



This is a repository copy of *Visual and semantic knowledge transfer for large scale semi-supervised object detection*.

White Rose Research Online URL for this paper:
<http://eprints.whiterose.ac.uk/123674/>

Version: Accepted Version

Article:

Tang, Y., Wang, J.K. orcid.org/0000-0003-0048-3893, Wang, X. et al. (4 more authors) (2017) Visual and semantic knowledge transfer for large scale semi-supervised object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. ISSN 0162-8828

<https://doi.org/10.1109/TPAMI.2017.2771779>

© 2017 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/ republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works. Reproduced in accordance with the publisher's self-archiving policy.

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

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

Yuxing Tang, Josiah Wang, Xiaofang Wang, Boyang Gao, Emmanuel Dellandréa, Robert Gaizauskas and Liming Chen *Senior Member, IEEE*

Abstract—Deep CNN-based object detection systems have achieved remarkable success on several large-scale object detection benchmarks. However, training such detectors requires a large number of labeled bounding boxes, which are more difficult to obtain than image-level annotations. Previous work addresses this issue by transforming image-level classifiers into object detectors. This is done by modeling the differences between the two on categories with both image-level and bounding box annotations, and transferring this information to convert classifiers to detectors for categories without bounding box annotations. We improve this previous work by incorporating knowledge about object similarities from visual and semantic domains during the transfer process. The intuition behind our proposed method is that visually and semantically similar categories should exhibit more common transferable properties than dissimilar categories, e.g. a better detector would result by transforming the differences between a dog classifier and a dog detector onto the cat class, than would by transforming from the violin class. Experimental results on the challenging ILSVRC2013 detection dataset demonstrate that each of our proposed object similarity based knowledge transfer methods outperforms the baseline methods. We found strong evidence that visual similarity and semantic relatedness are complementary for the task, and when combined notably improve detection, achieving state-of-the-art detection performance in a semi-supervised setting.

Index Terms—Object detection, convolutional neural networks, semi-supervised learning, transfer learning, visual similarity, semantic similarity, weakly supervised object detection.



1 INTRODUCTION

GIVEN an image, an object detection/localization method aims to recognize and locate objects of interest within it. It is one of the most widely studied problems in computer vision with a variety of applications. Most object detectors adopt strong supervision in learning appearance models of object categories, that is by using training images annotated with bounding boxes encompassing the objects of interest, along with their category labels. The recent success of deep convolutional neural networks (CNN) [1] for object detection, such as DetectorNet [2], OverFeat [3], R-CNN [4], SPP-net [5], Fast R-CNN [6], Faster R-CNN [7], YOLO [8] and SSD [9], is heavily dependent on a large amount of training data manually labeled with object localizations (e.g., PASCAL VOC [10], ILSVRC (subset of ImageNet) [11], and Microsoft COCO [12] datasets).

Although localized object annotations are extremely valuable, the process of manually annotating object bounding boxes is extremely laborious and unreliable, especially

for large-scale databases. On the other hand, it is usually much easier to obtain annotations at *image* level (e.g., from user-generated tags on Flickr or Web queries). For example, ILSVRC contains image-level annotations for 1,000 categories, while object-level annotations are currently restricted to only 200 categories. One could apply image-level classifiers directly to detect object categories, but this will result in a poor performance as there are differences in the statistical distribution between the training data (whole images) and the test data (localized object instances). Previous work by Hoffman et al. [13] addresses this issue, by learning a transformation between CNN classifiers and detectors of object categories with *both* image-level and object-level annotations (“strong” categories), and applying the transformation to adapt image-level classifiers to object detectors for categories with *only* image-level labels (“weak” categories). Part of this work involves transferring *category-specific* classifier and detector differences of visually similar “strong” categories equally to a classifier of a “weak” category to form a detector for that category (Fig. 1). We argue that more can potentially be exploited from such similarities in an informed manner to improve detection beyond using the measures solely for nearest neighbor selection (see Section 4.1). Moreover, since there is evidence that deep CNNs trained for image classification also learn proxies to objects and object parts [14], the transformation from CNN classifiers to detectors is reasonable and practicable.

Our main contribution in this paper is therefore to incorporate external knowledge about object similarities from visual and semantic domains in modeling the aforementioned category-specific differences, and subsequently transferring this knowledge for adapting an image classifier to an object

- Y. Tang is with Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, National Institutes of Health (NIH) Clinical Center, 10 Center Dr., Bethesda, 20814, USA. This work was done when he was a Ph.D. student in LIRIS, École Centrale de Lyon, France.
E-mail: yuxing.tang@ec-lyon.fr
- X. Wang, E. Dellandréa and L. Chen are with LIRIS, CNRS UMR 5205, École Centrale de Lyon, 36 avenue Guy de Collongue, Écully, F-69134, France.
E-mail: {xiaofang.wang, emmanuel.dellandrea, liming.chen}@ec-lyon.fr
- J. Wang and R. Gaizauskas are with Department of Computer Science, The University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, United Kingdom.
E-mail: {j.k.wang, r.gaizauskas}@sheffield.ac.uk
- B. Gao is with Department of Advanced Robotics (ADVR), Istituto Italiano di Tecnologia (IIT), Via Morego, Genova, 16163, Italy.
E-mail: boyang.gao@iit.it

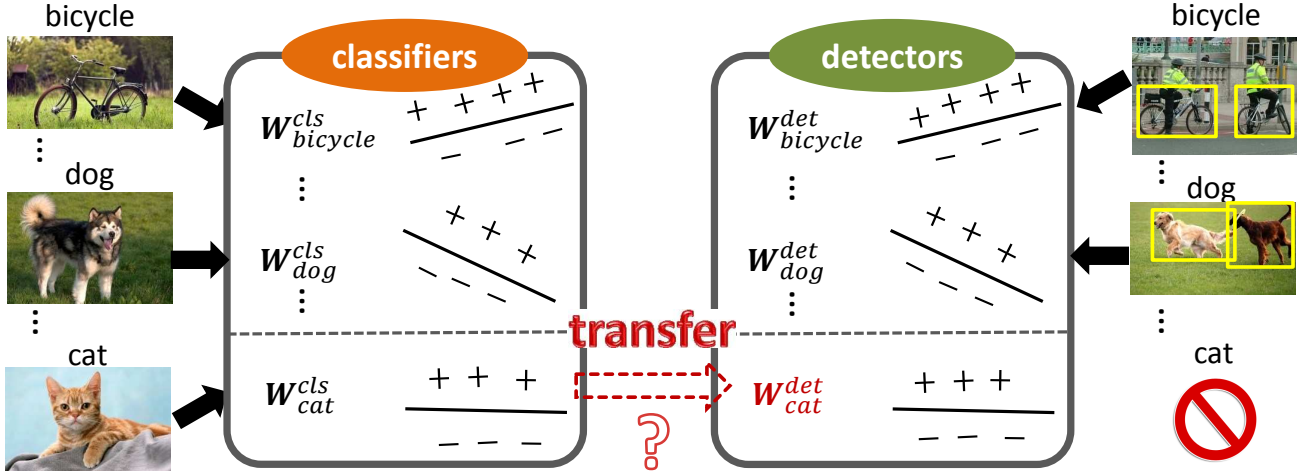


Fig. 1. In this work, we consider a dataset containing image-level labels for all the categories, while object-level bounding box annotations are only available for some of the categories (i.e., weakly labeled categories). How can we transform a CNN classification network into a detection network to detect the weakly labeled categories (e.g., the *cat* class)?

detector for a “weak” category. Our proposed method is motivated by the following observations: (i) category specific difference exists between a classifier and a detector [4], [13]; (ii) visually and semantically similar categories may exhibit more common transferable properties than visually or semantically dissimilar categories; (iii) visual similarity and semantic relatedness are shown to be correlated, especially when measured against object instances cropped out from images (thus discarding background clutter) [15]. Intuitively, we would prefer to adapt a *cat* classifier to a *cat* detector by using the category-specific differences between the classifier and the detector of a *dog* rather than of a *violin* or a *strawberry* (Fig. 2). The main advantage of our proposed method is that knowledge about object similarities can be obtained without requiring further object-level annotations, for example from existing image databases, text corpora and external knowledge bases.

Our work aims to answer the question: can knowledge about visual and semantic similarities of object categories (and the combination of both) help improve the performance of detectors trained in a weakly supervised setting (i.e., by converting an image classifier into an object detector for categories with only image-level annotations)? Our claim is that by exploiting knowledge about objects that are visually and semantically similar, we can better model the category-specific differences between an image classifier and an object detector and hence improve detection performance, without requiring bounding box annotations. We also hypothesize that the combination of both visual and semantic similarities can help further improve the detector performance. Experimental results on the challenging ILSVRC2013 dataset [11] validate these claims, showing the effectiveness of our approach of transferring knowledge about object similarities from both visual and semantic domains to adapt image classifiers into object detectors in a semi-supervised manner.

A preliminary version of this work appeared in [16]. In this paper, we provide more technical details of our models, introduce a bounding-box post-processing method based

on the transferability of regression models, and present extended results with more comparisons. The rest of the paper is organized as follows. We review related work in Section 2 and define the semi-supervised object detection problem in Section 3. In Section 4, we first review the LSDA framework, and then introduce our two knowledge transferring methods (i.e. visual similarity based method and semantic similarity based method). We present our experimental results and comparisons in Section 5. In Section 6, we conclude and describe future direction.

2 RELATED WORK

With the remarkable success of deep CNN on large-scale object recognition [1] in recent years, a substantial number of CNN-based object detection frameworks [2], [3], [4], [5], [6], [7], [8], [9] have emerged. However, these object detectors are trained in a fully supervised manner, where bounding box annotations are necessary during training. The requirement of bounding box annotations hinders the application of these methods in large-scale datasets where training images are weakly annotated.

