WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks

Thibaut DURAND, Nicolas THOME, Matthieu CORD
Sorbonne Universités, UPMC Univ Paris 06, LIP6, Paris, France

CVPR 2016

Context

Goal: image classification or ranking

▷ Learning of deep CNN on small datasets
▷ How to transfer on datasets with complex scenes?
▷ Problem: invariance (scale, translation)

OK for centered object

KO for "natural" image

▷ Efficient transfer: needs bounding boxes
▷ Full annotations expensive ⇒ weak supervision
  ▷ Select relevant regions → better prediction
  ▷ Baseline model: Latent SVM (LSVM)

Contributions

▷ New region aggregation strategy
▷ Structured ranking AP loss for WSL
▷ Fully convolutional architecture
▷ Experimental validation on 6 datasets

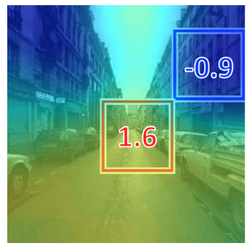
Region Aggregation

Baseline: max aggregation [1] (MIL, LSVM)

Our MANTRA aggregation [2]

▷ `max` + `min` pooling (negative evidence)
  ▷ `max`: indicator of the **presence** of the class
  ▷ `min`: indicator of the **absence** of the class

street model



highway model

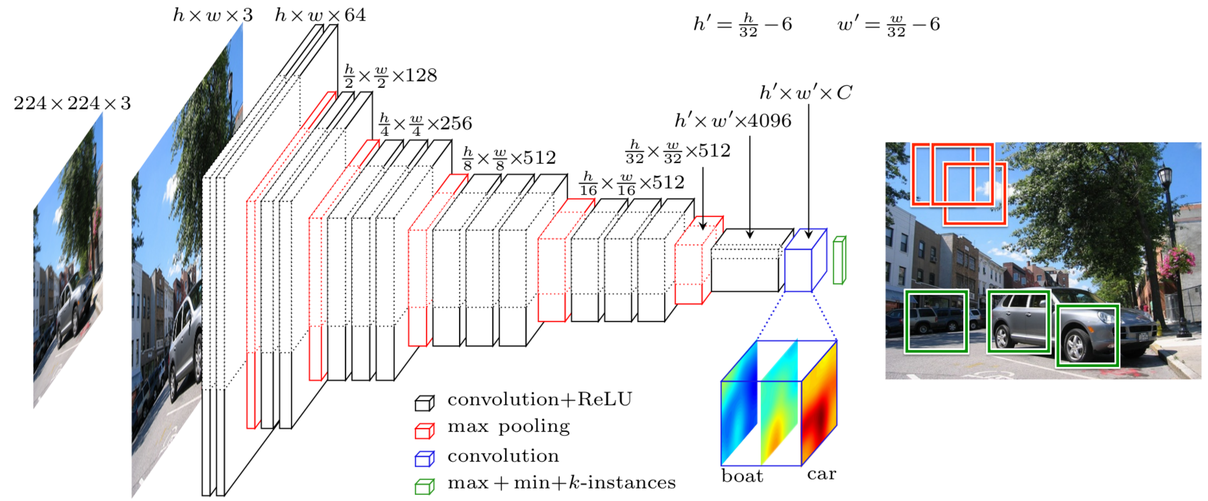
Our WELDON aggregation

▷ MANTRA extension to multiple regions

$$\max \rightarrow \frac{1}{k}\sum_{i=1}^{k} i\text{-th max} \qquad \min \rightarrow \frac{1}{k}\sum_{i=1}^{k} i\text{-th min}$$

▷ More robust to outliers

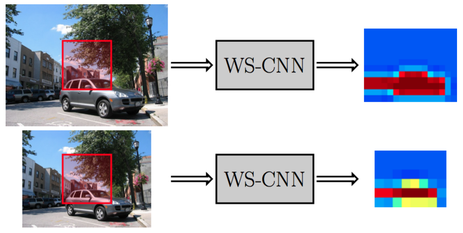
WELDON Architecture for Classification



▷ Fully connected layer → convolution layer
  ▷ Sliding window approach / shared features
▷ Spatial aggregation
  ▷ Object localization prediction

