Object-centric spatial pooling for image classification

Olga Russakovsky1  Yuanqing Lin2  Kai Yu3  Li Fei-Fei1
1 Stanford University  2 NEC Laboratories America  3 Baidu, Inc.

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 \( y_i \) and no object location information, the optimization goal is

\[
\min \frac{1}{N} \sum \left( \|y_i - \hat{y}_i\|_1 + C \sum \text{slack}_i \right) + \text{slack}_i \geq 1 - \text{slack}_i,
\]

where \( y_i \) are the labels, \( \hat{y}_i \) are the predicted labels, and \( \text{slack}_i \) are the slack variables.

Iteration 0

Step 1: Predict object location
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

Optimization scale

PASCAL VOC 2007, 20 object classes
Training examples:
- ~200 positive examples
- ~5000 negative examples x ~1500 regions per image (van de Sande 13)
- ~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): 57.3% classification mAP

Weakly supervised object localization
Detection AP on PASCAL07 test set (no bounding boxes during training)

Bibliography