Large Scale Semi-supervised Object Detection using Visual and Semantic Knowledge Transfer

Yuxing Tang[1], Josiah Wang[2], Boyang Gao[1,3], Emmanuel Dellandrea[1], Robert Gaizauskas[2], Liming Chen[1]

1. Ecole Centrale de Lyon, France 2. The University of Sheffield, UK 3. Istituto Italiano di Tecnologia, Italy

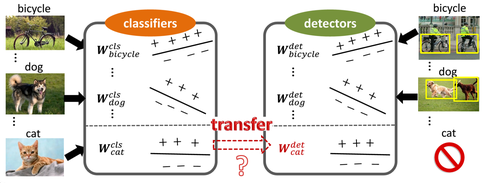
IEEE 2016 Conference on Computer Vision and Pattern Recognition

CVPR 2016

Overview

Goal:

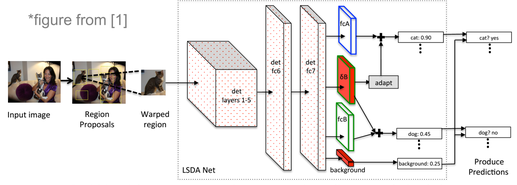
To convert the image classifiers into object detectors on weakly annotated categories (without bounding box annotations), by transferring knowledge from visually or/and semantically similar categories.



LSDA Background

LSDA: Large Scale Detection through Adaptation [1]

*figure from [1]



A: Weakly annotated categories (image labels)
B: Fully annotated categories (bounding boxes)

1. Pre-train an 8-layer CNN on ImageNet;
2. Fine-tune for classification on A+B;
3. Fine-tune for detection on B in R-CNN [2] manner (category-***invariant*** adaptation, layer 1-7);
4. Category-specific adaptation on A (layer 8).

Assumption: CLS and DET difference on a target category has a positive correlation with those of similar categories.

$$\forall j \in \mathcal{A} : \overrightarrow{w_j^d} = \overrightarrow{w_j^c} + \frac{1}{k}\sum_{i=1}^{k} \Delta_{\mathcal{B}_i^j}$$

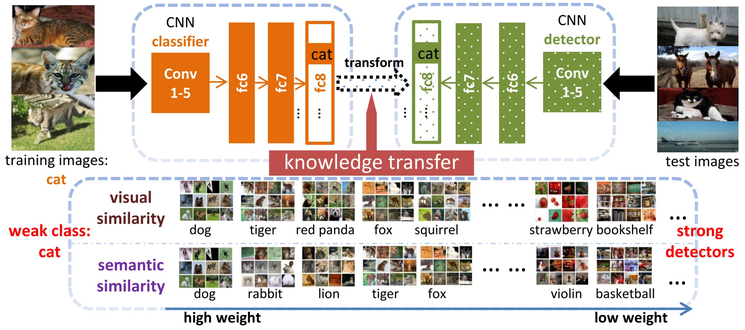
w: *fc8* weight, Δ: *fc8* weight change from CLS to DET of the neighbor category.

Neighbor definition: L2 distance between w_j^c and w_i^c.

Similarity-based Knowledge Transfer Model

Motivation:

1. Category-specific difference exists between classifier and detector;
2. Visually and semantically similar categories may exhibit more common transferable properties than dissimilar categories;
3. Visual similarity and semantic relatedness are shown to be correlated, especially when measured against objects cropped out from images (thus discarding background clutter).



Visual similarity measure s_v

Visual similarity between two categories:

$s_v(j,i) \propto \frac{1}{N}\sum_{n=1}^{N} CNN(I_n)_i$ on a balanced CLS validation set.

Semantic similarity measure s_s

Each category is a WordNet synset. 300-dimension vector using word2vec embedding + synset embedding.

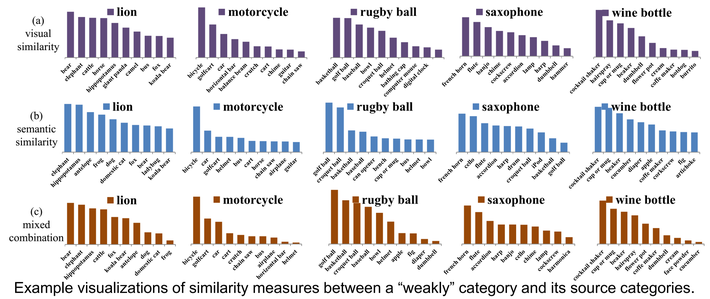
Semantic similarity (2 measures):

1. Inversely proportional to L2 distance of two feature vectors;
2. Coefficient of linear combination of vectors (sparse representation ≤ 20).

Mixture transfer model

$$s = intersect[\alpha s_v + (1-\alpha) s_{s:sparse}]$$

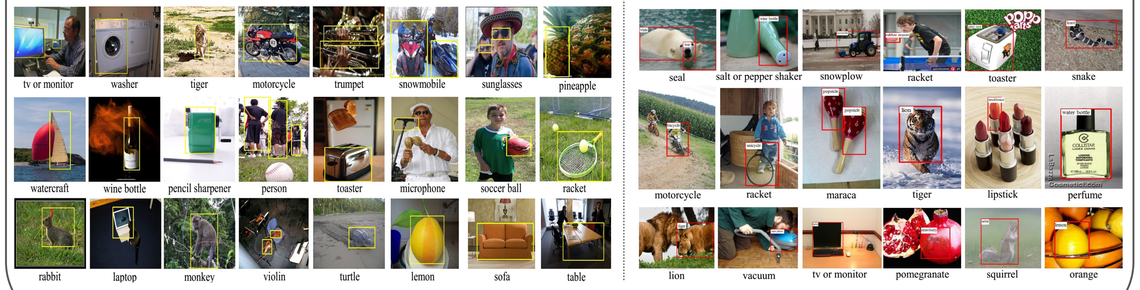
Experimental results



Example visualizations of similarity measures between a "weakly" category and its source categories.

Method	mAP
CLS network	10.31
LSDA (adapt 1-7)	15.85
LSDA (adapt 1-8)	16.33
Our visual KT	19.02
Our semantic KT1	18.32
Our semantic KT2	19.04
Our mixture KT	20.03
Full supervised	26.25

Detection mAP on 100 weakly labeled categories on ILSVRC2013 val2 subset.



Examples of correct detections (true positives) and incorrect detections of our mixture knowledge transfer model on ILSVRC2013 images.

Conclusion

- We investigated how knowledge about object similarities from both visual and semantic domains can be transferred to adapt an image classifier to an object detector.
- Both visual and semantic similarities play an essential role in improving the adaptation process, and the combination of the two modalities yielded better performance.

References

[1] J. Hoffman et al. LSDA: Large scale detection through adaptation. *NIPS* 2014

[2] R. Girshick et al. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR* 2014

Know more at:

https://liris.cnrs.fr/~ytang/