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

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

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

{yuxing.tang, emmanuel.dellandrea, liming.chen}@ec-lyon.fr
{j.k.wang, r.gaizauskas}@sheffield.ac.uk boyang.gao@iit.it

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.

1. Introduction

Object detection/localization in images is one of the most widely studied problems in computer vision. 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) [16] for object detection, such as DetectorNet [34], OverFeat [30], R-CNN [12], SPP-net [13], Fast R-CNN [11] and Faster R-CNN [24], is heavily dependent on a large amount of training data manually labeled with object localizations (*e.g.* PASCAL VOC [8], ILSVRC (subset of ImageNet) [29], and Microsoft COCO [17]).

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.* [14] addresses this issue, by learning a transformation between 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. 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.

Our main contribution in this paper is therefore to incor-

* This work was supported by the EU CHIST-ERA D2K Visual Sense (Visen) project (ANR-12-CHRI-0002-04 and EPSRC EP/K019082/1).

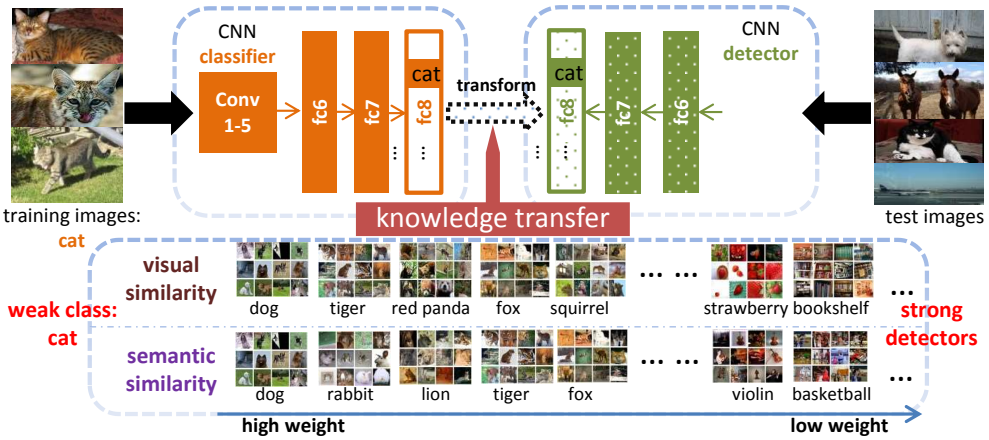

Figure 1. An illustration of our similarity-based knowledge transfer model. The question we investigate is whether knowledge about object similarities – visual and semantic – can be exploited to improve detectors trained in a semi-supervised manner. More specifically, to adapt the image-level classifier (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).

porate 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 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 [12, 14]; (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) [6]. 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* (Figure 1). 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 [29] 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.

2. Related Work

With the remarkable success of deep CNN on large-scale object recognition [16] in recent years, a substantial number of CNN-based object detection frameworks have emerged [11, 12, 13, 24, 30, 34]. However, these object detectors are trained in a fully supervised manner, where bounding box annotations are necessary during training.

Recently, there have been several studies in CNN-based object detection in a *weakly*-supervised setting [3, 4, 33, 36], *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 [35], BING [5], and EdgeBoxes [40]. Oquab *et al.* [22] develop a 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. Hoffman *et al.* [14] 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.

Another line of related work is to exploit knowledge transfer from various domains. Transfer learning (TL) [31] aims to transfer knowledge across different domains or tasks. Two general categories of TL have been proposed in previous work: *homogeneous* TL [7, 21, 14] in a single domain but with different data distributions in training and testing sets, and *heterogeneous* TL [25, 32, 39] 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. Rochan *et al*. [25] propose an appearance transfer method by transferring semantic knowledge (heterogeneous TL) from familiar objects to help localize novel objects in images and videos. Our work integrates knowledge transfer via both visual similarity (homogeneous TL) and semantic relatedness (heterogeneous TL) to help convert classifiers into detectors. Recently, Shu *et al*. [32] 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 [2]. DTN takes auxiliary text annotations (user tags and comments) and image pairs as input, while our semantic transfer method only needs image-level labels.