2.1 Weakly Supervised Learning

Weakly supervised learning methods for object detection attempt to learn localization cues from image-wide labels indicating the presence or absence of object instances of a category, thus reducing or removing the requirement of bounding box annotations [17], [18], [19], [20], [21], [22], [23], [24], [25], [26]. Recently, there have been several studies in CNN-based object detection in a *weakly*-supervised setting [27], [28], [29], [30], [31], i.e. using training images with only image-level labels and no bounding boxes. The common practice is to jointly learn an appearance model together with the latent object location from such weak annotations. Such approaches only adopt CNN as a feature extractor, and exhaustively mine image regions extracted by region proposal approaches, such as Selective Search [32], BING [33], and EdgeBoxes [34]. Oquab et al. [35] develop a

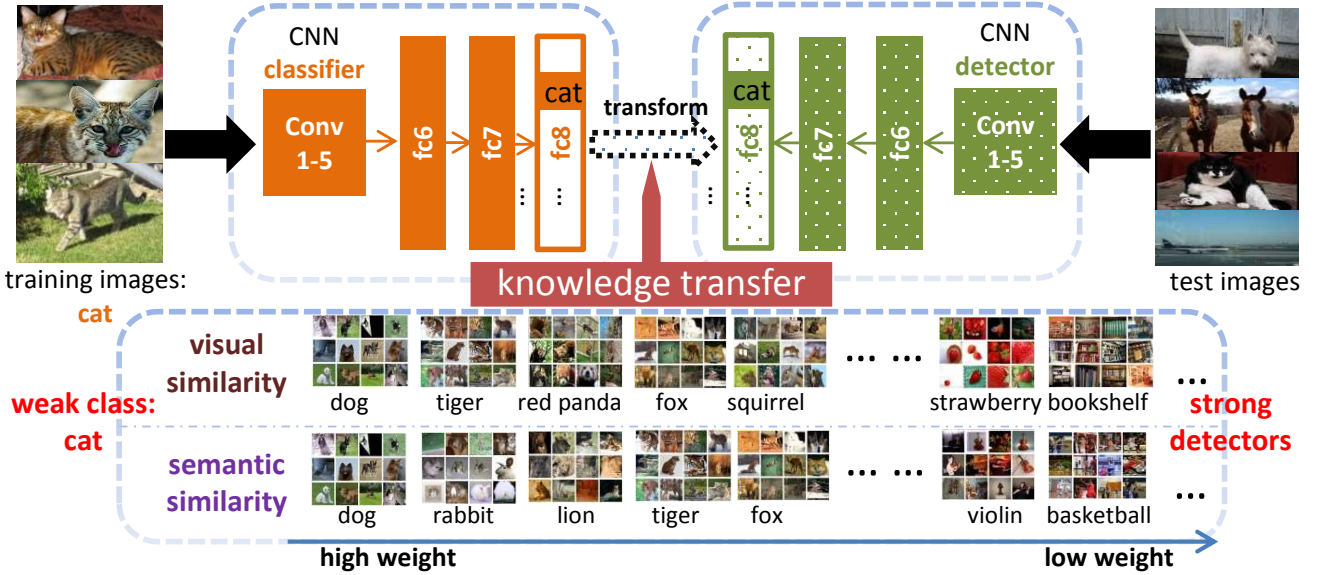


Fig. 2. An illustration of our similarity-based knowledge transfer model. The question we investigate is whether knowledge about object similarities – visual and semantic – can be exploited to improve detectors trained in a semi-supervised manner. More specifically, to adapt the image-level classifier (up-left) of a “weakly labeled” category (no bounding boxes) into a detector (up-right), we transfer information about the classifier and detector differences of “strong” categories (with image-level and bounding box annotations, bottom of the figure) by favoring categories that are more similar to the target category (e.g., transfer information from *dog* and *tiger* rather than *basketball* or *bookshelf* to produce a *cat* detector).

weakly supervised CNN end-to-end learning pipeline that learns from complex cluttered scenes containing multiple objects by explicitly searching over possible object locations and scales in the image, which can predict image labels and coarse locations (but not exact bounding boxes) of objects. Bilen and Vedaldi [36] propose a Weakly Supervised Deep Detection Network (WSDNN) method that extends a pre-trained network to a two-stream CNN: recognition and detection. The recognition and detection scores for region proposals are aggregated to predict the object category. Zhou et al. [37] adopt a classification-trained CNN to learn to localize object by generating Class Activation Maps (CAM) using global average pooling (GAP). Hoffman et al. [13] propose a Large Scale Detection through Adaptation (LSDA) algorithm that learns the difference between the CNN parameters of the image classifier and object detector of a “fully labeled” category, and transfers this knowledge to CNN classifiers for categories without bounding box annotated data, turning them into detectors. For LSDA, auxiliary object-level annotations for a subset of the categories are required for training “strong” detectors. This can be considered a semi-supervised learning problem (see Section 3). We improve upon LSDA, by incorporating knowledge about visual and semantic similarities of object categories during the transfer from a classifier to a detector.

2.2 Transfer Learning

Another line of related work is to exploit knowledge transfer from various domains. Transfer learning (TL) [38] aims to transfer knowledge across different domains or tasks. Two general categories of TL have been proposed in previous work: *homogeneous* TL [13], [39], [40] in a single domain but with different data distributions in training and testing sets, and *heterogeneous* TL [41], [42], [43] across different domains or modalities. LSDA treats the transfer from classifiers to

detectors as a homogeneous TL problem as the data distributions for image classification (whole image features) and object detection (image region features) are different. The adaptation from a classifier to a detector is, however, restricted to the visual domain. Lu et al. [44] propose a sparse representation based discriminative knowledge transfer method that leverages relatedness of various source categories with the target category, where only few training samples existed, to enhance learning of the target classifier. Rochan and Wang [41] propose an appearance transfer method by transferring semantic knowledge (heterogeneous TL) from familiar objects to help localize novel objects in images and videos. Singh et al. [45] transfer tracked object boxes from weakly labeled videos to weakly labeled images to automatically generate pseudo ground-truth bounding boxes. Our work integrates knowledge transfer via both visual similarity (homogeneous TL) and semantic relatedness (heterogeneous TL) to help convert classifiers into detectors. Frome et al. [46] present a deep visual-semantic embedding model learned to recognize visual objects using both labeled image data and semantic information collected from unannotated text. Shu et al. [42] propose a weakly-shared Deep Transfer Network (DTN) that hierarchically learns to transfer semantic knowledge from web texts to images for image classification, building upon Stacked Auto-Encoders [47]. DTN takes auxiliary text annotations (user tags and comments) and image pairs as input, while our semantic transfer method only requires image-level labels.

2.3 Semantic Similarity of Text

Semantic similarity is a well-explored area within the Natural Language Processing community. The main objective is to measure the distance between the *semantic* meanings of a pair of words, phrases, sentences, or documents. For example, the word “*car*” is more similar to “*bus*” than it is

to “cat”. The two main approaches to measuring semantic similarity are knowledge-based approaches and corpus-based, distributional methods. In the case of knowledge-based approaches, external resources such as thesauri (primarily WordNet [48]) or online knowledge bases are used to compute the similarity between the semantic meaning of two terms, for example using path-based similarity [49] or information-content similarity measures [50], [51]. The heavy reliance on knowledge bases, which tend to suffer from issues such as missing words, resulted in the development of *distributional* methods that rely instead on text corpora. In such methods, each term is represented as a *context* vector, and two terms are assumed to have similar vectors if they occur frequently in the same context (e.g. “car” and “truck” have similar vectors because they often co-occur with “drive” and “road”). Such context vectors are more often referred to in recent years as *word embeddings*. Recent advances in word embeddings trained on large-scale text corpora [52], [53] have helped progress research in distributional methods to semantic similarity, as it has been observed that semantically related word vectors tend to be close in the embedding space, and that the embeddings capture various linguistic regularities [54] (*King - Man + Woman ≈ Queen*). As such, we will concentrate on such state-of-the-art word embedding methods to measure the semantic similarity of terms.

3 TASK DEFINITION

In our semi-supervised learning case, we assume that we have a set of “fully labeled” categories and “weakly labeled” categories. For the “fully labeled” categories, a large number of training images with both image-level labels and bounding box annotations are available for learning the object detectors. For each of the “weakly labeled” categories, we have many training images containing the target object, but we do not have access to the exact locations of the objects. This is different from the semi-supervised learning proposed in previous work [55], [56], [57], where typically a small amount of fully labeled data with a large amount of weakly labeled (or unlabeled) data are provided for each category. In our semi-supervised object detection scenario, the objective is to transfer the trained image classifiers into object detectors on the “weakly labeled” categories.

4 SIMILARITY-BASED KNOWLEDGE TRANSFER

We first describe the Large Scale Detection through Adaptation (LSDA) framework [13], upon which our proposed approach is based (Section 4.1). We then describe our proposed knowledge transfer models with the aim of improving LSDA. Two knowledge domains are explored: (i) visual similarity (Section 4.2); (ii) semantic relatedness (Section 4.3). Next, we combine both models to obtain our mixture transfer model, as presented in Section 4.4. Finally, we propose to transfer the knowledge to bounding-box regression from fully labeled categories to weakly labeled categories in Section 4.5.

4.1 Background on LSDA

Let \mathcal{D} be the dataset of K categories to be detected. One has access to both image-level and bounding box annotations only for a set of m ($m \ll K$) “fully labeled” categories, denoted as \mathcal{B} , but only image-level annotations for the rest of the categories, namely “weakly labeled” categories, denoted as \mathcal{A} . Hence, a set of K image classifiers can be trained on the whole dataset \mathcal{D} ($\mathcal{D} = \mathcal{A} \cup \mathcal{B}$), but only m object detectors (from \mathcal{B}) can be learned according to the availability of bounding box annotations. The LSDA algorithm learns to convert $(K - m)$ image classifiers (from \mathcal{A}) into their corresponding object detectors through the following steps:

Pre-training: First, an 8-layer (5 convolutional layers and 3 fully-connected (*fc*) layers) *Alex-Net* [1] CNN is pre-trained on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 classification dataset [11], which contains 1.2 million images of 1,000 categories.

