Weakly Supervised Localization of Novel Objects Using Appearance Transfer

Mrigank Rochan
Department of Computer Science
University of Manitoba, Canada
mrochan@cs.umanitoba.ca

Yang Wang
Department of Computer Science
University of Manitoba, Canada
ywang@cs.umanitoba.ca

Abstract

We consider the problem of localizing unseen objects in weakly labeled image collections. Given a set of images annotated at the image level, our goal is to localize the object in each image. The novelty of our proposed work is that, in addition to building object appearance model from the weakly labeled data, we also make use of existing detectors of some other object classes (which we call "familiar objects"). We propose a method for transferring the appearance models of the familiar objects to the unseen object. Our experimental results on both image and video datasets demonstrate the effectiveness of our approach.

1. Introduction

How would one detect an object class, say "dog", in images? The de facto answer in computer vision is to collect a set of labeled training data (e.g. images with object bounding box annotations) for this object class and apply standard supervised machine learning to learn the appearance model for this object category. Then this appearance model can be used to detect dogs in any image. The key of this standard pipeline is that we need to have access to a large amount of manually labeled training data. In the past few years, the availability of large-scale annotated datasets (e.g. PASCAL VOC [6] and ImageNet [25]) has been one of the driving forces of much progress in visual recognition.

The most straightforward approach to create large datasets is to gather images/videos online and ask people to annotate them via crowd-sourcing. However, this is very expensive and time-consuming. The PASCAL dataset [6] has focused only on 20 common objects. ImageNet [25] covers more object classes, but is still limited to the objects defined in the WordNet hierarchy, and most of the images in ImageNet are not annotated with object bounding boxes. It is not clear how this straightforward approach would scale up when we need to deal with a large number of concepts emerging and changing over time, which is common for images/videos on the Internet.

Although it is difficult to collect training images annotated with object bounding boxes, it is usually much easier to collect weakly labeled data, where labels are only given at the image level. For example, many online data (Flickr images, YouTube videos) might come with user-generated tags describing the objects present in the images/videos. It is also possible to collect weakly labeled images of an object class via image search. In this paper, our goal is to develop techniques to localize the object in weakly labeled data (see Fig. 1). Given a collection of images labeled with an object category (e.g. "car"), our method will output the bounding box of this object in each image. Our method can also be applied in videos. In this case, we are given one single video of the novel object. Our method will treat the frames of the video as the image collection and localize the object in each frame.

Another weakness of traditional approaches in visual recognition is that even if we have appearance models for 1000 object classes, we have to start from scratch when building the appearance model for the 1001-th object class. This is somewhat unintuitive and unsatisfying – it should be easier to build the appearance model for a new object class if it is related to other known object categories.

Our work is motivated by the following observations. 1) Large datasets with bounding box annotations exist for some object categories, e.g. the 20 objects in PASCAL [6] and a subset of objects in ImageNet [25]. For these object categories (we will call them "familiar objects"), we have access to detectors with reasonably good performance. 2) For most of other object categories (we call them "novel objects"), fully annotated data are scarce. But it is easy to collect weakly labeled images/videos for them. In this paper, we use the term "novel objects" to denote objects for which we do not have fully annotated data. It is different from the "novel object" used in object discovery (e.g. [18]). 3) Recent work in text analysis has produced valuable resources on word semantics. For example, a word is represented as a fixed length vector (called "word embedding") in [19]. The embedding vectors of words are learned from large collections of text documents. Semantically related words (e.g.

a collection of images labeled as “car” a video labeled as “car”

Figure 1. Our goal is to localize objects in weakly labeled data. (Left) Given a collection of images labeled as “car”, our algorithm will localize the car in each image of the collection. (Right) By applying our algorithm on a single video labeled as “car”, we can localize the specific instance of “car” in this video. The red bounding boxes in this figure are outputs of our algorithm.

“cat” and “dog”) are being mapped closer in this embedding space. The word embedding provides a way for us to infer how two object classes are related. 4) Objects that are semantically close often have similar visual appearances. We acknowledge that some people might not agree with the last point – indeed one can find object categories that are semantically close, but visually very different. But previous work (e.g. [3, 7, 17]) in computer vision has demonstrated that semantic knowledge can still be useful for solving vision-related tasks, even when it is constructed from non-visual information. In this paper, we will show that it is possible to transfer appearance model from one object class to another based on their semantic relationship in term of the word vectors.

The main contribution of this paper is to incorporate knowledge transfer into weakly supervised learning (WSL) of object classes. Although weakly supervised learning has been previously explored in computer vision, there has not been much work on exploiting familiar objects to help learning new object categories.