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 [20, 27, 37], 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 [14], 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). Finally, we combine both models to obtain our mixture transfer model, as presented in Section 4.4.

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* [16] CNN is pre-trained on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 classification dataset [29], 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 [12] framework. As in R-CNN, a background class ($fc8_{\mathcal{BG}}$) is added to the output layer and fine-tuned using bounding boxes from a region proposal algorithm, *e.g*. Selective Search [35]. 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.

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

where $\Delta_{\mathcal{B}_i^j}$ is the $fc8$ weight variation of the i^{th} nearest neighbor category in set $\mathcal{B}$ for category $j \in \mathcal{A}$. $\overrightarrow{w^c}$ and $\overrightarrow{w^d}$ 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*. [14] 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 $fc8_{\mathcal{A}}$ 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 $fc7$ layer 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 [1, 15, 38]: 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:

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

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 $CNN(I_n)_i$ 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 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 $fc7$ were to be used (two methods produce similar range of results).

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:

$$\forall j \in \mathcal{A} : \overrightarrow{w^d_{j\,v}} = \overrightarrow{w^c_j} + \sum_{i=1}^{m} s_v(j,i)\Delta_{\mathcal{B}^j_i} \tag{3}$$

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 [6, 25, 26], we observe that visual similarity is correlated with semantic relatedness. According to [6], this relationship is particularly strong when measurements are focused on the category instances themselves, 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: Semantic similarity is a well-explored area within the Natural Language Processing community. Recent advances in word embeddings trained on large-scale text corpora [18, 23] have helped progress research in this area, 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 [19]. Thus, we encode each of the K categories as a word vector, more specifically a 300-dimensional word2vec embedding [18]. As each category is a WordNet [9] 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 [28] to encode the categories as *synset embeddings* in the original word2vec embedding space. AutoExtend 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 only report the results for 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 can 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 [25]. 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 $V = [v_1; v_2; \ldots; v_i; \ldots; v_m]$, and $\Gamma_j = [\gamma_j^1, \gamma_j^2, \ldots, \gamma_j^i, \ldots, \gamma_j^m]$ is a set of coefficients of the linear combination. We are motivated to find the solution $\Gamma_j^\star$ 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 $\Gamma_j^\star$ can be formulated as the following optimization:

$$\Gamma_j^\star = \underset{\Gamma_j > 0}{\arg\min}(\|v_j - \Gamma_j V\|_2 + \lambda\|\Gamma_j\|_0) \tag{4}$$

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 [10]. Therefore, the final transformation via semantic transferring is formulated as:

$$\forall j \in \mathcal{A} : \overrightarrow{w^d_{j\,s}} = \overrightarrow{w^c_j} + \sum_{i=1}^{m} s_s(j,i)\Delta_{\mathcal{B}^j_i} \tag{5}$$

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:

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

where $intersect[\cdot]$ 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.

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 [14], 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 [12]. 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 [12, 14]). Finally, we evaluate our knowledge transfer framework on the val2 dataset (9,917 images in total).

5.2. Implementation details

In all the experiments, we consider LSDA [14] as our baseline model and follow their main settings. For the semantic representation, we use word2vec CBoW embeddings pre-trained on part of the Google News dataset comprising about 100 billion words [18]. We train AutoExtend [28] 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 nonzero components to 20, since a target category has strong correlation with a small number of source categories.

5.3. Quantitative evaluation on the "weakly labeled" categories

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 steadily 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 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 sim-

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. - $\leq$20	28.18	19.04	23.66
Ours (mixture transfer model)	-	28.04	**20.03** ↑**3.88**	24.04
Oracle: Full Detection Network	-	29.72	26.25	28.00

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.