Fine-tuning for classification: The final weight layer (1,000 linear classifiers) of the pre-trained CNN is then replaced with K linear classifiers. This weight layer is randomly initialized and the whole CNN is then fine-tuned on the dataset \mathcal{D} . This produces a classification network that can classify K categories (*i.e.*, K -way softmax classifier), given an image or an image region as input.

Category-invariant adaptation: Next, the classification network is fine-tuned into a detector with bounding boxes of \mathcal{B} as input, using the R-CNN [4] framework. As in R-CNN, a background class ($fc8_{BG}$) is added to the output layer and fine-tuned using bounding boxes from a region proposal algorithm, e.g., Selective Search [32]. The $fc8$ layer parameters are category *specific*, with 4,097 weights ($fc7$ output: 4,096, plus a bias term) in each category, while the parameters of layers 1-7 are category *invariant*. Note that object detectors are not able to be directly trained on \mathcal{A} , since the fine-tuning and training process requires bounding box annotations. Therefore, at this point, the category specific output layer $fc8_{\mathcal{A}}$ stays unchanged. The variation matrix of $fc8_{\mathcal{B}}$ after fine-tuning is denoted as $\Delta_{\mathcal{B}}$.

Category-specific adaptation: Finally, each classifier of categories $j \in \mathcal{A}$ is adapted into a corresponding detector by learning a category-specific transformation of the model parameters. This is based on the assumption that the difference between classification and detection of a target object category has a positive correlation with those of similar (close) categories. The transformation is computed by adding a bias vector to the weights of $fc8_{\mathcal{A}}$. This bias vector for category j is measured by the average weight change of its k nearest neighbor categories in set \mathcal{B} , from classification to detection.

$$\forall j \in \mathcal{A} : \vec{w}_j^d = \vec{w}_j^c + \frac{1}{k} \sum_{i=1}^k \Delta_{\mathcal{B}_i^j} \quad (1)$$

where $\Delta_{\mathcal{B}_i^j}$ is the $fc8$ weight variation of the i^{th} nearest neighbor category in set \mathcal{B} for category $j \in \mathcal{A}$. \vec{w}_j^c and \vec{w}_j^d are, respectively, $fc8$ layer weights for the fine-tuned classification and the adapted detection network. The nearest neighbor categories are defined as those with nearest L_2 -norm (Euclidean distance) of $fc8$ weights in set \mathcal{B} .

The fully adapted network is able to detect all K categories in test images. In contrast to R-CNN, which trains SVM classifiers on the output of the $fc7$ layer followed by bounding box regression on the extracted features from the $pool5$ layer of all region proposals, LSDA directly outputs the score of the softmax “detector”, and subtracts the background score from this as the final score. This results in a small drop in performance, but enables direct adaptation from a classification network into a detection network on the “weakly labeled” categories, and significantly reduces the training time.

Hoffman et al. [13] demonstrated that the adapted model yielded a 50% relative mAP (mean average precision) boost for detection over the classification-only framework on the “weakly labeled” categories of the ILSVRC2013 detection dataset (from 10.31% to 16.15%). They also showed that category-specific adaptation (final LSDA step) contributes least to the performance improvement (16.15% with *vs.* 15.85% without this step), with the other features (adapted layers 1-7 and background class) being more important. However, we found that by properly adapting this layer, a significant boost in performance can be achieved: an mAP of 22.03% can be obtained by replacing the semi-supervised $fc8_A$ weights with their corresponding supervised network weights and leaving the other parameters fixed. Thus, we believe that adapting this layer in an informed manner, such as making better use of knowledge about object similarities, will help improve detection.

In the next subsections, we will introduce our knowledge transfer methods using two different kinds of similarity measurements to select the nearest categories and weight them accordingly to better adapt the $fc8$ layer, which can efficiently convert an image classifier into an object detector for a “weakly labeled” category.

4.2 Knowledge Transfer via Visual Similarity

Intuitively, the object detector of an object category may be more similar to those of visually similar categories than of visually distinct categories. For example, a cat detector may approximate a dog detector better than a strawberry detector, since cat and dog are both mammals sharing common attributes in terms of shape (both have four legs, two ears, two eyes, one tail) and texture (both have fur). Therefore, given a “fully labeled” dataset \mathcal{B} and a “weakly labeled” dataset \mathcal{A} , our objective is to model the visual similarity between each category $j \in \mathcal{A}$ and all the other categories in \mathcal{B} , and to transfer this knowledge for transforming classifiers into detectors for \mathcal{A} .

Visual similarity measure: Visual similarity measurements are often obtained by computing the distance between feature distributions such as the $fc6$ or $fc7$ output of a CNN, or in the case of LSDA the $fc8$ layer parameters. In our work, we instead forward propagate an image through the whole fine-tuned classification network (created by the second step in Section 4.1) to obtain a K -dimensional classification score vector. This score vector encodes the probabilities of an image being each of the K object categories. Consequently, for all the positive images of an object category $j \in \mathcal{A}$, we can directly accumulate the scores of each dimension, on a balanced validation dataset. We assume that the normalized

accumulated scores (range [0,1]) imply the similarities between category j and other categories: the larger the score, the more it visually resembles category j . This assumption is supported by the analysis of deep CNNs [58], [59], [60]: CNNs are apt to confuse visually similar categories, on which they might have higher prediction scores. The visual similarity (denoted s_v) between a “weakly labeled” category $j \in \mathcal{A}$ and a “fully labeled” category $i \in \mathcal{B}$ is defined as:

$$s_v(j, i) \propto \frac{1}{N} \sum_{n=1}^N CNN_{softmax}(I_n)_i \quad (2)$$

where I_n is a positive image from category j of the validation set of \mathcal{A} , N is the number of positive images for this category, and $CNN_{softmax}(I_n)_i$ is the i^{th} CNN output of the softmax layer on I_n , namely, the probability of I_n being category $i \in \mathcal{B}$ as predicted by the fine-tuned classification network. $s_v(j, i) \in [0, 1]$ is the degree of similarity after normalization on all the categories in \mathcal{B} .

Note that we adopt the $fc8$ outputs (classification scores) since most of the computation is integrated into the end-to-end *Alex-Net* framework except for the accumulation of classification scores in the end, saving the extra effort otherwise required for distance computation if $fc6$ or $fc7$ outputs were to be used. The idea of using the L_2 distance of the $fc8$ weights (linear classifier parameters) as a visual similarity measurement in LSDA is closely related to ours. However, in addition to the $fc8$ weights, our visual similarity measurement is assumed to leverage the powerful and supplementary feature representations generated by the prior layers of the neural network by combining both, given the fact that the $fc8$ outputs are obtained by taking the inner-product of $fc7$ outputs (visual features) and $fc8$ weights. Experimental results in Section 5 validate this intuition.

Weighted nearest neighbor scheme: Using Eq. (1), we can transfer the model parameters based on a category’s k nearest neighbor categories selected by Eq. (2). This allows us to directly compare our visual similarity measure to that of LSDA which uses the Euclidean distance between the $fc8$ parameters. An alternative to Eq. (1) is to consider a *weighted* nearest neighbor scheme, where weights can be assigned to different categories based on how visually similar they are to the target object category. This is intuitive, as different categories will have varied degrees of similarity to a particular class, and some categories may have only a few (or many) visually similar classes. Thus, we modify Eq. (1) and define the transformation via visual similarity based on the proposed weighted nearest neighbor scheme as:

$$\forall j \in \mathcal{A} : \vec{w}_{j_v}^d = \vec{w}_j^k + \sum_{i=1}^m s_v(j, i) \Delta_{\mathcal{B}_i^j} \quad (3)$$

It is worth noting that Eq. (1) is a special case of Eq. (3), where $m = k$ and $s_v(j, i) = 1/k$.

4.3 Knowledge Transfer via Semantic Relatedness

Following prior work [15], [41], [61], we observe that visual similarity is correlated with semantic relatedness. According to [15], this relationship is particularly strong when measurements are focused on the category instances themselves,

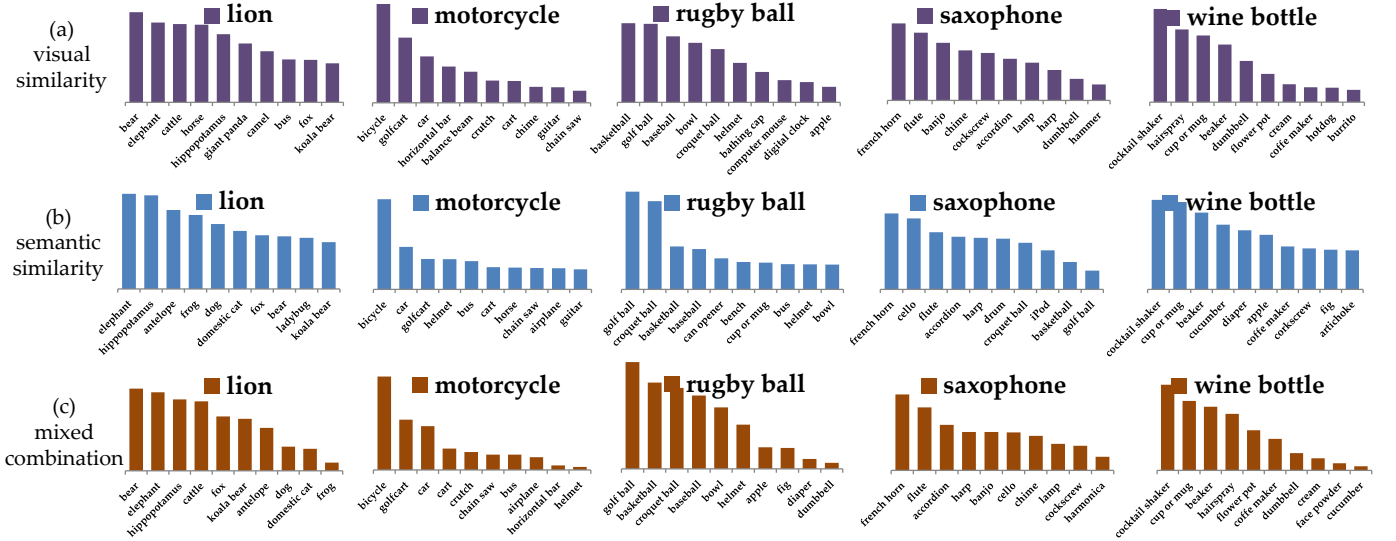


Fig. 3. Some example visualizations of (a) visual similarity (first row in the figure), (b) semantic similarity (middle row) and (c) mixture similarity (last row) between a target “weakly labeled” category and its source categories from which knowledge is to be transferred. For each target category, the top-10 weighted nearest neighbor categories are shown. The magnitude of each column bar shows the relative weight (degree of similarity s_v , s_s , s in Eq. (6), where α is set to 0.6).