2. Related Work

In recent years, weakly supervised learning has emerged as a powerful way of reducing the effort required in collecting fully-annotated datasets. Several methods have been proposed to localize or segment objects in weakly labeled images and videos, where the labels are only provided at the image/video level. Nguyen et al. [20] use a variant of multiple instance learning (MIL) to localize objects in images without bounding box annotations. Deselaers et al. [4] propose a CRF model for learning object appearance while localizing the objects in weakly labeled images. Similar models [22] have been used for learning object detectors from videos. Tang et al. [31] propose a method for object co-localization in noisy Internet images. The goal is to simultaneously localize the object of interest in a collection of images. Joulin et al. [15] extend the co-localization framework to videos. Weakly supervised learning has also been used to segment objects in images and videos. Vezhnevets et al. [35] propose a CRF model for semantic segmentation of object with weak supervision. There are also methods [11, 32] for segmenting objects in YouTube videos.

Our work of using familiar objects to help learning novel objects is related to transfer learning. The goal is to use the knowledge learned in one task to help the learning of related tasks. Earlier work [8] on one-shot learning in vision aims to transfer the knowledge from some object categories to a new object class, so learning a new object class only requires a small number of training images. Stark et al. [30] transfer shape knowledge between related object classes. Tommasi and Caputo [33] use the SVM parameters learned from one object class as the prior for a new related object class. Farhadi et al. [7] and Lampert et al. [17] propose to transfer knowledge about object categories via attributes. There is also work [23, 24] on discovering attributes from online knowledge base. Another type of transfer learning is to adapt the model learned for one vision task to another, e.g. Hoffman [12] propose a method for adapting classification models to detection.

Our work is also related to a line of research on using linguistic knowledge (in particular, word embedding) in computer vision. In the natural language processing (NLP) community, there has been work [13, 19] on learning word embedding from a text corpus. The goal is to produce a vector representation for each word. If two words (e.g. “dog” and “cat”) are semantically close, their word vectors will tend to be similar. The learned word vectors have been used in many NLP applications, e.g. information retrieval, document classification, etc. Recently, they have been used in computer vision applications as well. Frome et al. [9] learn to embed both words and images jointly in a semantic space for image classification. Andrej et al. [16] use word vectors for learning to translate between images and sentences.

3. Problem Statement

We assume that we have a set of “familiar object” categories and some “novel object” categories. For familiar objects, we have access to pre-trained appearance models (i.e. detectors) for them, or equivalently, a set of training images with bounding box annotations for learning the appearance

models. For example, familiar objects might correspond to the 20 object categories in the PASCAL VOC challenge or the subset of synsets in ImageNet with annotated bounding boxes. For these familiar objects, a large collection of annotated images are available. There are also many pre-trained object detectors available for these familiar objects. For a novel object class, we have a set of images containing the novel object, but we do not have the bounding boxes of the object in those images. Our goal is to localize the novel object in each image. In this work, we will make the following assumptions about the image collection: 1) the novel object appears in every image of the collection; 2) For each object name (either familiar or novel), we have a word vector describing the semantic information of this object. These word vectors can be obtained from the NLP community [13, 19].

4. Our Approach

An overview of our approach is illustrated in Fig. 2. To localize a novel object in a collection of weakly labeled images, we build two initial appearance models. The first appearance model is obtained from the image collection using object proposals (Sec. 4.1). The second appearance model is obtained by transferring knowledge from other familiar objects (Sec. 4.2). Our final appearance model of the novel object is a combination of these two initial models. We then use the final appearance model to localize the novel object in each image of the collection.

4.1. Appearance model from object proposals

Given a collection of weakly labeled images of a novel object, the first step of our approach is to generate a set of object proposals in each image. Each object proposal is a bounding box which might contain *any* object class. There has been several recent work on generating object proposals in the form of bounding boxes [1, 36] or image segments [5]. In our work, we use the edge boxes method in [36] for generating bounding boxes as our object proposals. This method is based on the observation that the number of contours in a bounding box is indicative of how likely this bounding box contains an object of any class. Based on this observation, it defines a box objectness score that measures the number of edges in the box minus those that are members of contours that overlap the box's boundary. Using efficient data structures, the edge boxes algorithm can evaluate millions of candidate boxes quickly and return the top scoring boxes in a given image. The score (called "objectness score") associated with each bounding box indicates the likelihood of the box containing an object. Figure 3 shows the results of applying the edge boxes algorithm on sample images.

We assume that the novel object is reasonably salient in most images in the collection. Admittedly, this assumption does not always hold. But we believe this is a reasonable assumption in many cases. For example, if we collect images by querying the name of the novel object from search engines, the novel object tends to be salient in the images returned by search engines.

Based on this assumption, we can train an initial model for the novel object from the object proposals in the image collection. We select object proposals with high objectness scores and consider them as positive examples of the novel object. We then select a set of negative examples by randomly generating bounding boxes from images that do not correspond to the novel object. Given these positive and negative examples, we learn an appearance model for this novel object using a linear SVM. Let $\mathbf{x}$ denote the feature vector of an image patch, the appearance model is represented by a parameter vector $\mathbf{w}_p$. The dot product $\mathbf{w}_p^\top \mathbf{x}$ (without loss of generality, we assume a linear SVM model without the bias term) indicates the likelihood of $\mathbf{x}$ being the novel object.

