Object-centric spatial pooling for image classification

Olga Russakovsky[1] Yuanqing Lin[2] Kai Yu[3] Li Fei-Fei[1]

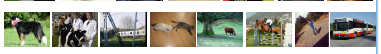
[1] Stanford University [2] NEC Laboratories America [3] Baidu, Inc.

NEC Laboratories America
Relentless passion for innovation

Image classification

Training: cars

not cars



Testing: Does this image contain a car?

Pipeline:

Standard: SPM pooling

The Spatial Pyramid Matching (SPM) approach forms the image representation by pooling visual features over pre-defined coarse spatial bins.

SPM-based pooling results in inconsistent image representations when the object of interest appears in different locations within the image.

Object-centric spatial pooling

In contrast, our object-centric spatial pooling (OCP) approach forms the image representation by

(1) localizing the object of interest, and then
(2) pooling foreground visual features separately from the background features.

Motivation

If localization is perfect, classifiers trained using the OCP representation outperform those trained using the SPM representation by up to 15% in mean average precision on the PASCAL07 classification task.

Challenges

- No localization information is available during training
- Inferring inaccurate localization will hurt classification accuracy

Training

Given N images with labels $y_1 .. y_N \in \{-1,+1\}$ and no object location information

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 + C\sum_i \text{slack}_i$$

$$\text{s.t. } y_i \max_{\substack{\text{regions}\\ \text{of Image}_i}} [\mathbf{w}^T F_{\text{region}} + b] \geq 1 - \text{slack}_i \ \ \forall i$$

Credit: Nguyen '09

Iteration 0

Step 1: Predict object location is the full image
Step 2: Learn appearance model

Iteration 1

Step 1: Carefully update location prediction
Step 2: Learn new appearance model

...

Upon convergence

Have an appearance model which can accurately predict object location

Representation

The foreground-background representation of a region:

low-level image descriptors pooled over the foreground
concatenated with
low-level image descriptors pooled over the background

This representation can be seen as both a bounding box representation (for detection) and an image-level representation (for classification)

Optimization scale

PASCAL VOC 2007, 20 object classes
Training examples:
~200 positive examples +
~5000 negative images x ~1500 regions per image (van de Sande '11) =
> 7M examples
Feature representation:
Codebook of 8192 x (10 foreground + 1 background bin) =
90K features
Data size: 700Gb
Training time: inner loop in 7-8 hours on single machine, total 3 days

Results

Dataset

PASCAL VOC 2007, 20 object classes
DHOG features with LLC coding (codebook size 8192, k=5) and max pooling
1x1,3x3 SPM pooling on foreground + 1 background bin

Joint classification and localization

Classification and localization are mutually beneficial

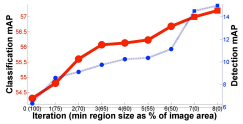


Image classification

Baseline SPM on full image: 54.3% classification mAP
Object-centric pooling (OCP): **57.2%** classification mAP

Method	aero	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dining	dog	horse	mot	person	plant	sheep	sofa	train	tv
SPM	72.5	56.3	**49.5**	63.5	22.4	60.1	76.4	57.5	**51.9**	42.2	48.9	38.1	75.1	62.8	82.9	20.5	38.1	46.0	**71.7**	50.5
OCP	**74.2**	**63.1**	45.1	**65.9**	**29.5**	**64.7**	**79.2**	**61.4**	51.0	**45.0**	**54.8**	**45.4**	**76.3**	**67.1**	**84.4**	**21.8**	**44.3**	**48.8**	70.7	**51.7**

Baseline with 4-level SPM: 54.8% classification mAP
OCP foreground-only: 55.7% classification mAP
OCP with state-of-the-art detector: 56.9% classification mAP

Weakly supervised object localization

Detection AP on PASCAL07 test set (no bounding boxes during training)

Method	aeroplane		bicycle		boat		bus		horse		motorbike		average
	left	right	left	right	left	right	left	right	left	right	left	right	
Deselaers '10	9.1	23.6	33.4	49.4	0.0	0.0	0.0	16.4	9.6	9.1	20.9	16.1	16.0
Pandey '11	7.5	21.1	38.5	44.8	0.3	0.5	0	0.3	45.9	17.3	43.8	27.2	20.8
OCP	**30.8**		25.0		**3.6**		**26.0**		21.3		29.9		**22.8**

In contrast to prior work, we accurately localize objects on PASCAL07 without splitting them up by viewpoint.

Bibliography

1. Nguyen, M.H., Torresani, L., de la Torre, F., Rother, C.: Weakly supervised discriminative localization and classification: a joint learning process. In ICCV 209.
2. van de Sande, K.E.A., Uijlings, J.R.R., Gevers, T., Smeulders, A.W.M.: Segmentation as selective search for object recognition. In ICCV 2011.
3. Deselaers, T., Alexe, B., Ferrari, V.: Localizing objects while learning their appearance. In ECCV 2010.
4. Pandey, M., Lazebnik, S.: Scene recognition and weakly supervised object localization with deformable part-based models. In ICCV 2011.