ignoring image backgrounds. This observation is quite intriguing for object detection, where the main focus is on the target objects themselves. Hence, we draw on this fact and propose transferring knowledge from the natural language domain to help improve semi-supervised object detection.

Semantic similarity measure: As mentioned in Section 2.3, we use word embeddings to represent each category and to measure the semantic similarity between categories. Each of the K categories is embedded as a word vector, more specifically a 300-dimensional word2vec embedding [52]. Since each category is a WordNet [48] synset, we represent each category as the sum of the word vectors for each term in its synset, normalized to unit vector by its L_2 -norm. Out-of-vocabulary words are addressed by attempting to match case variants of the words (lowercase, Capitalized), e.g., “aeroplane” is not in the vocabulary, but “Aeroplane” is. Failing that, we represent multiword phrases by the sum of the word vectors of each in-vocabulary word of the phrase, normalized to unit vector (“baby”+“bed” for *baby bed*). In several cases, we also augment synset terms with any category label defined in ILSVRC2013 that is not among the synset terms defined in WordNet (e.g. “bookshelf” for the WordNet synset *bookcase*, and “tv” and “monitor” for *display*).

Word embeddings often conflate multiple senses of a word into a single vector, leading to an issue with polysemous words. We observed this with many categories, for example *seal* (animal) is close to *nail* and *tie* (which, to further complicate matters, is actually meant to refer to its clothing sense); or the stationery *ruler* being related to *lion*. Since ILSVRC2013 categories are actually WordNet synsets, it makes perfect sense to exploit WordNet to help disambiguate the word senses. Thus, we integrate corpus-based representations with semantic knowledge from WordNet, by using AutoExtend [62] to encode the categories as *synset embeddings* in the original word2vec embedding space. Au-

toExtend exploits the interrelations between synsets, words and lexemes to learn an auto-encoder based on these constraints, as well as constraints on WordNet relations such as hypernyms (encouraging *poodle* and *dog* to have similar embeddings). We observed that AutoExtend has indeed helped form better semantic relations between the desired categories: *seal* is now clustered with other animal categories like *whale* and *turtle*, and the nearest neighbors for *ruler* are now *rubber eraser*, *power drill* and *pencil box*. In our detection experiments (Section 5), we found that while the ‘naive’ word embeddings performed better than the baselines, the synset embeddings yielded even better results. Thus, we concentrate on reporting the results of the latter.

We represent each category $j \in \mathcal{A}$ and $i \in \mathcal{B}$ with their synset embeddings, and compute the L_2 -norm of each pair $d_s(j, i)$ as their semantic distance. The semantic similarity $s_s(j, i)$ is inversely proportional to $d_s(j, i)$. We then transfer the semantic knowledge to the appearance model using Eq. (3) or its special case Eq. (1) as before.

As our semantic representations are in the form of vectors, we explore an alternative similarity measure as used in [41]. We assume that each vector of a “weakly labeled” category $j \in \mathcal{A}$ (denoted as v_j) can be approximately represented by a linear combination of all the m word vectors in \mathcal{B} : $v_j \approx \Gamma_j V$, where $V = [v_1; v_2; \dots; v_i; \dots; v_m]$, and $\Gamma_j = [\gamma_j^1, \gamma_j^2, \dots, \gamma_j^i, \dots, \gamma_j^m]$ is a set of coefficients of the linear combination. We are motivated to find the solution Γ_j^* which contains as few non-zero components as possible, since we tend to reconstruct category j with fewer categories from \mathcal{B} (sparse representation). This optimal solution Γ_j^* can be formulated as the following optimization:

$$\Gamma_j^* = \arg \min_{\Gamma_j > 0} (\|v_j - \Gamma_j V\|_2 + \lambda \|\Gamma_j\|_0) \quad (4)$$

Note that $\Gamma_j > 0$ is a positive constraint on the coefficients, since negative components of sparse solutions for semantic transferring are meaningless: we only care about the most

similar categories and not dissimilar categories. We solve Eq. (4) by using the positive constraint matching pursuit (PCMP) algorithm [63]. Therefore, the final transformation via semantic transferring is formulated as:

$$\forall j \in \mathcal{A} : \vec{w}_j^d = \vec{w}_j^c + \sum_{i=1}^m s_s(j, i) \Delta_{\mathcal{B}_i^j} \quad (5)$$

where $s_s(j, i) = \gamma_j^i$ in the sparse representation case.

4.4 Mixture Transfer Model

We have proposed two different knowledge transfer models. Each of them can be integrated into the LSDA framework independently. In addition, since we consider the visual similarity at the whole image level and the semantic relatedness at object level, they can be combined simultaneously to provide complementary information. We use a simple but very effective combination of the two knowledge transfer models as our final mixture transfer model. Our mixture model is a linear combination of the visual similarity and the semantic similarity:

$$s = \text{intersect}[\alpha s_v + (1 - \alpha) s_s] \quad (6)$$

where $\text{intersect}[\cdot]$ is a function that takes the intersection of cooccurring categories between visual and sparse semantic related categories. $\alpha \in [0, 1]$ is a parameter used to control the relative influence of the two similarity measurements. α is set to 1 when only considering visual similarity transfer, and 0 for the semantic similarity transfer. We will analyze this parameter in Section 5.3.

4.5 Transfer on Bounding-box Regression

The detection windows generated by the region based detection models are the highest scoring proposals (e.g., Selective Search). In order to improve localization performance, a bounding-box regression stage [4] is commonly adopted to post-process the detection windows. This process needs bounding box annotations in training the regressors, which is an obstacle for “weakly labeled” categories in our case. Hence, we propose to transfer the class-specific regressors from “fully labeled” categories to “weakly labeled” categories based on the aforementioned similarity measures.

To train a regressor for each “fully labeled” category, we select a set of N training pairs $\{(\vec{P}^i, \vec{G}^i)\}_{i=1, \dots, N}$, where $\vec{P}^i = (P_x^i, P_y^i, P_w^i, P_h^i)$ is a vector indicating the center coordinates (P_x^i, P_y^i) of proposal P^i together with P^i 's width and height (P_w^i, P_h^i) . $\vec{G}^i = (G_x^i, G_y^i, G_w^i, G_h^i)$ is the corresponding ground-truth bounding box. Except where needed to avoid confusion we omit the superscript i . The goal is to learn a mapping function $f(P) = (f_x(P), f_y(P), f_w(P), f_h(P))$ which maps a region proposal P to a ground-truth window G . Each function within $f(P)$ is modeled as a linear function of the *pool5* features (namely the feature map after the last convolutional and pooling block of the ConvNet): $f(P) = \mathbf{w}_*^T F_5(P)$, where \mathbf{w}_* is a vector of learnable parameters, $F_5(P)$ is the *pool5* feature of region proposal P . \mathbf{w}_* can be learned by optimizing the following least squares objective function:

$$\mathbf{w}_* = \arg \min_{\mathbf{w}_*} \sum_{i=1}^N (\hat{\mathbf{w}}_*^T F_5(P^i) - t_*^i)^2 + \lambda_0 \|\hat{\mathbf{w}}_*\|^2 \quad (7)$$

where $t_* = (t_x, t_y, t_w, t_h)$ is the regression target for the training pair (P, G) defined as:

$$\begin{aligned} t_x &= (G_x - P_x)/P_w, \\ t_y &= (G_y - P_y)/P_h, \\ t_w &= \log(G_w/P_w), \\ t_h &= \log(G_h/P_h). \end{aligned} \quad (8)$$

The first two equations specify a scale-invariant translation of the center of the bounding box, while the remaining two specify the log-space translation of the width and height of the bounding box. After learning the parameters of the transformation function, a detection window (region proposal) P can be transformed into a new prediction $\hat{P} = (\hat{P}_x, \hat{P}_y, \hat{P}_w, \hat{P}_h)$ by applying:

$$\begin{aligned} \hat{P}_x &= P_x + P_w f_x(P), \\ \hat{P}_y &= P_y + P_h f_y(P), \\ \hat{P}_w &= P_w \exp(f_w(P)), \\ \hat{P}_h &= P_h \exp(f_h(P)). \end{aligned} \quad (9)$$

The training pair (P, G) is selected if the proposal P has maximum IoU overlap with ground-truth bounding box G . The pair (P, G) is discarded if the maximum IoU overlap is less than a threshold (which is set to be 0.6 using a validation set).