4.2. Appearance model from familiar objects

The appearance model and the localization of the novel object appear to be a chicken-and-egg problem. If we have an appearance model of the novel object, we can use the appearance model to localize the object in an image. Conversely, if we know the ground-truth locations of the novel object in some images, we can simply learn an appearance model of this object. Then we can use the appearance model to localize the object in other images. Sec. 4.1 provides one way of getting the appearance model $\mathbf{w}_p$. In this section, we propose another way of constructing the appearance model by transferring knowledge from other familiar objects. First, we use the word vectors associated with the novel object and familiar objects to establish their semantic relatedness. Then we transfer the appearance models of familiar objects based on their relatedness to the novel object.

Word vectors: We use the word vectors learned in [13]. These word vectors are learned in an unsupervised fashion from a large corpus using a neural-network-based language model. The model learns the semantics of words from their local and global context in the corpus. As a result, the model produces a vector space representation for each English word as a D-dimensional vector ($D = 200$ in our experiments). These vectors can then be used as features in various applications in text analysis, e.g. information retrieval, document classification, parsing, etc. In our work, we use the word vectors as a source of semantic knowledge to bridge the familiar and novel objects.

Figure 4 shows a visualization of the word vectors by projecting them on a 2D space using t-SNE [34]. We can see that words similar in their semantic meanings are close in term of their word vectors. For example, words corresponding to various music instruments are mapped together

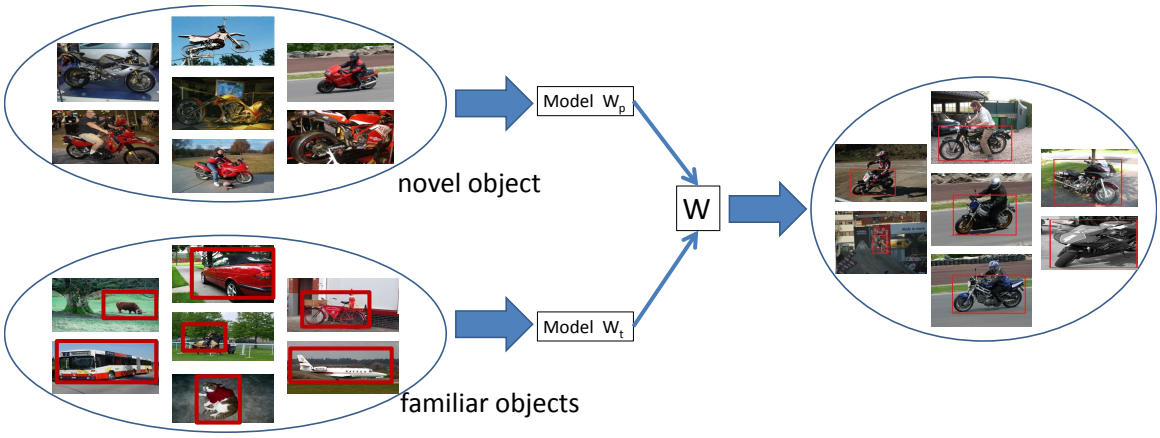

Figure 2. An overview of our approach. (Top left) Given a collection of weakly labeled images of a novel object (e.g. motorbike), we learn an appearance model $\mathbf{w}_p$ from the object proposals (see Sec. 4.1). (Bottom left) We also have access to fully annotated data (or pre-trained models) for a set of familiar objects, e.g. car, bus, dog, etc. We transfer the knowledge of familiar objects to obtain another appearance model $\mathbf{w}_t$ for the novel object (see Sec. 4.2). (Middle) The final appearance model $\mathbf{w}$ for the novel object is a combination of $\mathbf{w}_p$ and $\mathbf{w}_t$. (Right) We can then use $\mathbf{w}$ to localize the novel object in the image collection (see Sec. 4.3).

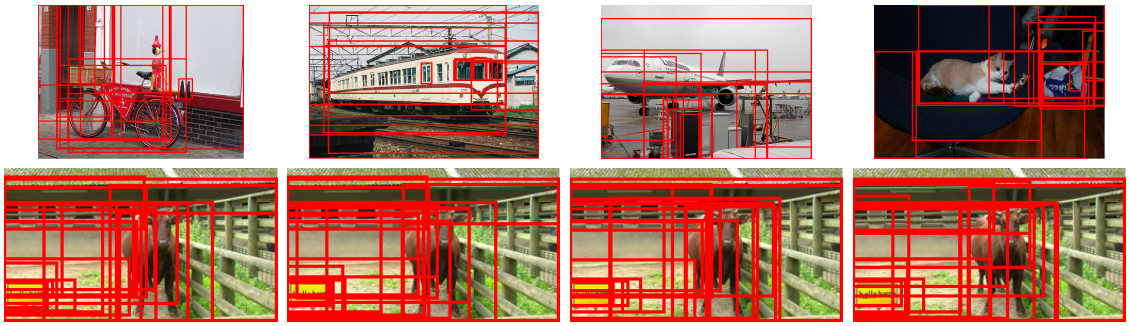

Figure 3. Examples of applying the edge boxes algorithm on images (1st row) and videos (2nd row). Objects in the images tend to be covered by one or more bounding boxes produced by the edge boxes algorithm.

in the upper left corner in Fig. 4.