Experiments

▷ VGG16 pre-trained on ImageNet
▷ Multi-scale: 8 scales (Object Bank fusion)



Multi-label (mAP)	VOC 2007	VOC 2012
VGG16	84.5	82.8
Deep WSL MIL [1]		81.8
WELDON	**90.2**	**88.5**

Multi-label (mAP)	VOC12 Action	COCO
VGG16	67.1	59.7
Deep WSL MIL [1]		62.8
WELDON	**75.0**	**68.8**

Multi-class (acc)	15 Scene	MIT67
VGG16	91.2	69.9
MOP CNN [4]		68.9
Negative parts [3]		77.1
WELDON	**94.3**	**78.0**

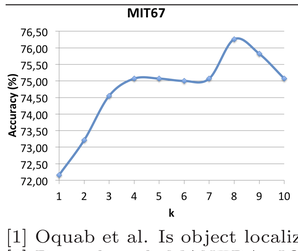
VOC 07/12

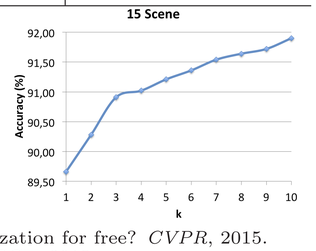
COCO

MIT67

▷ Analysis of improvements (mono-scale)

`max`	+k=3	+`min`	+AP	VOC07	VOCAct
✓				83.6	53.5
✓	✓			86.3	62.6
✓		✓		87.5	68.4
✓		✓	✓	88.4	71.7
✓	✓	✓		87.8	69.8
✓	✓	✓	✓	**88.9**	**72.6**





[1] Oquab et al. Is object localization for free? *CVPR*, 2015.
[2] Durand et al. MANTRA. *ICCV*, 2015.
[3] Parizi et al. Automatic discovery of parts. *ICLR*, 2015.
[4] Gong et al. Multi-scale orderless pooling. *ECCV*, 2014.

Learning

▷ Stochastic gradient descent training
▷ Back-propagation of the **selected windows**

Class *car* is present

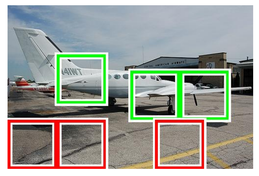
Class *boat* is absent

▷ Optimized ranking metrics (Average Precision)
  ▷ Surrogate upper-bound loss definition
  ▷ Generalized MANTRA ranking instantiation

Visual results

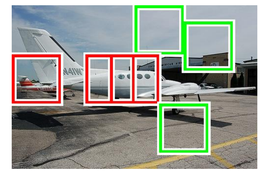
Correct predictions

Correct model

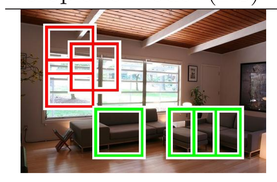


Aeroplane model (1.8)

Incorrect model



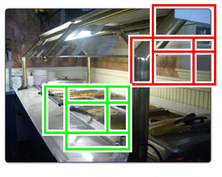
Bus model (-0.4)



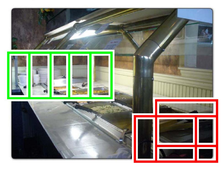
Sofa model (1.2)

Horse model (-0.6)

Failing example



Buffet model (1.5)



Restaurant kitchen (1.6)

Latent Variables Models

LSSVM: $\max_{h\in\mathcal{H}} f(x,y,h)$ (maximization)
HCRF: $\sum_{h\in\mathcal{H}} \exp(f(x,y,h))$ (marginalization)
WELDON: $\sum_{h\in\Omega\subseteq\mathcal{H}} f(x,y,h)$