For a “weakly labeled” category j , the transformation function cannot be explicitly learned due to the absence of ground-truth bounding boxes. However, we can still transfer this knowledge from similar categories in the “fully labeled” subset \mathcal{B} :

$$\forall j \in \mathcal{A} : \mathbf{w}_j = \sum_{i=1}^m s_* \mathbf{w}_i \quad (10)$$

where s_* indicates any one of the aforementioned similarity measures.

5 EXPERIMENTS

5.1 Dataset Overview

We investigate the proposed knowledge transfer models for large scale semi-supervised object detection on the ILSVRC2013 detection dataset covering 200 object categories. The training set is not exhaustively annotated because of its sheer size. There are also fewer annotated objects per training image than the validation and testing image (on average 1.53 objects for training *vs.* 2.5 objects for validation set). We follow all the experiment settings as in [13], and simulate having access to image-level annotations for all 200 categories and bounding box annotations only for the first 100 categories (alphabetical order). We separate the dataset into classification and detection sets. For the classification data, we use 200,000 images in total from all 200 categories of the training subset (around 1,000 images per category) and their image-level labels. The validation set is roughly split in half: val1 and val2 as in [4]. For the detection training set, we take the images with their bounding boxes from only the first 100 categories (\mathcal{B}) in val1 (around 5,000 images in total). Since the validation dataset is relatively small, we then augment val1 with 1,000 bounding box annotated images per class from the training set (following the same protocol of [4], [13]). Finally, we evaluate our knowledge

TABLE 1

Detection mean average precision (mAP) on ILSVRC2013 val2. The first row shows the basic performance of directly using all classification parameters for detection, without adaptation or knowledge transfer (*i.e.*, weakly supervised learning). The last row shows results of an oracle detection network which assumes that bounding boxes for all 200 categories are available (*i.e.*, supervised learning). The second row shows the baseline LSDA results using only feature adaptation. Rows 3-5 show the performance of LSDA for adapting both the feature layers (layer 1-7) and the class-specific layer (layer 8), by considering different numbers of neighbor categories. Rows 6-8, 9-12 and row 13 show the results of our visual transfer, semantic transfer and mixture transfer model, respectively. Row 14 shows our results after bounding-box regression. For all methods, the same “*AlexNet*” CNN is adopted.

Method	Number of Nearest Neighbors	mAP on \mathcal{B} : “Fully labeled” 100 Categories	mAP on \mathcal{A} : “Weakly labeled” 100 Categories	mAP on \mathcal{D} : All 200 Categories
Classification Network	-	12.63	10.31	11.90
LSDA (only class invariant adaptation)	-	27.81	15.85	21.83
LSDA (class invariant & specific adapt)	avg/weighted - 5	28.12 / -	15.97 / 16.12	22.05 / 22.12
	avg/weighted - 10	27.95 / -	16.15 / 16.28	22.05 / 22.12
	avg/weighted - 100	27.91 / -	15.96 / 16.33	21.94 / 22.12
Ours (visual transfer)	avg/weighted - 5	27.99 / -	17.42 / 17.59	22.71 / 22.79
	avg/weighted - 10	27.89 / -	17.62 / 18.41	22.76 / 23.15
	avg/weighted - 100	28.30 / -	17.38 / 19.02	22.84 / 23.66
Ours (semantic transfer)	avg/weighted - 5	28.01 / -	17.32 / 17.53	22.67 / 22.77
	avg/weighted - 10	28.00 / -	16.67 / 17.50	22.31 / 22.75
	avg/weighted - 100	28.14 / -	17.04 / 18.32	23.23 / 23.28
	Sparse rep. - ≤ 20	28.18	19.04	23.66
Ours (mixture transfer)	-	28.04	20.03 $\uparrow 3.88$	24.04
Ours (mixture transfer + BB reg.)	-	31.85	21.88	26.87
Oracle: Full Detection Network (no BB reg.)	-	29.72	26.25	28.00
Oracle: Full Detection Network (BB reg.)	-	32.17	29.46	30.82

transfer framework on the val2 dataset (9,917 images in total).

5.2 Implementation Details

In all the experiments, we consider LSDA [13] as our baseline model and follow their main settings. Following [13], we first use the Caffe [59] implementation of the “*AlexNet*” CNN. A pre-trained CNN on ILSVRC 2012 dataset is then fine-tuned on the classification training dataset (see Section 5.1). This CNN is then fine-tuned again for detection on the labeled region proposals of the first 100 categories (subset \mathcal{B}) of val1. Selective Search [32] with “fast” mode is adopted to generate the region proposals from all the images in val1 and val2. We also report results using two deeper models of “*VGG-Nets*” [64], namely, the 16-layer model (*VGG-16*) and the 19-layer model (*VGG-19*), *GoogLeNet* [65] and two *ResNets* [66] (34-layer and 50-layer) with the Caffe toolbox. For the semantic representation, we use word2vec CBoW embeddings pre-trained on part of the Google News dataset comprising about 100 billion words [52]. We train AutoExtend [62] using WordNet 3.0 to obtain synset embeddings, and using equal weights for the synset, lexeme and WordNet relation constraints ($\alpha = \beta = 0.33$). As all categories are nouns, we use only hypernyms as the WordNet relation constraint. For the sparse representation of a target word vector in Eq. (4), we limit the maximum number of non-zero components to 20, since a target category has strong correlation with a small number of source categories. We set $\lambda = 100$ in Eq. (4) and $\lambda_0 = 1000$ in Eq. (7) based on a

validation set. Other detailed information regarding training and detection can be found in Section 4.1.

5.3 Quantitative Evaluation on the “Weakly Labeled” Categories with “*AlexNet*”

Setting LSDA as the baseline, we compare the detection performance of our proposed knowledge transfer methods against LSDA. The results are summarized in Table 1. As we are concerned with the detection of the “weakly labeled” categories, we focus mainly on the second column of the table (mean average precision (mAP) on \mathcal{A}). Rows 1-5 in Table 1 are the baseline results for LSDA. The first row shows the detection results by applying a classification network (*i.e.*, weakly supervised learning, and without adaptation) trained with only classification data, achieving only an mAP of 10.31% on the “weakly labeled” 100 categories. The last row shows the results of an oracle detection network which assumes that bounding boxes for all 200 categories are available (*i.e.*, supervised learning). This is treated as the upper bound (26.25%) of the fully supervised framework. We observed that the best result obtained by LSDA is to adapt both category independent and category specific layers, and transforming with the weighted *fc8* layer weight change of its 100 nearest neighbor categories (**weighted-100** with 16.33% in Table 1). Our “*weighted*” scheme works consistently better than its “*average*” counterpart.

For our **visual knowledge transfer model**, we show steady improvement over the baseline LSDA methods when considering the average weight change of both 5 and 10

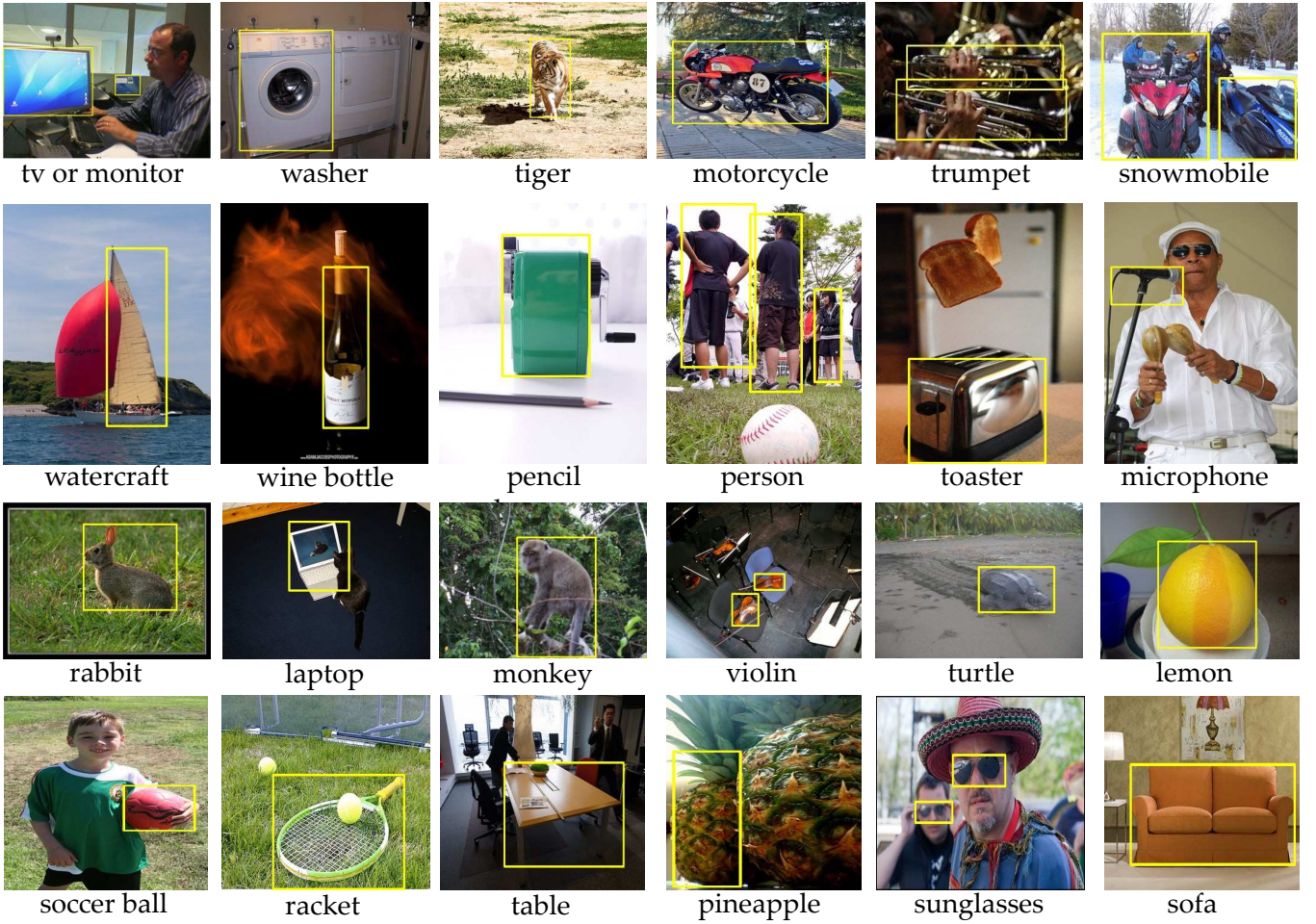


Fig. 4. Examples of correct detections (true positives) of our mixture knowledge transfer model on ILSVRC2013 images. For each image, only detections for the “weakly labeled” target category (text below image) are listed.