Novel object as sparse reconstruction: We are given a set of K familiar object classes. We use $\mathbf{v}_i$ to denote the word vector associated with the i-the object class and $\mathbf{u}_i$ to denote the corresponding appearance model. For simplicity, we assume the appearance model (object detector) has a linear form:

$$f_i(\mathbf{x}) = \mathbf{u}_i^\top \mathbf{x} \tag{1}$$

where $\mathbf{x}$ represents the feature vector of an image patch. Given an input image, we can apply Eq. 1 to sub-windows at various positions and scales to detect the i-th object in the image.

For a novel object class, we denote its word vector as $\mathbf{v}$. Our goal is to obtain an appearance model (we denote it as $\mathbf{w}_t$) for this novel object class. Our approach is based on two assumptions. First of all, the word vectors and appearance models of objects are related – if two objects i and j are similar in terms of their word vectors $\mathbf{v}_i$ and $\mathbf{v}_j$, they tend to be similar in terms of their appearance models $\mathbf{u}_i$ and $\mathbf{u}_j$. Secondly, for a novel object, we can approximate its word vector $\mathbf{v}$ as a linear combination of those of familiar objects, i.e.:

$$\mathbf{v} \approx \theta_1 \mathbf{v}_1 + \theta_2 \mathbf{v}_2 + ...\theta_K \mathbf{v}_K \tag{2}$$

where the parameters θ_i $(i = 1, 2, ..., K)$ are the coefficients of the linear combination.

We estimate the coefficient vector $\Theta = [\theta_1, \theta_2, ..., \theta_K]^\top$ by solving the following optimization problem:

$$\min_{\Theta > 0} ||\mathbf{v} - (\theta_1 \mathbf{v}_1 + \theta_2 \mathbf{v}_2 + ...\theta_K \mathbf{v}_K)||_2^2 + \lambda ||\Theta||_1 \tag{3}$$

The first term in Eq. 3 minimizes the reconstruction error of the linear approximation, while the second term minimizes the L_1 norm of the parameter Θ. The L_1 norm will encourage Θ to be sparse, since we prefer to reconstruct the novel object using a small number of familiar objects.

Transferring appearance model: By solving Eq. 3, we get the parameter vector $\Theta = [\theta_1, \theta_2, ..., \theta_K]^\top$. If we assume that the semantic relatedness of object classes (in term

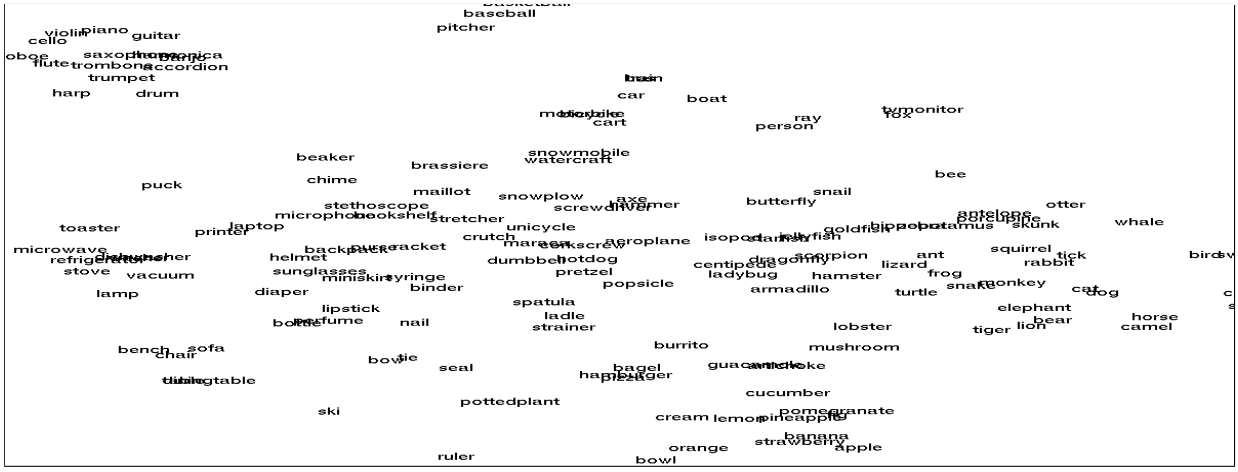

Figure 4. (Best viewed in PDF with magnification) Visualization of the word vectors in 2D using t-SNE [34]. The t-SNE algorithm finds a 2D embedding of the word vectors.

of word vectors) is similar to that of appearance models, we can use the same Θ to represent the appearance model of the novel object as:

$$\mathbf{w}_t = \theta_1 \mathbf{u}_1 + \theta_2 \mathbf{u}_2 + ...\theta_K \mathbf{u}_K \tag{4}$$

Note that we do not require any training data of the novel object in order to get $\mathbf{w}_t$. As long as we have the word vectors of object classes (both familiar and novel) and pre-trained appearance models for familiar objects, we can use Eq. 3 and Eq. 4 to compute $\mathbf{w}_t$. In other words, we have transferred the appearance models from familiar objects to the novel object.

