DP-Net: Learning Discriminative Parts for image recognition

Ronan Sicre, Hanwei Zhang, Julien Dejasmin, Chiheb Daaloul, Stéphane Ayache, Thierry Artières

LIS - Ecole Centrale Méditerranée
Marseille - France

2023

Pre-CNN image classification

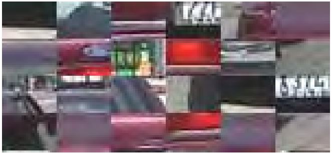
Pre-CNN classification pipelines

Feature extraction: SIFT, HOG

Feature encoding: Fisher vectors, VLAD

Pooling: Spatial pyramids

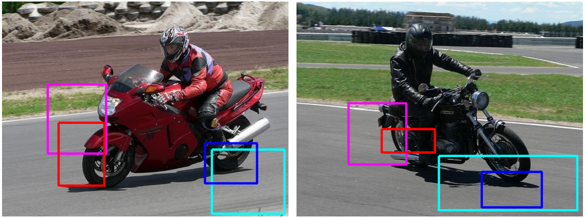
Learning and classification: SVMs



The standard pipeline can benefits from mid-level information.

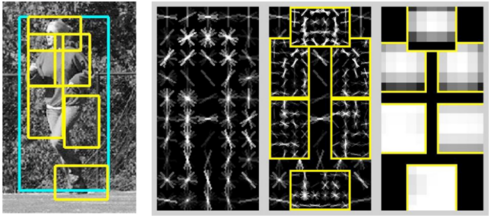
Part-based models: mid-level features

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



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

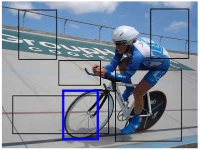
Two-stage optimization with specific definition of parts and constraints.

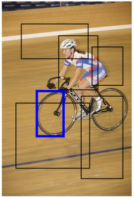
A bit of history

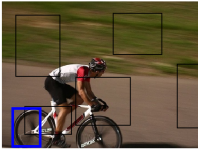
Prototype based architectures:

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

Discriminative part model for visual recognition







Stage 1: Learn parts of images that are relevant for a specific class: distinctive and generative.

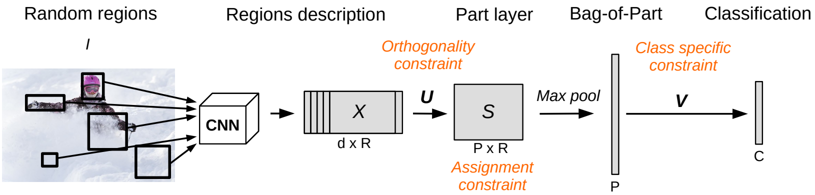
Generative: the part occurs often in the positive set
Distinctive: the part occurs rarely in the negative set

We learn parts iteratively inspired from the *soft-assign* algorithm:

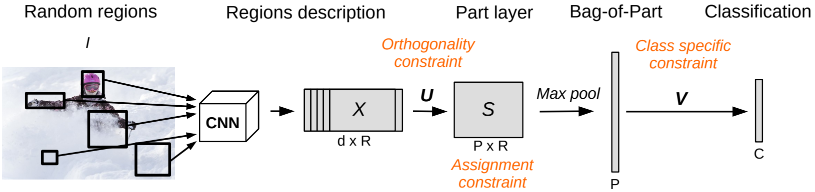
Stage 2: compute image descriptors based on part response.

Replace with a dedicated architecture that:
extract regions - compute parts activation - classify

DP-Net: Discriminative Part Network

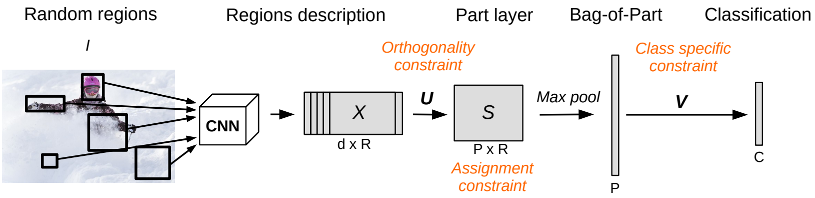


Part constraints



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.

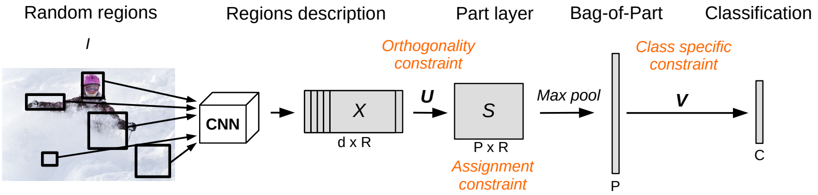
Part constraints



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.