TABLE 2
Comparison of mean average precision (mAP) for semantic similarity measures/representations, using **Weighted - 100**.

Method	Path Similarity	Lin Similarity	Naive Embeddings	AutoExtend (this paper)
mAP	17.08	17.31	17.83	18.32

visually similar categories, with 1.45% and 1.47% increase in mAP, respectively. This proves that our proposed visual similarity measure is superior to that of LSDA, showing that category-specific adaptation can indeed be improved based on knowledge about the visual similarities between categories. Further improvement is achieved by modeling individual weights of all 100 source categories according to their degree of visual similarities to the target category (**weighted-100** with 19.02% in the table). This verifies our supposition that the transformation from a classifier to a detector of a certain category is more related to visually similar categories, and is proportional to their degrees of similarity. For example, *motorcycle* is most similar to *bicycle*. Thus the weight change from a *bicycle* classifier to detector has the largest influence on the transformation of *motorcycle*. The influence of less visually relevant categories, such as

cart and *chain saw*, is much smaller. For visually dissimilar categories (*apple*, *fig*, *hotdog*, etc.), the influence is extremely insignificant. We show some examples of visual similarities between a target category and its source categories in the first row of Fig. 3. For each target category, the top-10 weighted nearest neighbor categories with their similarity degrees are visualized.

Our **semantic knowledge transfer model** also showed marked improvement over the LSDA baseline (Table 1, Rows 9-12), and is comparable to the results of the visual transfer model. This suggests that the cross-domain knowledge transfer from semantic relatedness to visual similarity is very effective. The best performance for the semantic transfer model (19.04%) is obtained by sparsely reconstructing the target category with the source categories using the synset embeddings. We also compare the results of using other semantic similarity measures, as shown in Table 2. The result of using synset embeddings (18.32%, using weighted-100, the same below) are superior to using ‘naive’ word2vec embeddings (17.83%) and WordNet based measures such as path-based similarity (17.08%) and Lin similarity [51] (17.31%). Several examples visualizing the related categories of the 10 largest semantic reconstruction coefficients are shown in the middle row of Fig. 3. We observe that semantic

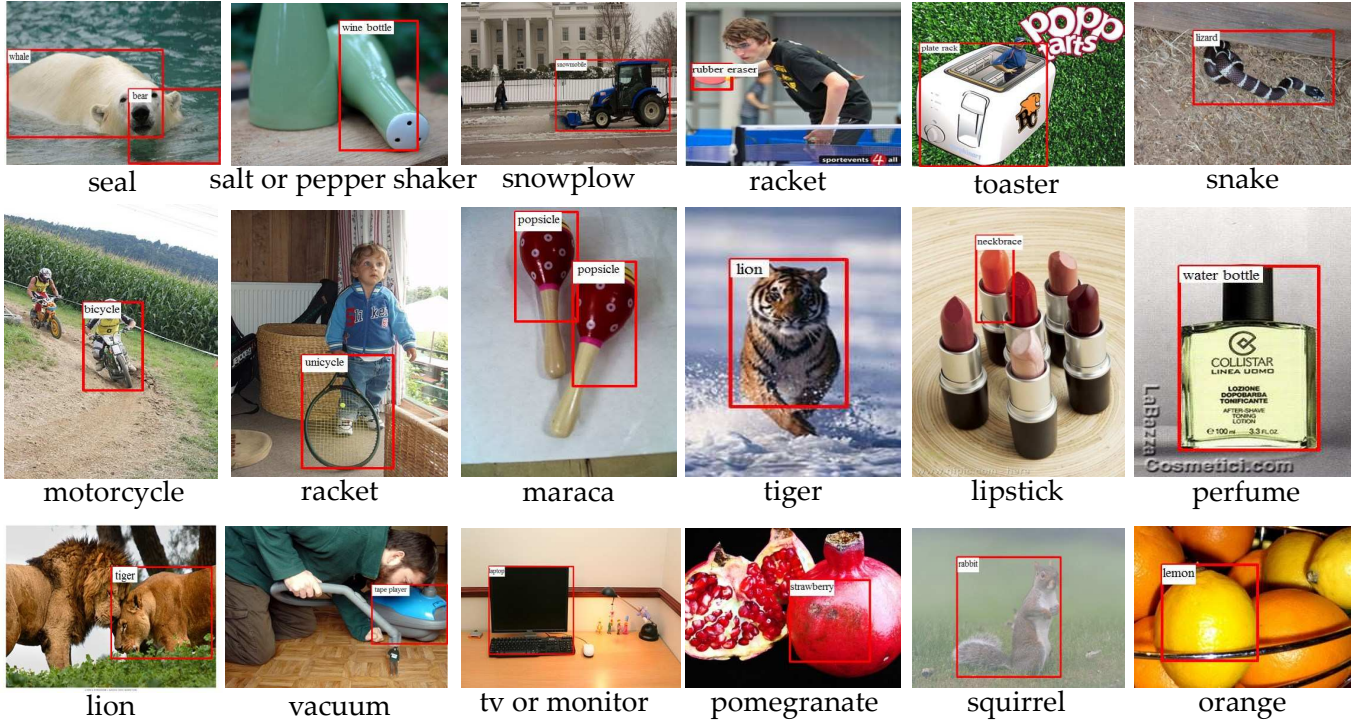


Fig. 5. Examples of incorrect detections (confusion with other objects) of our mixture knowledge transfer model on ILSVRC2013 images. The detected object label is shown in the top-left of its bounding box.

relatedness indeed correlates with visual similarity.

The state-of-the-art result using the 8-layer “Alex-Net” for semi-supervised detection on this dataset is achieved by our **mixture transfer model** which combines visual similarity and semantic relatedness. A boost in performance of 3.88% on original split ($3.82\% \pm 0.12\%$, based on 6 different splits of the dataset) is achieved over the best result reported by LSDA on the “weakly labeled” categories. We show examples of transferred categories with their corresponding weights for several target categories in the bottom row of Fig. 3. The parameter α in Eq. (6) for the mixture model weights is set to 0.6 for final detection, where $\alpha \in \{0, 0.2, 0.4, 0.5, 0.6, 0.8, 1\}$ is chosen via cross-validation on the val1 detection set (Fig. 6). This suggests that the transferring of visual similarities is slightly more important than semantic relatedness, although both are indeed complementary. We do not tune α for each category separately, though this can be expected to further improve our detection performance. Fig. 4 and Fig. 5 show some examples of correct and incorrect detections respectively. Although our proposed mixture transfer model achieves the state of the art in detecting the “weakly labeled” categories, it is still occasionally confused by visually similar categories.

5.4 Experimental Results with Deeper Neural Networks

Previous work [64], [66], [67] found that region based CNN detection performance is significantly influenced by the choice of CNN architecture. In Table 3, we show some detection results using the 16-layer and 19-layer deep “VGG-Nets” proposed by Simonyan and Zisserman [64], “GoogLeNet” (Inception-v2 [65]), together with the 34-layer and 50-layer “ResNets” [66]. The VGG-16 network consists

of 13 convolutional layers of very small (3×3) convolution filters, with 5 max pooling layers interspersed, and topped with 3 fully connected layers (namely, $fc6$, $fc7$ and $fc8$). The VGG-19 network extends VGG-16 by inserting 3 more convolutional layers, while keeping other layer configurations unchanged. The state-of-the-art residual networks (ResNets) make use of identity shortcut connections that enable flow of information across layers without decay. In the aforementioned deep neural nets, we transfer the parameters of the last fully connected layer ($fc8$ layer for VGG-Nets and the only fc layer in GoogLeNet and ResNets).

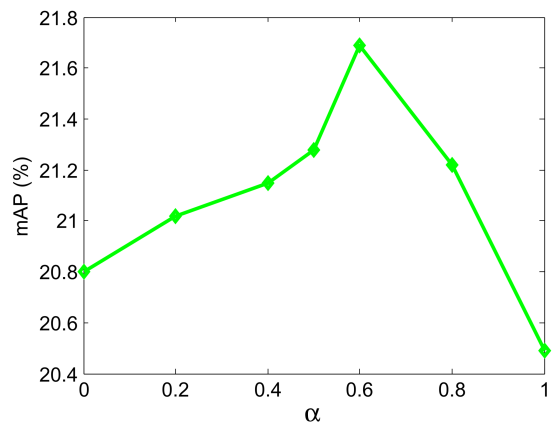


Fig. 6. Sensitivity of parameter α vs. mAP for detection of “weakly labeled” categories on the validation (val1) dataset. $\alpha \in [0, 1]$ is a parameter used to control the relative influence of the two similarity measurements. α is set to 1 when only considering visual similarity transfer, and 0 for semantic similarity transfer.

TABLE 3

Comparisons of detection mean average precision (mAP) on the “weakly labeled” categories of ILSVRC2013 val2, using “**VGG-Nets**”, “**GoogLeNet**” and “**ResNets**”. For LSDA, our visual similarity and semantic relatedness transfer models, **Weighted - 100** scheme is adopted.

