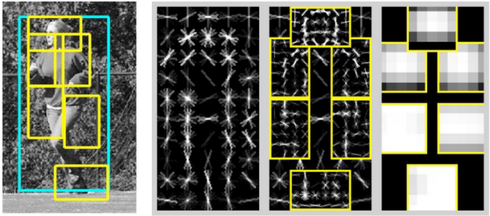
Figure 2. The network architecture.

Parts and prototypes



A bit of history

Deformable Part Models:
Object detection with discriminatively trained part-based models, 2010

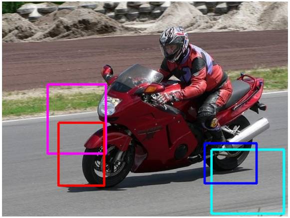


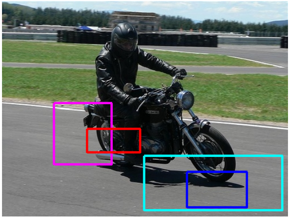
Blocks That Shout: Distinctive Parts for Scene Classification, 2013
Mid-level Visual Element Discovery as Discriminative Mode Seeking, 2013
Discriminative part model for visual recognition, 2014-2016
Automatic discovery and optimization of parts for image classif., 2014
No spare parts: Sharing part detectors for image categorization, 2016

Two-stage optimization with specific definition of parts and constraints.

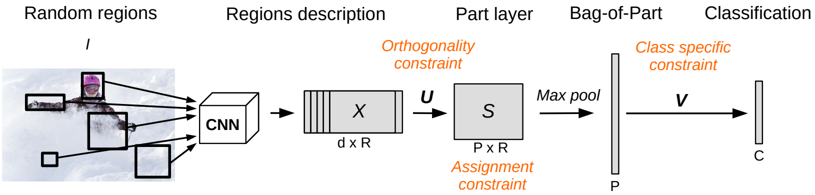
Part-based models: mid-level features

Learning a set of discriminative parts per class.
Detect parts in an image to produce a part-based description

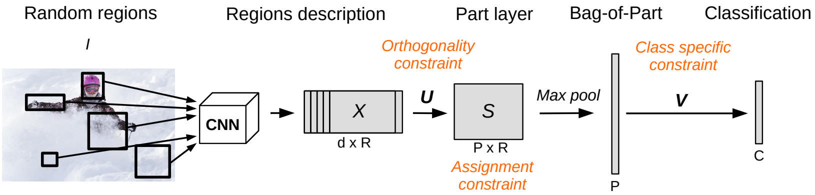




DP-Net: Discriminative Part Network

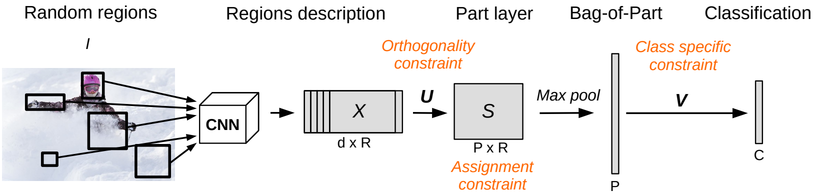


Part constraints



1) Parts should be complementary, *i.e*. parts should be different one from another.
2) Parts should cover as much as possible the diversity of regions extracted from images.
3) Parts should be discriminative with respect to classes.
4) Parts should be specific to categories.

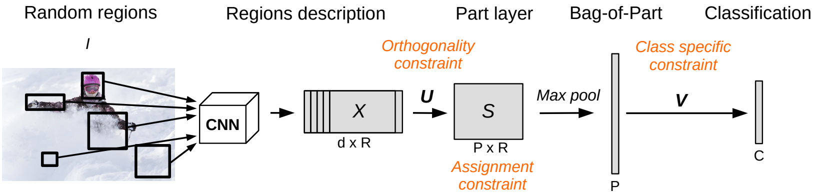
Part constraints



1) Parts should be complementary, *i.e.* parts should be different one from another.
2) Parts should cover as much as possible the diversity of regions extracted from images.
3) Parts should be discriminative with respect to classes.
4) Parts should be specific to categories.

Categorical Cross entropy loss

Part constraints

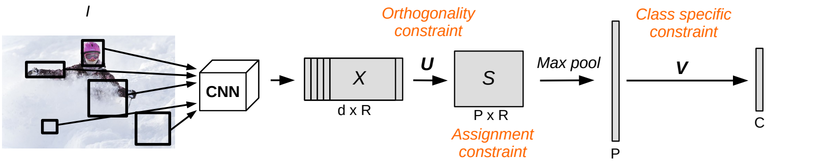


1) Parts should be complementary, *i.e.* parts should be different one from another.
2) Parts should cover as much as possible the diversity of regions extracted from images.
3) Parts should be discriminative with respect to classes.
4) Parts should be specific to categories.