4.3. Modeling and localizing the novel object

Sec. 4.1 and Sec. 4.2 provide two different ways of learning the appearance model of the novel object. Let $\mathbf{w}_p$ and $\mathbf{w}_t$ denote the two appearance models learned in Sec. 4.1 and Sec. 4.2, respectively. Our final appearance model $\mathbf{w}$ for the novel object is a linear combination of these two:

$$\mathbf{w} = \gamma \mathbf{w}_p + \mathbf{w}_t \tag{5}$$

where γ is a parameter that controls the relative importance of $\mathbf{w}_p$ and $\mathbf{w}_t$.

Intuitively, the parameter γ should vary depending on the "transferability" of the novel object. If a lot of familiar objects are closely related to the novel object, it should be easier to transfer the appearance model to the novel object. In this case, we like γ to be small, so $\mathbf{w}_t$ will have a higher influence. Conversely, if the novel object is vastly different from all the familiar objects, we like γ to be large. So we do not rely too much on transferring appearance model from the familiar objects.

One way to define the "transferability" of an novel object is to examine the reconstruction error in Eq. 3. Let $\Theta^* = [\theta_1^*, \theta_2^*, ..., \theta_K^*]^\top$ be the solution to Eq. 3, the reconstruction error is:

$$E(\Theta^*) = ||\mathbf{v} - (\theta_1^* \mathbf{v}_1 + \theta_2^* \mathbf{v}_2 + ...\theta_K^* \mathbf{v}_K)||_2^2 \tag{6}$$

We then set $\gamma = \beta E(\Theta^*)$, where β is a free parameter. I.e. our final appearance model is computed as:

$$\mathbf{w} = \beta \cdot E(\Theta^*) \cdot \mathbf{w}_p + \mathbf{w}_t \tag{7}$$

Notice that if a novel object can be easily represented as a linear combination of familiar objects, i.e. it is easy to do the transfer learning, the reconstruction error $E(\Theta^*)$ will be small. In this case, the appearance model $\mathbf{w}_t$ obtained from the transferring learning will have a larger effect in Eq. 7.

We can then use this appearance model $\mathbf{w}$ to re-score the object proposals generated in Sec. 4.1. Let $\mathbf{x}$ be the feature vector extracted from the image patch of a proposal, we use $\mathbf{w}^\top \mathbf{x}$ to measure the score of this proposal belonging to the novel object. The top scored bounding box in each image will be our localization result.

An interesting special case is when the image collection consists of frames from a single video. This is potentially useful for video retrieval. For example, if we query a novel object, say "tiger" in YouTube. Instead of just returning the videos containing tigers, we can also localize the tiger in each video. If we apply our method on a *single* video of a novel object, we will get an appearance model for the specific instance of the object in this particular video. In other words, our approach can automatically adapt to different videos of the same novel object.

5. Experiments

We evaluate our approach on one image dataset (Sec. 5.2) and two video datasets (Sec. 5.3 and 5.4).

method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	avg
transfer only	48.32	48.97	17.58	55.25	6.15	32.26	15.85	40.36	**28.54**	70.92	4.5	15.91	43.55	34.69	13.75	3.26	51.04	28.38	46.74	19.92	31.3
proposal only	77.31	55.55	62.73	40.88	**21.31**	77.96	72.1	54.9	14.83	68.79	29.5	56.29	70.38	74.69	43.18	27.35	47.91	26.2	70.88	67.19	53
combined	**78.57**	**63.37**	**66.36**	**56.35**	19.67	**82.26**	**74.75**	**69.13**	22.47	**72.34**	**31**	**62.95**	**74.91**	**78.37**	**48.61**	**29.39**	**64.58**	**36.24**	**75.86**	**69.53**	**58.84**

Table 1. CorLoc results on the PASCAL VOC 2007 dataset. We compare three different methods: (1st row) using only the appearance model transferred from familiar objects $\mathbf{w}_t$; (2nd row) using only the appearance model from the object proposals $\mathbf{w}_p$; (3rd row) using the combined appearance model $\mathbf{w}$.