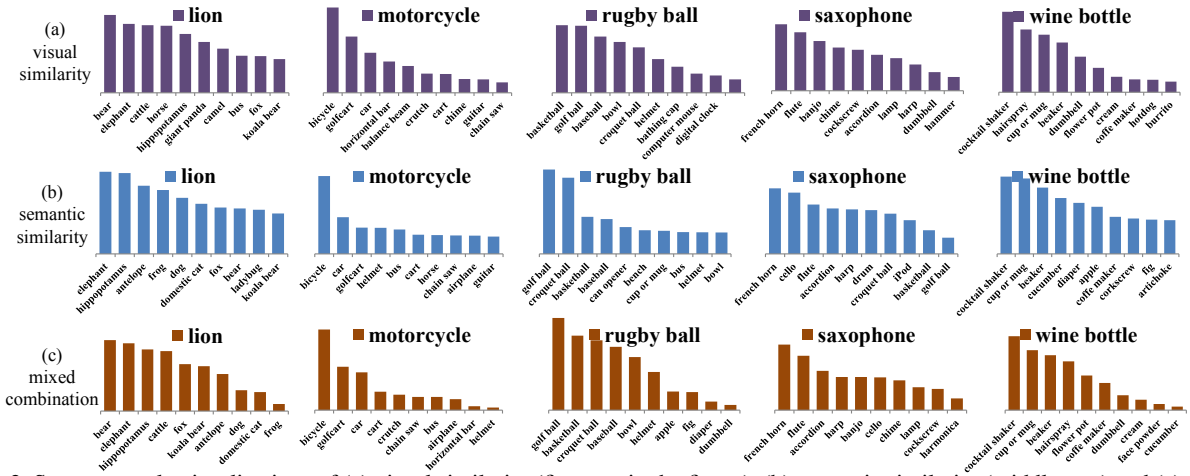

Figure 2. Some example visualizations of (a) visual similarity (first row in the figure), (b) semantic similarity (middle row) and (c) mixture similarity (last row) between a target "weakly labeled" category and its source categories from which to transfer knowledge. For each target category, the top-10 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).

ilarity. 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 Figure 2. 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. The results of using synset embeddings (18.32%, using weighted-100, the same below) are superior to using 'naive' word2vec embeddings

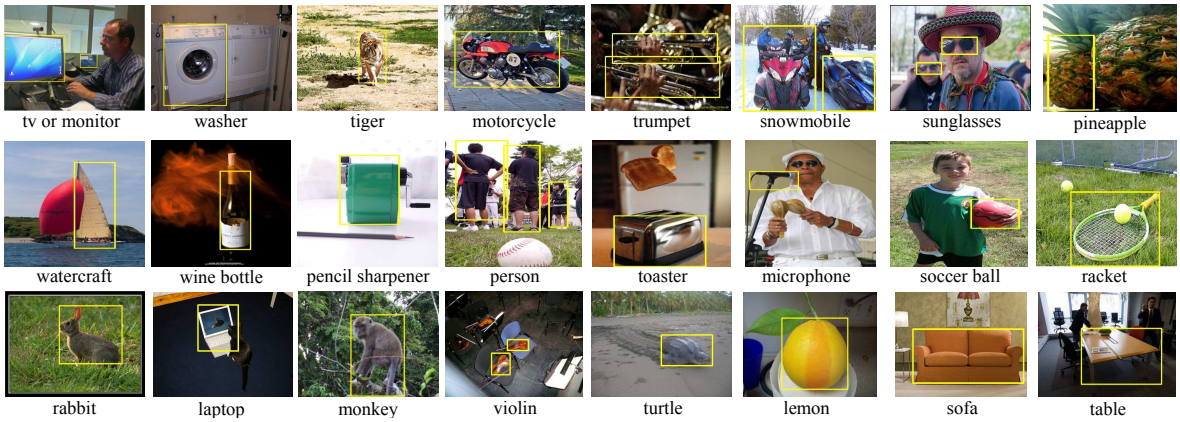

Figure 3. Examples of correct detections (true positives) of our mixture knowledge transfer model on ILSVRC2013 images. For each image, only detections for the "weakly labeled" target category (text below image) are listed.