$$C_{\perp}(U) = -\frac{1}{P^2}\sum_{i=1}^{P}\sum_{j=1, j\neq i}^{P}(u_i^T u_j)^2$$

u_p is assumed to be $l2$-normalized

Part constraints

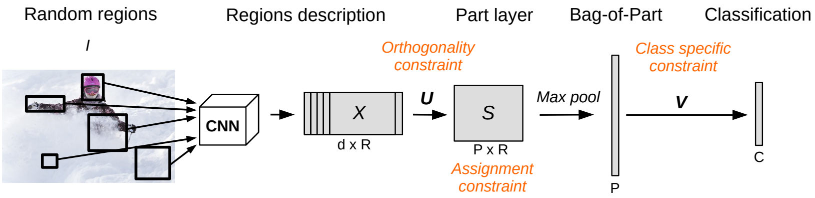


1) Parts should be complementary, *i.e.* parts should be different one from another.
2) Parts should cover as much as possible the diversity of regions extracted from images.
3) Parts should be discriminative with respect to classes.
4) Parts should be specific to categories.

$$C_{Assign}(U) = -\sum_{r=1}^{R}\sum_{p=1}^{P} s_{p,r} log(s_{p,r})$$

Softmax is first applied on the columns of the matrix S and u_p is assumed to be $l2$-normalized

Part constraints



1) Parts should be complementary, *i.e.* parts should be different one from another.
2) Parts should cover as much as possible the diversity of regions extracted from images.
3) Parts should be discriminative with respect to classes.
4) Parts should be specific to categories.

$$CS(V) = \frac{1}{P(C-1)} \sum_{i=1}^{C} \sum_{j=1, j \notin [q(i-1), qi]}^{P} V_{i,j}$$

Results

Table: DP-Net without constraints on parts and global representations

Dataset	MIT		Birds		ImageNet	
Network	VGG	RN50	VGG	RN50	VGG	RN50
Global	76.2	78.1	66.4	81.5	61.0	70.8
Parts	76.9	79.7	76.1	84.9	69.0	74.6

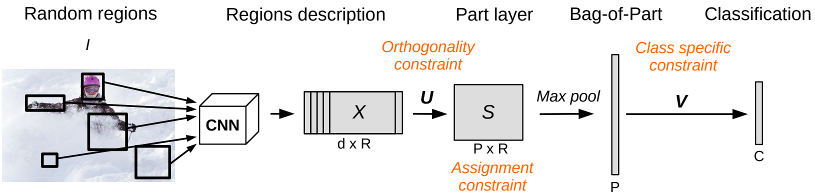
Table: Accuracy when using the constraints, with ResNet-50.

Dataset	Constraints			
	wo	$\perp$	Assign	CS
Birds	84.9	84.6	84.6	84.5
MIT	79.7	79.1	80.3	79.5
	$\perp$+Assign	CS+$\perp$	CS+Assign	CS+$\perp$+Assign
Birds	85.1	84.4	84.3	85.0
MIT	80.3	78.8	79.9	80.5

Interpretability

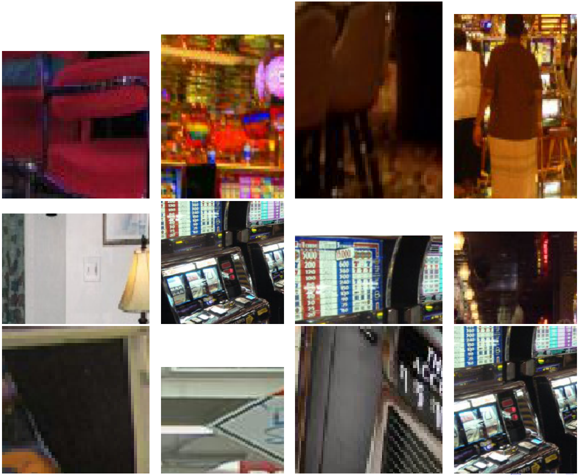
Class-level: what is the participation of each part.

Image-level: what is the participation of each part (as Class Activation Maps (CAM)).
A part can be linked to its most activating region in a given image.



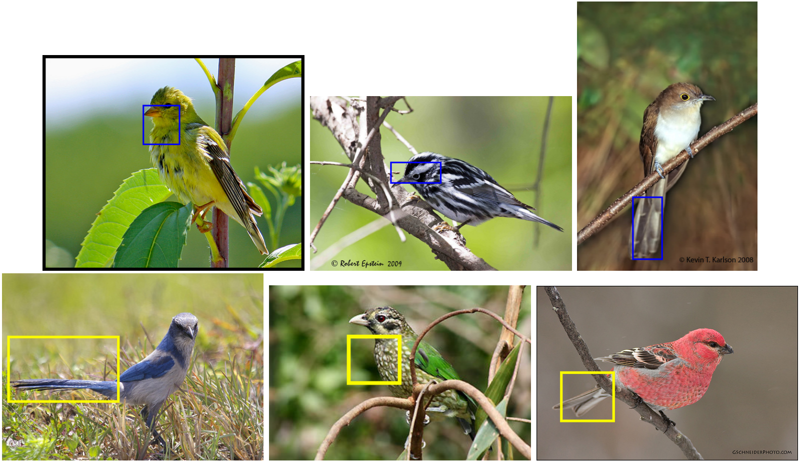
Interpretability - Casino parts

no constraints | orthogonal | sparse | class specific



Interpretability - heatmaps

Interpretability - best box



Part conclusions

Evaluation focused on accuracy and qualitative results.

Simpler explanations with specific constraints.

Ongoing works

Gradient denoising for better interpretability

Cross attention for CNNs

Improving insertion/deletion

Interpretability of models classifying gene data.

Thank you!