Method	Only classification	LSDA class invariant	LSDA class invariant & specific	Ours visual	Ours semantic	Ours mixed	Ours mixed + BB reg.
<i>Alex-Net</i>	10.31	15.85	16.33	19.02	18.32	20.03	21.88
<i>VGG-16</i>	14.89	18.24	18.86	21.75	21.07	23.21	24.91
<i>VGG-19</i>	16.22	20.38	21.02	23.89	23.10	25.07	27.32
<i>GoogLeNet</i>	16.12	–	–	23.62	23.46	25.06	26.62
<i>ResNet-34</i>	17.09	–	–	24.28	24.64	26.17	28.05
<i>ResNet-50</i>	17.34	–	–	24.94	24.75	26.77	28.30

As can be seen from Table 3, the very deep ConvNets *VGG-16*, *VGG-19* and *GoogLeNet* significantly outperform *Alex-Net* for all the adaptation methods. Our knowledge transfer models using the very deep *VGG-Nets* with different similarity measures show consistent improvement over the LSDA baseline method. The relative overall improvement over performance using *VGG-Nets* is similar with that of *AlexNet*. *GoogLeNet* obtains similar results to *VGG-19*, while the best performance is achieved by *ResNet-50*.

5.5 Experimental Results with Bounding-box Regression

The results in Table 1 and Table 3 show that the transferred bounding-box regression from “fully labeled” categories fixes a large number of incorrectly localized detections, boosting mAP by about 2 points for the “weakly labeled” categories. The bounding-box regression process could boost mAP by 3 to 4 points if the bounding box annotations for all the categories were provided. We show some example detections before and after bounding box regression on the “weakly labeled” categories in Fig. 7, using *VGG-16*.

In addition to the results reported above using the default Intersection-over-Union (IoU) threshold 0.5, we evaluate detection performance by setting different IoU overlap ratios {0.6, 0.7} before and after bounding box regression using the best performing network, i.e., *ResNet-50*. The mAP@IoU=0.6 is 23.21 *v.s.* 26.02 before and after regression, mAP@IoU=0.7 is 17.89 *v.s.* 20.64, respectively. These results validate that the proposed regression transfer is very effective in moving bounding box boundaries so as to cover more foreground object regions, by transferring the class-specific regressors from “fully labeled” categories to “weakly labeled” categories based on the proposed similarity measures. Note that results in other parts of this paper are reported as mAP@IoU=0.5, unless specified otherwise.

5.6 Experimental Results with Fast R-CNN

The proposed knowledge transfer method is applied to Fast R-CNN [6], without modifying much of the framework. Our Fast R-CNN based transfer framework is much faster than the R-CNN based approach, since in Fast R-CNN, in Fast R-CNN, an image is first fed into a CNN to create a convolution feature map and a single feature vector is then extracted from a Region of Interest (RoI) pooling layer

TABLE 4
Detection performance using Fast R-CNN on the “weakly labeled” categories of ILSVRC2013 val2.

Method	<i>Alex-Net</i>	<i>VGG-16</i>	<i>ResNet-50</i>
2-stage	22.36	25.47	29.09
end-to-end	22.79	26.22	29.71

for each region proposal, while in R-CNN, each region proposal in an image is fed into a CNN to extract feature independently, which is considerably more computationally expensive than Fast R-CNN.

We investigate two different bounding box regression strategies in Fast R-CNN. For the first strategy, we remove the built-in bounding-box regression layer in the Fast R-CNN pipeline and transfer the regressor off-line after detection as in Section 4.5, like the *SPP-Net* [5], which we call “2-stage” Fast R-CNN. For the second strategy, we use the built-in bounding-box regression layer, which is actually a class-specific fully-connected layer with $4C$ neurons (4 indicates the coordinates of a bounding box position, C indicates the number of categories). This class-specific layer can be therefore transferred from “fully labeled” categories to “weakly labeled” categories using the proposed similarity measures in a way similar to that described in Section 4.5. This strategy is called “end-to-end”. Table 4 shows the detection performance using these Fast R-CNN based ConvNets. As can be seen from the table, Fast R-CNN achieves consistently better performance over the R-CNN based approach, and the “end-to-end” joint training/testing is superior to the “2-stage” pipeline. State-of-the-art detection performance on the weakly labeled categories (29.71) is obtained by Fast R-CNN based *ResNet-50*, which is very close to that of the fully supervised R-CNN based *AlexNet* (30.82).

6 CONCLUSION

In this paper, 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 in a semi-supervised setting. We experimented with different CNN architectures, found clear evidence that both visual and semantic similarities play an essential role in improving the adaptation process, and that the combination of the two modalities yielded state-of-the-art performance,

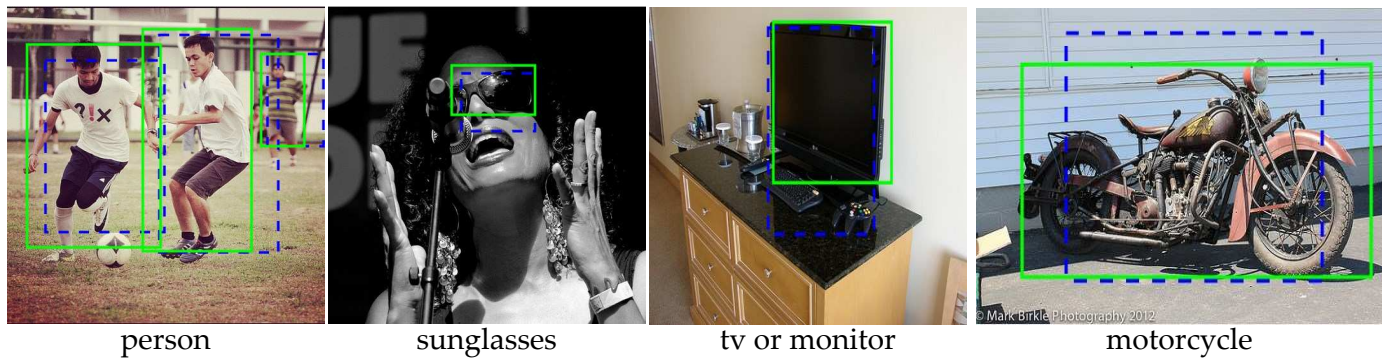


Fig. 7. Some example detections before and after bounding box regression on the “weakly labeled” categories. Boxes before (resp. after) bounding-box regression are shown in dashed blue (resp. green).

suggesting that knowledge inherent in visual and semantic domains is complementary. Future work includes extracting more knowledge from different domains, using better representations, and investigating the possibility of using category-invariant properties, e.g., the difference between feature distributions of whole images and target objects, to help knowledge transfer. We believe that the combination of knowledge from different domains is key to improving semi-supervised object detection.

ACKNOWLEDGMENTS

This work was partly supported by the French Research Agency, Agence Nationale de Recherche (ANR), through the VideoSense Project under Grant 2009 CORD 026 02, and the Visen project under Grants ANR-12-CHRI-0002-04 and UK EPSRC EP/K019082/1 within the framework of the ERA-Net CHIST-ERA, and by the Partner University Foundation through the 4D Vision project. The authors thank NVIDIA for providing the Titan X GPU.

REFERENCES

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Neural Information Processing Systems (NIPS)*, 2012.
- [2] C. Szegedy, A. Toshev, and D. Erhan, “Deep neural networks for object detection,” in *Neural Information Processing Systems (NIPS)*, 2013.
- [3] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, “Overfeat: Integrated recognition, localization and detection using convolutional networks,” in *International Conference on Learning Representations (ICLR)*, 2014.
- [4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [5] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 37, no. 9, pp. 1904–1916, 2015.
- [6] R. Girshick, “Fast R-CNN: towards real-time object detection with region proposal networks,” in *International Conference on Computer Vision (ICCV)*, 2015.
- [7] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN,” in *Neural Information Processing Systems (NIPS)*, 2015.
- [8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. Berg, “SSD: Single shot multibox detector,” in *European Conference on Computer Vision (ECCV)*, 2016.
- [10] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (VOC) challenge,” *International Journal of Computer Vision (IJCV)*, vol. 88, no. 2, pp. 303–338, 2010.
- [11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision (IJCV)*, vol. 0, no. 0, pp. 1–42, April 2015.
- [12] T. Lin, M. Maire, S. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: common objects in context,” *CoRR*, vol. abs/1405.0312, 2014.
- [13] J. Hoffman, S. Guadarrama, E. Tzeng, R. Hu, J. Donahue, R. Girshick, T. Darrell, and K. Saenko, “LSDA: Large scale detection through adaptation,” in *Neural Information Processing Systems (NIPS)*, 2014.
- [14] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Object detectors emerge in deep scene CNNs,” in *International Conference on Learning Representations (ICLR)*, 2015.
- [15] T. Deselaers and V. Ferrari, “Visual and semantic similarity in imagenet,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.
- [16] Y. Tang, J. Wang, B. Gao, E. Dellandrea, R. Gaizauskas, and L. Chen, “Large scale semi-supervised object detection using visual and semantic knowledge transfer,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [17] D. Crandall and D. Huttenlocher, “Weakly supervised learning of part-based spatial models for visual object recognition,” in *European Conference on Computer Vision (ECCV)*, 2006.
- [18] O. Chum and A. Zisserman, “An exemplar model for learning object classes.”
- [19] C. Galleguillos, B. Babenko, A. Rabinovich, and S. Belongie, “Weakly supervised object recognition and localization with stable segmentations,” in *European Conference on Computer Vision (ECCV)*, 2008.
- [20] M. Nguyen, L. Torresani, F. de la Torre, and C. Rother, “Weakly supervised discriminative localization and classification: a joint learning process,” in *International Conference on Computer Vision (ICCV)*, 2009.
- [21] P. Siva and T. Xiang, “Weakly supervised object detector learning with model drift detection,” in *International Conference on Computer Vision (ICCV)*, 2011.
- [22] M. Pandey and S. Lazebnik, “Scene recognition and weakly supervised object localization with deformable part-based models,” in *International Conference on Computer Vision (ICCV)*, 2011.
- [23] P. Siva, C. Russell, and T. Xiang, “In defence of negative mining for annotating weakly labelled data,” in *European Conference on Computer Vision (ECCV)*, 2012.
- [24] T. Deselaers, B. Alexe, and V. Ferrari, “Weakly supervised localization and learning with generic knowledge,” *International Journal of Computer Vision (IJCV)*, vol. 100, no. 3, pp. 275–293, 2012.
- [25] Z. Shi, T. M. Hospedales, and T. Xiang, “Bayesian joint topic modelling for weakly supervised object localisation,” in *International Conference on Computer Vision (ICCV)*, 2013.
- [26] Y. Tang, X. Wang, E. Dellandrea, S. Masnou, and L. Chen, “Fusing generic objectness and deformable part-based models for weakly

