Weakly Supervised Localization of Novel Objects Using Appearance Transfer

Mrigank Rochan, Yang Wang
Department of Computer Science, University of Manitoba, Canada.

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. The main weakness of these approaches 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. In this paper, we show that it is possible to transfer appearance model from one object class to another based on their semantic relationship.

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/video 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. 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.

An overview of our approach is illustrated in Fig. 1. 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. The second appearance model is obtained by transferring knowledge from other familiar objects. 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.

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. We use the edge boxes method in [2] for generating bounding boxes as our object proposals. We 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 $\mathbf{w}_p^\top \mathbf{x}$ indicates the likelihood of $\mathbf{x}$ being the novel object.

Next, we propose another way of constructing the appearance model by transferring knowledge from other familiar objects. First, we use the word vectors [1] 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.

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.:

$$\mathbf{v} \approx \theta_1 \mathbf{v}_1 + \theta_2 \mathbf{v}_2 + \ldots \theta_K \mathbf{v}_K \qquad (1)$$

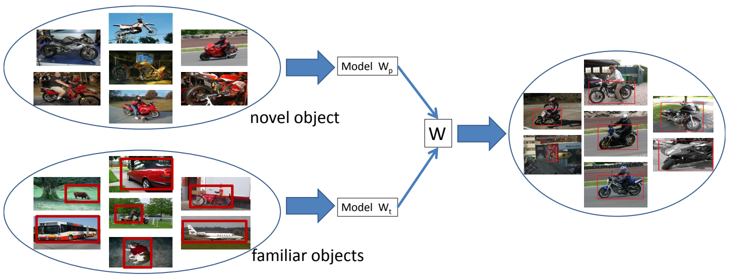

Figure 1: An overview of our approach. (Top left) Given a collection of weakly labeled images of a novel object (e.g. motorbike), we learn an appearance model $\mathbf{w}_p$ from the object proposals. (Bottom left) We also transfer the knowledge of pre-trained models for a set of familiar objects, e.g. car, bus, dog, etc. and obtain another appearance model $\mathbf{w}_t$ for the novel object. (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.

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:

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

By solving Eq. 2, 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 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:

$$\mathbf{w}_t = \theta_1 \mathbf{u}_1 + \theta_2 \mathbf{u}_2 + \ldots \theta_K \mathbf{u}_K \qquad (3)$$

Our final appearance model $\mathbf{w}$ for the novel object is a linear combination of these two:

$$\mathbf{w} = \gamma \mathbf{w}_p + \mathbf{w}_t \qquad (4)$$

where γ is a parameter that controls the relative importance of $\mathbf{w}_p$ and $\mathbf{w}_t$. It also vary depending on the “transferability” of the novel object. We examine the reconstruction error in Eq. 2 to determine the “transferability”. Let $\Theta^* = [\theta_1^*, \theta_2^*, ..., \theta_K^*]^\top$ be the solution to Eq. 2, the reconstruction error is:

$$E(\Theta^*) = ||\mathbf{v} - (\theta_1^* \mathbf{v}_1 + \theta_2^* \mathbf{v}_2 + \ldots \theta_K^* \mathbf{v}_K)||_2^2 \qquad (5)$$

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

$$\mathbf{w} = \beta \cdot E(\Theta^*) \cdot \mathbf{w}_p + \mathbf{w}_t \qquad (6)$$

We then use this appearance model $\mathbf{w}$ to re-score the object proposals generated in novel object images. 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.

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 the effectiveness of our approach.

[1] Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. Improving word representations via global context and multiple word prototypes. In *ACL*, 2012.

[2] C. Lawrence Zitnick and Piotr Doll. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.