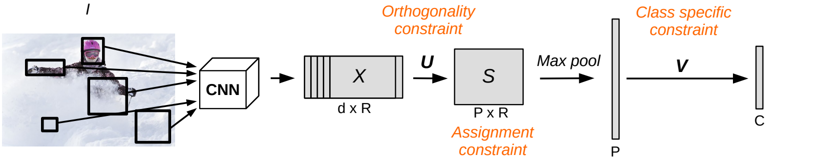
3) Parts should be discriminative with respect to classes.

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

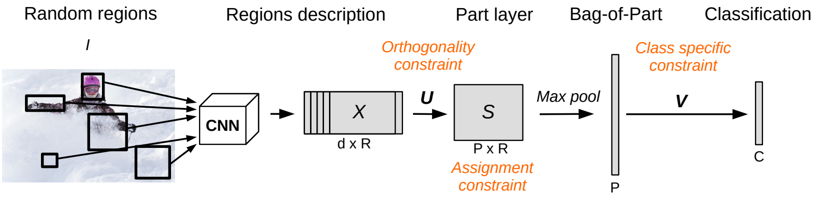


1) Parts should be complementary, *i.e.* parts should be different one from another.
2) Parts should cover as much as possible the diversity of regions extracted from images.
3) Parts should be discriminative with respect to classes.
4) Parts should be specific to categories.

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

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.

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

Interpretability

Table: Tables comparing our DP-Net without constraints on parts and global represen- tations

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

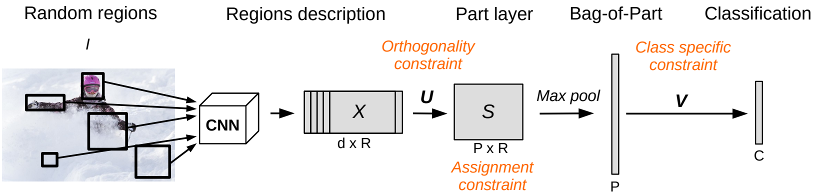
Table: Accuracy with ResNet 50 when using the constraints (wo = without constaint).

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

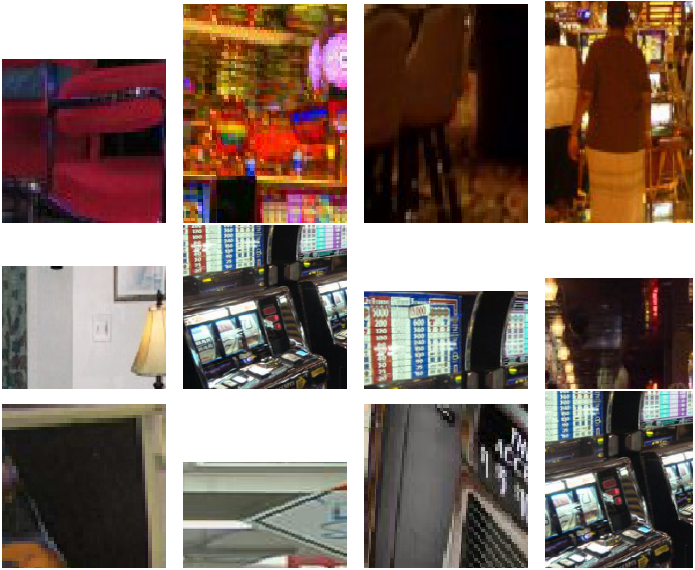
Class-level: what is the participation of each part.

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.



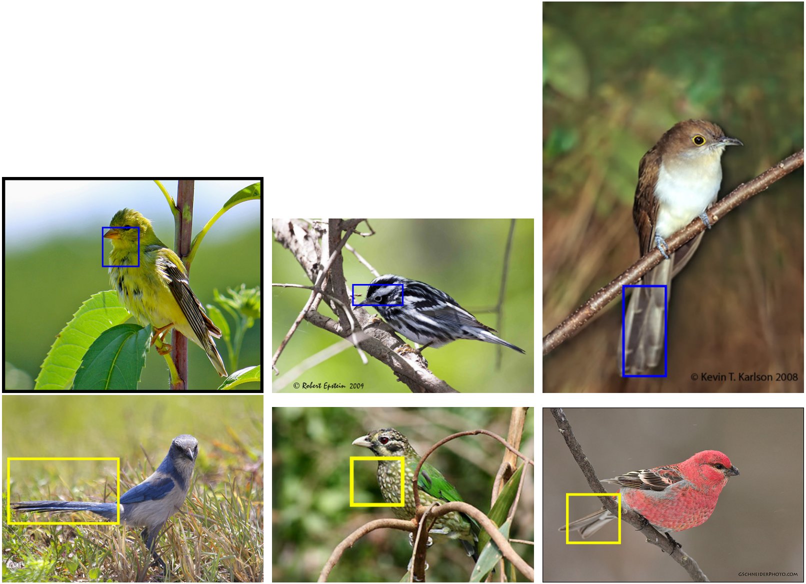
Interpretability - Casino parts

no constraints | orthogonal | sparse | class specific



Interpretability - heatmaps

Interpretability - best box



QUESTIONS