- supervised object detection," in *IEEE International Conference on Image Processing (ICIP)*, 2014.
- [27] H. Bilen, M. Pedersoli, and T. Tuytelaars, "Weakly supervised object detection with posterior regularization," in *British Machine Vision Conference (BMVC)*, 2014.
- [28] H. O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell, "On learning to localize objects with minimal supervision," in *International Conference on Machine Learning (ICML)*, 2014.
- [29] H. Bilen, M. Pedersoli, and T. Tuytelaars, "Weakly supervised object detection with convex clustering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [30] C. Wang, K. Huang, W. Ren, J. Zhang, and S. Maybank, "Large-scale weakly supervised object localization via latent category learning," *IEEE Transactions on Image Processing (TIP)*, vol. 24, no. 4, pp. 1371–1385, April 2015.
- [31] Y. Tang, X. Wang, E. Dellandrea, and L. Chen, "Weakly supervised learning of deformable part-based models for object detection via region proposals," *IEEE Transactions on Multimedia (TMM)*, vol. PP, no. 99, pp. 1–1, 2016.
- [32] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision (IJCV)*, vol. 104, no. 2, pp. 154–171, 2013.
- [33] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [34] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European Conference on Computer Vision (ECCV)*, 2014.
- [35] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free? - weakly-supervised learning with convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [36] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [37] B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [38] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, vol. 26, no. 5, pp. 1019–1034, May 2015.
- [39] J. Donahue, J. Hoffman, E. Rodner, K. Saenko, and T. Darrell, "Semi-supervised domain adaptation with instance constraints," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [40] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [41] M. Roich and Y. Wang, "Weakly supervised localization of novel objects using appearance transfer," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [42] X. Shu, G.-J. Qi, J. Tang, and J. Wang, "Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation," in *ACM International Conference on Multimedia (MM)*, 2015.
- [43] Y. Zhu, Y. Chen, Z. Lu, S. J. Pan, G.-R. Xue, Y. Yu, and Q. Yang, "Heterogeneous transfer learning for image classification," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2011.
- [44] Y. Lu, L. Chen, A. Saidi, E. Dellandrea, and Y. Wang, "Discriminative transfer learning using similarities and dissimilarities," *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, vol. PP, no. 99, pp. 1–14, 2017.
- [45] K. K. Singh, F. Xiao, and Y. J. Lee, "Track and transfer: Watching videos to simulate strong human supervision for weakly-supervised object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [46] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *Neural Information Processing Systems (NIPS)*, 2013.
- [47] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Neural Information Processing Systems (NIPS)*, 2007.
- [48] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.
- [49] C. Leacock and M. Chodorow, "Combining local context and WordNet similarity for word sense identification," in *WordNet: An Electronic Lexical Database*, C. Fellbaum, Ed. MIT Press, 1998, pp. 265–283.
- [50] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of the International Joint Conference for Artificial Intelligence (IJCAI-95)*, 1995.
- [51] D. Lin, "An information-theoretic definition of similarity," in *International Conference on Machine Learning (ICML)*, 1998.
- [52] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Neural Information Processing Systems (NIPS)*, 2013.
- [53] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.
- [54] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2013.
- [55] I. Misra, A. Shrivastava, and M. Hebert, "Watch and learn: Semi-supervised learning of object detectors from videos," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [56] C. Rosenberg, M. Hebert, and H. Schneiderman, "Semi-supervised self-training of object detection models," in *IEEE Workshops on Application of Computer Vision (WACV)*, 2005.
- [57] Y. Yang, G. Shu, and M. Shah, "Semi-supervised learning of feature hierarchies for object detection in a video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [58] P. Agrawal, R. Girshick, and J. Malik, "Analyzing the performance of multilayer neural networks for object recognition," in *European Conference on Computer Vision (ECCV)*, 2014.
- [59] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM International Conference on Multimedia (MM)*, 2014.
- [60] M. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision (ECCV)*, 2014.
- [61] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele, "What helps where – and why? semantic relatedness for knowledge transfer," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.
- [62] S. Rothe and H. Schütze, "Autoextend: Extending word embeddings to embeddings for synsets and lexemes," in *the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL)*, 2015.
- [63] B. Gao, E. Dellandrea, and L. Chen, "Music sparse decomposition onto a midi dictionary of musical words and its application to music mood classification," in *International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2012.
- [64] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.
- [65] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [66] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [67] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 38, no. 1, pp. 142–158, 2016.



Yuxing Tang is currently a postdoctoral research fellow at the National Institutes of Health, USA. He received the B.S. and M.S. degrees from the Department of Information and Telecommunication Engineering, Beijing Jiaotong University, Beijing, China, in 2009 and 2011, respectively, and the Ph.D. degree in computer science from the Department of Mathematics and Computer Science, École Centrale de Lyon, France in 2016.

His research interests include computer vision and machine learning, in particular deep learning techniques for visual category recognition, object detection, and their application in medical imaging.



Josiah Wang is currently a postdoctoral researcher at the University of Sheffield, England. His main research interest is at the intersection of Computer Vision and Natural Language Processing, with an emphasis on integrating both modalities for better text and image understanding. He received his Ph.D. from the University of Leeds in 2013, working with the late Dr. Mark Everingham (from the Vision group) and Prof. Katja Markert (from the Natural Language Processing group) to develop algorithms for visual

object recognition by learning from textual descriptions.



Xiaofang Wang is currently assistant lecturer and researcher in Ecole Centrale Lyon. She has received the B.S. and M.S. degrees in biomedical engineering from Central South University, Changsha, China, and the Ph.D. degree in computer science from École Centrale de Lyon, France in 2015.

Her current research interests include image/video processing, machine learning (transfer learning, deep learning), computer vision (semantic image segmentation, object localization

and recognition, etc.).



Boyang Gao is currently a postdoctoral researcher at Istituto Italiano di Tecnologia (IIT), Italy. His research interests are deep learning, 3D reconstruction, object recognition, and speech recognition. He received the B.S. degree in computer science from Xi'an Jiaotong University in 2006 and M.S. degree in computer science from Harbin Institute of Technology in 2009, and the Ph.D. degree in computer science from École Centrale de Lyon, France in 2014.



Emmanuel Dellandréa received his Master and Engineering degrees in computer science from the University of Tours, Tour, France, in 2000, and the Ph.D. degree in computer science from the University of Tour in 2003.

He then joined École Centrale de Lyon, Écully, France, in 2004 as an Associate Professor. His research interests include multimedia analysis, image and audio understanding, and affective computing, including recognition of affect from image, audio and video signals.



Robert Gaizauskas is Professor of Computer Science and Head of the Natural Language Processing (NLP) research group within the Department of Computer Science, University of Sheffield, where he has worked since 1993. He studied Mathematics and Physics at the University of Toronto from 1972-74, then moved to Carleton University in Ottawa where he received an Honours BA in Philosophy in 1975 and an MA in Philosophy (with distinction) in 1978. He obtained a DPhil in Computational Logic from

the School of Cognitive and Computing Sciences, University of Sussex, in 1992. His research interests lie in developing computational techniques to access and manipulate information as expressed in natural language, particularly in written form. These techniques may then be utilized in applications such as document retrieval, text indexing, information extraction, text summarization, question answering, dialogue processing, machine translation, natural language generation, and so forth. He has published over 170 papers in peer-reviewed journals and conference proceedings and has served on the programme committees of numerous leading international conferences and workshops in the area of computational linguistics and as a reviewer for all the major journals in this area.



Liming Chen received the joint B.Sc. degree in mathematics and computer science from the University of Nantes, Nantes, France in 1984, and the M.Sc. and Ph.D. degrees in computer science from the University of Paris 6, Paris, France, in 1986 and 1989, respectively.

He first served as an Associate Professor with the Université de Technologie de Compiègne, before joining École Centrale de Lyon, Écully, France, as a Professor in 1998, where he leads an advanced research team on multimedia computing and pattern recognition. From 2001 to 2003, he also served as Chief Scientific Officer in a Paris-based company, Avivias, specializing in media asset management. In 2005, he served as Scientific Multimedia Expert for France Telecom R&D China, Beijing, China. He was the Head of the Department of Mathematics and Computer Science, École Centrale de Lyon, from 2007 through 2016. His current research interests include computer vision, machine learning, image and video analysis and categorization, face analysis and recognition, and affective computing. He has over 250 publications and successfully supervised over 35 PhD students. He has been a grant holder for a number of research grants from EU FP program, French research funding bodies and local government departments. He has so far guest-edited 3 journal special issues. He is an associate editor for Eurasip Journal on Image and Video Processing and a senior IEEE member.