Figure 5. Qualitative examples of our approach on the PASCAL VOC 2007 dataset.

method	aero	bird	boat	car	cat	cow	dog	horse	bike	train	avg
[15] (video)	25.12	31.18	27.78	38.46	41.18	28.38	33.91	35.62	23.08	25	30.97
[21]	**65.4**	**67.3**	38.9	65.2	**46.3**	40.2	**65.3**	48.4	39	25	50.1
transfer only	35.27	10.75	31.75	30.77	19.66	83.78	26.96	50.68	50.56	46.43	38.66
proposal only	51.69	54.84	32.54	**85.71**	14.53	75.68	55.65	53.42	51.69	39.29	51.5
combined	56.04	30.11	**39.68**	**85.71**	24.79	**87.83**	55.65	**60.27**	**61.8**	**51.79**	**55.37**

Table 3. CorLoc results on the YouTube-Objects dataset. Similar to the PASCAL VOC 2007 dataset, we compare three different methods: (3rd row) using only the appearance model transferred from familiar objects $\mathbf{w}_t$; (4th row) using only the appearance model from the object proposals $\mathbf{w}_p$; (5th row) using the combined appearance model $\mathbf{w}$. We also compare with previous work [15] (1st row) and [21] (2nd row) that uses the same dataset.

method	aero	bird	boat	car	cat	cow	dog	horse	bike	train	avg
transfer only	40.34	43.86	**40.41**	28.42	26.35	47.15	34.37	28.67	26.12	24.28	34
proposal only	42.23	51.24	29.54	**67.76**	14.75	50.2	**47.02**	22.18	16.44	18.84	36.02
combined	**45.74**	**55.47**	39.51	58.75	**26.51**	**55**	43.51	**33.71**	**32.76**	**25.63**	**41.66**

Table 4. CorLoc results of different methods on the YouTube-Objects-Subset dataset.

5.1. Implementation details

We use the 4096 dimensional CNN-feature implemented in Caffe [14] as our feature representation for an object proposal. This feature has been proved to be one of the state-of-the-art feature representations in many visual recognition tasks. In order to construct the set of familiar objects, we use the 200 object classes in [10]. These object models are trained from a subset of the ImageNet images with bounding box annotations using the Caffe-based CNN features. Some of the object classes do not have word vectors asso-