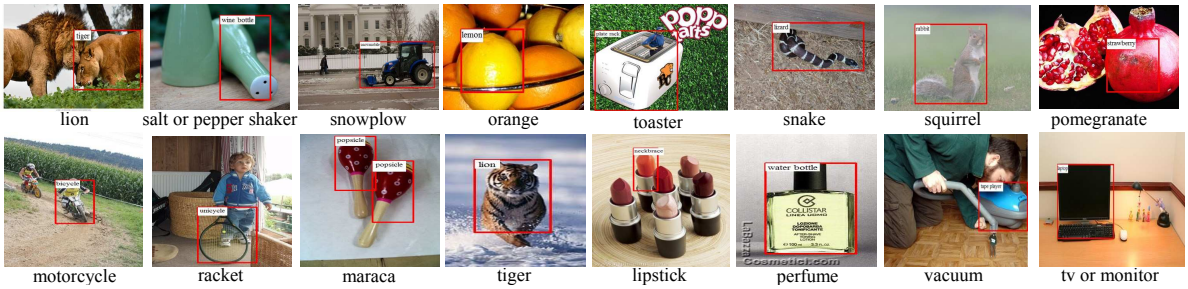

Figure 4. Examples of incorrect detections (confusion with other objects) of our mixture knowledge transfer model on ILSVRC2013 images. The detected object label is shown in the top-left of its bounding box.

(17.83%) and WordNet based measures such as path-based similarity (17.08%) and Lin similarity [**?**] (17.31%). Several examples visualizing the related categories of the 10 largest semantic reconstruction coefficients are shown in the middle row of Figure 2. We observe that semantic relatedness indeed correlates with visual similarity.

The state-of-the-art result 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%±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 Figure 2. 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 (Figure 5). This suggests that the transferring of visual similarities is slightly more important than semantic relatedness, although both are indeed complementary. Figures 3 and 4 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.

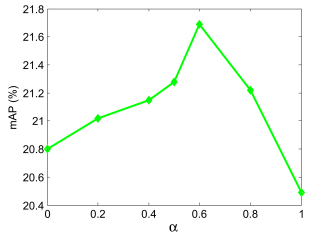

Figure 5. Sensitivity of parameter α vs. mAP.

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 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, 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.

References

[1] 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. 4

[2] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, U. D. Montral, and M. Qubec. Greedy layer-wise training of deep networks. In *Neural Information Processing Systems (NIPS)*, 2007. 3

[3] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with posterior regularization. In *British Machine Vision Conference (BMVC)*, 2014. 2

[4] 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. 2

[5] 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. 2

[6] T. Deselaers and V. Ferrari. Visual and semantic similarity in imagenet. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 2, 5

[7] 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. 3

[8] 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)*, 88(2):303–338, 2010. 1

[9] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998. 5

[10] 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. 5

[11] R. Girshick. Fast R-CNN: towards real-time object detection with region proposal networks. In *International Conference on Computer Vision (ICCV)*, 2015. 1, 2

[12] 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. 1, 2, 3, 6

[13] 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)*, 2015. 1, 2

[14] 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. 1, 2, 3, 4, 6

[15] 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. 4

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems (NIPS)*. 2012. 1, 2, 3

[17] 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*, abs/1405.0312, 2014. 1

[18] 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. 5, 6

[19] 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. 5

[20] 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. 3

[21] 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. 3

[22] 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. 2

[23] 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. 5

[24] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN. In *Neural Information Processing Systems (NIPS)*, 2015. 1, 2

[25] M. Rochan and Y. Wang. Weakly supervised localization of novel objects using appearance transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3, 5

[26] 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. 5

[27] 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. 3

[28] 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. 5, 6

[29] 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)*, 0(0):1–42, April 2015. 1, 2, 3

[30] 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. 1, 2

[31] L. Shao, F. Zhu, and X. Li. Transfer learning for visual categorization: A survey. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 26(5):1019–1034, May 2015. 3

[32] X. Shu, G.-J. Qi, J. Tang, and J. Wang. Weekly-shared deep transfer networks for heterogeneous-domain knowledge propagation. In *ACM International Conference on Multimedia (MM)*, 2015. 3

[33] 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. 2

[34] C. Szegedy, A. Toshev, and D. Erhan. Deep neural networks for object detection. In *Neural Information Processing Systems (NIPS)*. 2013. 1, 2

[35] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *International Journal of Computer Vision (IJCV)*, 104(2):154–171, 2013. 2, 3

[36] 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)*, 24(4):1371–1385, April 2015. 2

[37] 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. 3

[38] M. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, 2014. 4

[39] 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. 3

[40] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision (ECCV)*, 2014. 2