Praveen Kulkarni[1,2], Frédéric Jurie[2], Joaquin Zepeda[1], Patrick Pérez[1], Louis Chevalier[1]

[1]Technicolor, Rennes, France

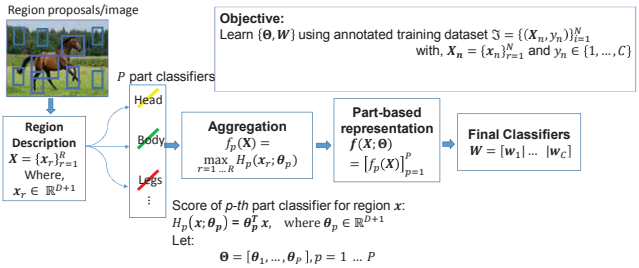
[2]University of CAEN CAEN, France

Email: {first name}.{last name} @technicolor.com, {first name}.{last name} @unicaen.fr

SPLeaP: Soft Pooling of Learned Parts for Image Classification

Introduction

Part-Based Models (PBMs) in the context of Image Classification.



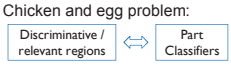
Motivation

Motivation for using PBMs:

1. By breaking object / image into parts at-least some of these are visible and recognizable.
2. Parts can be recombined in many ways to construct an object / scene.
3. Parts are distinctive of a particular class.
4. Part-based representation are compact.

Challenges:

1. How to train the part classifiers ?
2. Difficult problem – No annotations of relevant regions within an image.
3. Chicken and egg problem:



4. Initialization critical.

Approach

Novel joint learning framework for learning parameters in PBMs:

1. We describe each part classifier as a linear combination of weak non-linear classifiers.
2. We introduce a parameter, referred as the "per-part softness pooling coefficient" inside the optimization process.
3. Learning is done by a block-wise Stochastic Gradient Descent (SGD).

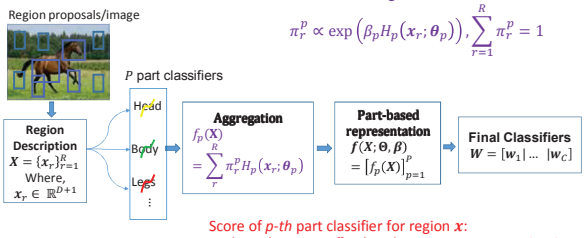
Contributions:

1. A boosted non-linear part classifiers.

2. Parametric soft-max Pooling.

Normalized weights:

$$\pi_r^p \propto \exp\left(\beta_p H_p(\boldsymbol{x}_r; \boldsymbol{\theta}_p)\right), \sum_{r=1}^{R} \pi_r^p = 1$$



Score of p-th part classifier for region $\boldsymbol{x}$:

$H_p(\boldsymbol{x}; \boldsymbol{\theta}_p) = \sum_{k=1}^{K} a_k^p \boldsymbol{\sigma}(\boldsymbol{u}_k^T \boldsymbol{x})$, where $\boldsymbol{\theta}_p \in \mathbb{R}^{K(D+2)}$

And let:

$$\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_P], p = 1 \dots P$$

Objective:

Learn $\{\boldsymbol{\Theta}, \boldsymbol{W}, \boldsymbol{\beta}\}$ using $\Im = \{(\boldsymbol{X}_n, y_n)\}_{i=1}^{N}$

with, $\boldsymbol{X}_n = \{\boldsymbol{x}_n\}_{r=1}^{N}, y_n \in \{1, \dots, C\}$ and pooling coefficent, $\boldsymbol{\beta} = [\beta_p]_{p=1}^{P}$

Multi-Class Classification Loss:

Using Logistic regression, class label for an input image $\boldsymbol{X}$ is predicted as:

$$Pr(y = c | \boldsymbol{X}; \boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{W}) = \frac{exp\left(\boldsymbol{w}_c^T \boldsymbol{f}(\boldsymbol{X}; \boldsymbol{\Theta}, \boldsymbol{\beta})\right)}{\sum_{d=1}^{C} exp\left(\boldsymbol{w}_d^T \boldsymbol{f}(\boldsymbol{X}; \boldsymbol{\Theta}, \boldsymbol{\beta})\right)}$$

$$\min_{\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{W}} -\sum_{n}\sum_{c} [y_n = c] \ln Pr(c | \boldsymbol{X}_n; \boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{W}) + \mu \|\boldsymbol{\Theta}\|^2 + \delta \|\boldsymbol{W}\|^2$$

Results

Experimental Settings:

Datasets:

1. Pascal-VOC-2007 (20 classes) - animals (e.g., dog, cat), vehicles (e.g., aeroplane, car) and other manufactured objects (e.g., tv monitor, chair).
2. MIT Indoor 67 (67 classes) - e.g., nursery, movie theater, casino or meeting room.
3. Willow dataset (7 classes) – e.g. play instrument, walk.