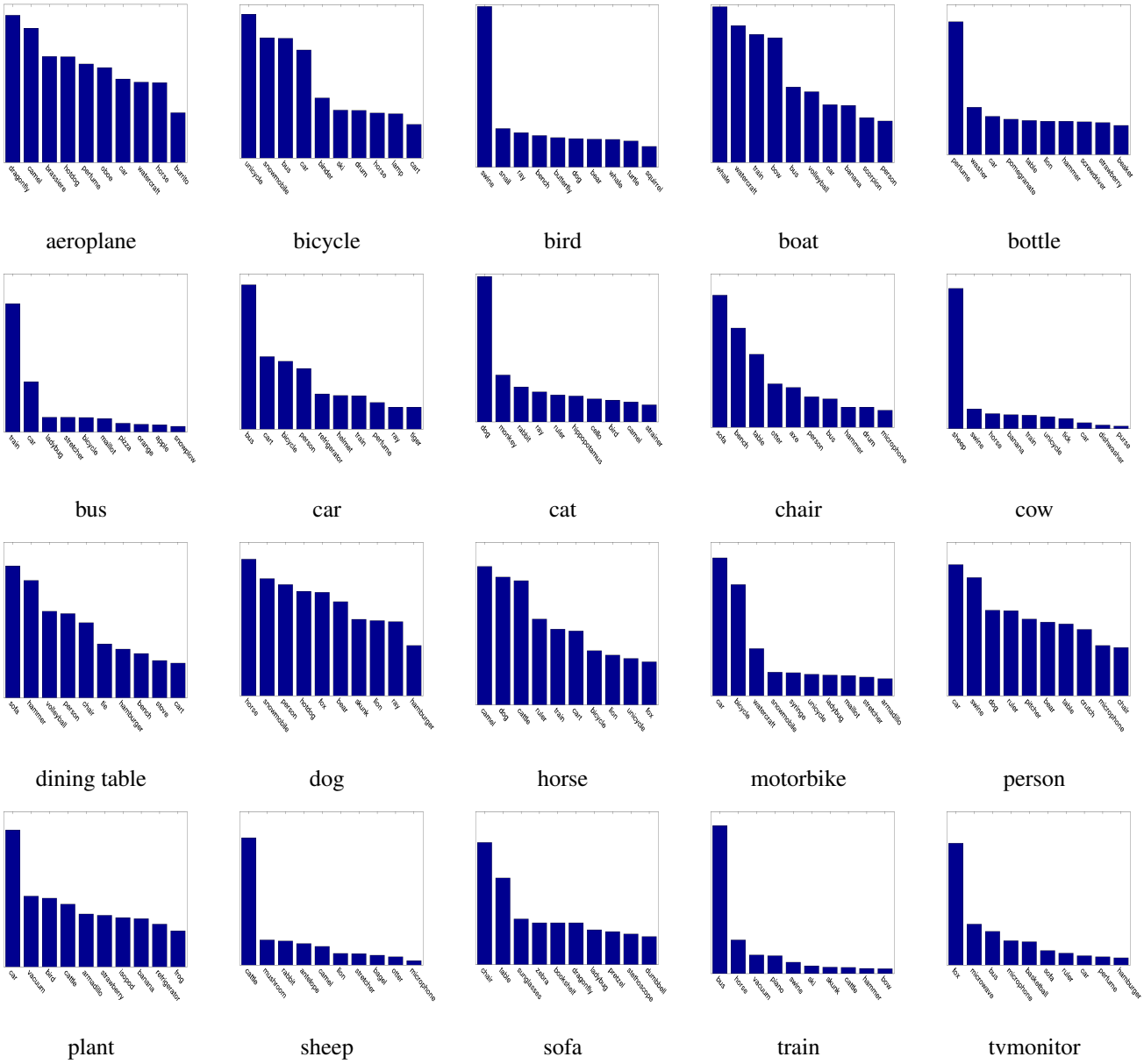

Figure 6. (Best viewed in PDF with magnification) Visualization of the Θ parameters for novel object classes. For each novel object class, we show the top 10 familiar objects with the corresponding θ values.

ciated with them, possibly because they does not appear in the corpus used for learning the word vectors. We filter out those object classes and select 142 familiar object classes in the end.

We set the free parameters of our method by validating over a small set of images/videos. For the images in the PASCAL VOC 2007 dataset, we extract 100 object proposals on each image and set $\lambda = 1$ and $\beta = 0.3$. For videos in the YouTube-Objects dataset, we extract 20 object proposals on each frame and set $\lambda = 1$ and $\beta = 0.1$.

5.2. PASCAL VOC 2007

The dataset contains images of 20 object classes from the train+val subsets of PASCAL VOC 2007 dataset. We consider each of them as the novel object and apply our algorithm on the images that contain at least one instance of this novel object. Since the object classes in PASCAL overlap with those of the 142 familiar objects, we remove the novel object class from the set of familiar objects when doing the appearance transfer. For example, when we consider "dog" as the novel object, we remove the "dog" model from

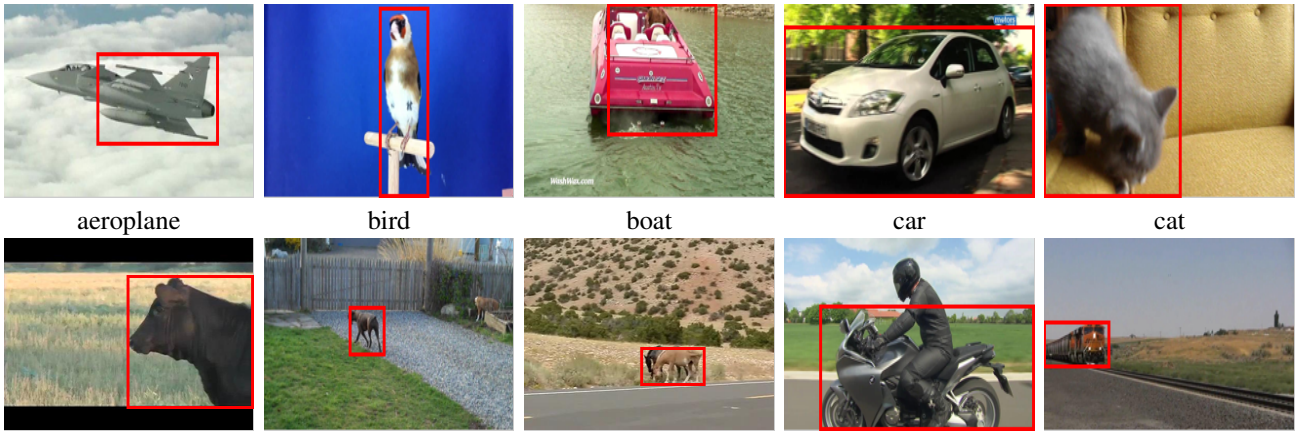

Figure 7. Qualitative examples of our approach on the YouTube-Objects-Subset dataset.

method	CorLoc
[15] (image model)	24.6
[27]	30.2
[29]	30.4
[28]	32.0
[26]	36.2
[2]	38.8
ours	58.84

Table 2. Comparison with previous work on the PASCAL VOC 2007 dataset in term of the average CorLoc.

the 142 familiar object classes.

We use the CorLoc defined in [4] to measure the performance. It is defined as the percentage of images in which a method correctly localizes the novel object according to the PASCAL criterion $\frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} > 0.5$, where B_p is the localized bounding box and B_{gt} is a ground-truth bounding box. Table 1 shows the CorLoc results of three methods: 1) using only the transferred appearance $\mathbf{w}_t$; 2) using only the appearance model from the object proposals $\mathbf{w}_p$; 3) using the combined appearance model $\mathbf{w}$. The results of using the combined appearance model achieve the best performance on 18 out of the 20 object classes.

Table 2 shows the comparison with other published results. Our approach significantly outperforms others. Some examples of our localization results are shown in Fig 5.

Figure 6 visualizes the Θ parameters obtained via Eq. 3. For each novel class, we show the top 10 familiar object classes according to the descending order of their corresponding θ values.

5.3. YouTube-Objects

The Youtube-Objects dataset [22] consists of videos of 10 object classes. For each class, bounding box annotations are provided for one frame per shot for 100-290 shots. We apply our method on each video in the dataset by considering the frames in this video as the image collection. Similarly, we remove the novel object class from the set of familiar objects when doing the appearance transfer. Only frames with annotations are considered in the evaluation. In Table 3, we compare our results with previous work that uses the same dataset.

5.4. YouTube-Objects-Subset

We also evaluate our method on the subset of the YouTube-Objects dataset collected in [32]. This dataset contains ground-truth segment-level object annotations on all frames in many video shots. The results on this dataset are shown in Table 4. Fig. 7 shows some qualitative results on this dataset.

6. Conclusion

We have proposed an approach for localizing novel objects from weakly labeled data. The novelty of our work is that in addition to learning appearance models from the weakly labeled data, we also exploit appearance models available from other familiar objects that are related to the novel object. Our experimental results demonstrate that our proposed method outperforms other baseline approaches. As future work, we plan to use the proposed method as a building block for large-scale incremental learning of object models from Internet data.

Acknowledgement

This work was supported by NSERC and the University of Manitoba Research Grants Program (URGP). We gratefully acknowledge the support of NVIDIA Corporation with the GPU donation used in this research.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010.

[2] R. G. Cinbis, J. Verbeek, and C. Schmid. Multi-fold mil training for weakly supervised object localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[3] J. Deng, A. C. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10,000 image categories tell us? In *European Conference on Computer Vision*, 2010.

[4] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *International Journal of Computer Vision*, 100(3):257–293, 2012.

[5] I. Endres and D. Hoiem. Category independent object proposals. In *European Conference on Computer Vision*, 2010.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[7] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009.

[8] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, April 2006.

[9] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. DeViSE: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, 2013.

[10] R. B. 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*, 2014.

[11] G. Hartmann, M. Grundmann, J. Hoffman, D. Tsai, V. Kwatra, O. Madani, S. Vijayanarasimhan, I. A. Essa, J. Rehg, and R. Sukthankar. Weakly supervised learning of object segmentations from web-scale video. In *ECCV Workshop on Web-scale Vision and Social Media*, 2012.

[12] J. Hoffman, S. Guadarrama, E. S. Tzeng, J. Donahue, T. Darrell, K. Saenko, and R. B. Girshick. LSDA: Large scale detection through adaptation. In *Advances in Neural Information Processing Systems*. MIT Press, 2014.

[13] E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng. Improving word representations via global context and multiple word prototypes. In *Annual Meeting of the Asscoation for Computational Linguistics*, 2012.

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014.

[15] A. Joulin, K. Tang, and L. Fei-Fei. Efficient image and video co-localization with frank-wolfe algorithm. In *European Conference on Computer Vision*, 2014.

[16] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image-sentence mapping. In *Advances in Neural Information Processing Systems*, 2014.

[17] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009.

[18] Y. J. Lee and K. Grauman. Object-graphs for context-aware category discovery. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010.

[19] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. MIT Press, 2013.

[20] M. H. Nguyen, L. Torresani, F. de la Torre, and Carsten. Weakly supervised discrimiative localization and classificaition: a joint learning approach. In *IEEE International Conference on Computer Vision*, 2009.

[21] A. Papazoglou and V. Ferrari. Fast object segmentation in unconstrained video. In *IEEE International Conference on Computer Vision*, 2013.

[22] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[23] M. Rohrbach, M. Stark, and B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[24] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps where - and why? semantic relatedness for knowledge transfer. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010.

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. arXiv, 2014.

[26] Z. Shi, T. M. Hospedales, and T. Xiang. Bayesian joint topic modelling for weakly supervised object localization. In *IEEE International Conference on Computer Vision*, 2013.

[27] P. Siva, C. Russell, and T. Xiang. In defense of negative mining for annotating weakly labelled data. In *European Conference on Computer Vision*, 2012.

[28] P. Siva, C. Russell, T. Xiang, and L. Agapito. Looking beyong the image unsupervised learning for object saliency and detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[29] P. Siva and T. Xiang. Weakly supervised object detector learning with model drift detection. In *IEEE International Conference on Computer Vision*, 2011.

[30] M. Stark, M. Goesele, and B. Schiele. A shape-based object class model for knowledge transfer. In *IEEE International Conference on Computer Vision*, 2009.

[31] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei. Co-localization in real-world images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[32] K. Tang, R. Sukthankar, J. Yagnik, and L. Fei-Fei. Discriminative segment annotation in weakly labeled video. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013.

[33] T. Tommasi and B. Caputo. The more you know, the less you learn: from knowledge transfer to one-shot learning of object categories. In *British Machine Vision Conference*, 2009.

[34] L. van der Maaten and G. E. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[35] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly supervised semantic segmentation with a multi-image model. In *IEEE International Conference on Computer Vision*, 2011.

[36] C. L. Zitnick and P. Doll. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, 2014.