Performance measure:

1.mean Average Precision (*mAP*) for Pascal-VOC and Willow.
2.Accuracy (*acc*) for MIT dataset.

Region proposal scheme: Selective search.

Region feature extraction:

1. 128-D from 13-layer architecture (VGG-128)
2. 4096-D from 16-layer architecture (VD-16)
3. Krizhevsky's architecture pre-trained using ImageNet + Places database (HybridCNN).

Comparison with state-of-the-art methods:

Pascal-VOC-2007

Methods	*mAP*
VGG128-G	75.5
Oquab *et al.*	77.31
Li *et al.*	77.90
Cimpoi *et al.*	79.50
CNN-S fine tuned [5]	82.42
SPLeaP-VGG-128	**84.68**

13-layer architecture

Methods	*mAP*
VD-16-G	81.73
Cimpoi *et al.*	85.1
SPLeaP-VD-16	**88.01**

16-layer architecture

MIT indoor67 (left) and Willow (right)

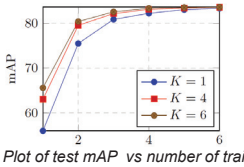
Methods	*acc*
Gong *et al.*	68.80
Li *et al.*	69.69
HybridCNN-G	72.54
Parizi *et al.*	73.30
SPLeaP-HybidCNN	**73.45**

Methods	*mAP*
Sicre *et al.*	81.90
VD-16-G	85.12
VD-16 (dense evaluation)	85.67
SPLeaP – VD-16	**88.47**

1. Parametric Soft-Max Pooling:

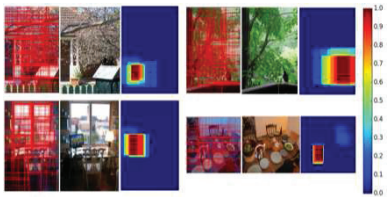
Average pooling	Max Pooling	Cross Valid. $\beta_p = \beta$	Learned β_p
80.77	83.23	84.31	**84.68**

2. A boosted non-linear part classifiers:



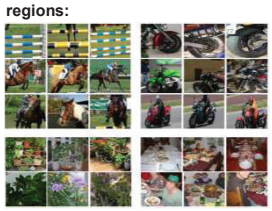
Plot of test mAP vs number of training epochs

Selectivity of Part Classifiers using heatmaps



Heatmaps for images Pascal-VOC-2007 of classes (clockwise from top-left) "potted plant", "bird", "bottle" and "TV monitor".

Specialization of Part Classifiers to discriminative regions:



Discriminative parts for the four classes (clockwise from top-left) "horse", "motorbike", "dining table", and "potted plant".

Conclusion

1. We introduce SPLeaP, a novel part-based model for image classification.
2. Based on non-linear part classifiers combined with part-dependent soft pooling - both being trained jointly with the image classifiers surpasses standard pooling approaches and other PBMs on several challenging classification tasks.
3. Our method does not need any particular initialization of the parts.

References

1. Parizi *et al.*: Automatic Discovery and Optimization of Parts for Image Classification. In: ICLR. (2015)
2. Li *et al.*: Mid-level deep pattern mining. In: CVPR (2015).
3. Gong *et al.*: Multi-scale orderless pooling of deep convolutional activation features ECCV (2014).
4. Cimpoi *et al.*: Deep filter banks for texture recognition and segmentation. In: CVPR. (2015)
5. Chatfield *et al.*: Return of the Devil in the Details: Delving Deep into Convolutional Nets. In BMVC 2014.
6. Oquab *et al.* : Learning and transferring mid-level image representations using convolutional neural networks.
7. Sicre *et al.*: Discovering and aligning discriminative mid-level features for image classification