Transfer Learning by Ranking for Weakly Supervised Object Annotation

Zhiyuan Shi
zhiyuan.shi@eecs.qmul.ac.uk
Parthipan Siva
psiva@eecs.qmul.ac.uk
Tao Xiang
txiang@eecs.qmul.ac.uk

School of Electronic Engineering and Computer Science,
Queen Mary, University of London,
London E1 4NS, UK

Object detectors [5] locate objects of interest in images and have many applications including image tagging, consumer photography, and surveillance. Most existing object detectors take a fully supervised learning (FSL) approach, where all the training images are manually annotated with the object location. However, manual annotation of hundreds of object categories is time-consuming, laborious, and subjective to human bias. To reduce the amount of manual annotation, a weakly supervised learning (WSL) [3, 6] approach is desired. In WSL, the training set is only annotated with a binary label indicating the presence or absence of the object of interest, not the location or extent of the object (Fig. 1(a)).

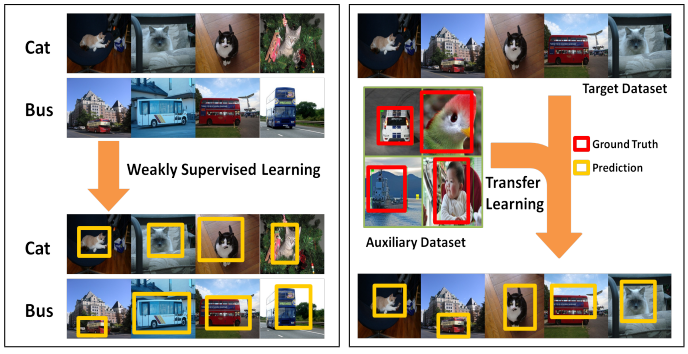

Figure 1: (a) Weakly supervised learning approach for automatic annotation of objects [3, 6]. (b) Our transfer learning approach for automatic annotation of objects.

Typically three information cues, saliency, inter-class, and intra-class, are used to locate or annotate the object of interest in images known to contain the object of interest (positive images). Saliency information ensures that the annotated region is a foreground region. Inter-class information ensures that the annotated regions look dissimilar to all images without the object of interest (negative images). Intra-class information ensures that the annotated regions in all positive images look similar to each other. Methods that use saliency alone [1] select salient regions in each positive image independently. Methods that use inter-class and intra-class information [3, 6] typically use saliency to limit the search space of each image by only looking at the most salient regions; then they select one of these salient regions by maximising the inter-class and intra-class information.

In this paper we utilise a fourth information cue (Fig. 1(b)) which is typically neglected by other approaches: an auxiliary fully annotated dataset. While we want to reduce manual annotation when learning new object categories, we cannot ignore the fact that there exist many datasets which already have manual annotation of object locations [4]. However, these auxiliary datasets seem unhelpful since they often contain object categories that are unrelated to the target object category we wish to annotate. For example, an auxiliary dataset might contain annotations of cars, birds, boats and person but a target object category might be cats and buses (Fig. 1(b)). So what information can we actually transfer? When adopting the strategy of selecting the optimal object location from a set of candidate salient regions [3, 6], the performance of the selection can obviously be measured by examining the degree of overlap between the selected region and the ground truth region (Fig. 2(a)). One can safely assumes that the more a salient region overlaps the ground truth region, the more similar the two's appearances are. In other words, there exists a mapping relationship between the degree of overlap (hence the accuracy of annotation) and the appearance similarity. This relationship should hold true regardless of the object category and is what we propose to learn and transfer to the target data. To quantify this mapping relationship, one must take into consideration the high dimensionality typical for representing object appearance and the inevitable noise. To this end, we formulate a ranking based transfer learning model which, once learned, takes appearance similarity as input and predicts the ranking order among all the candidate salient regions according to their degree of overlap with the (unknown) true object location.

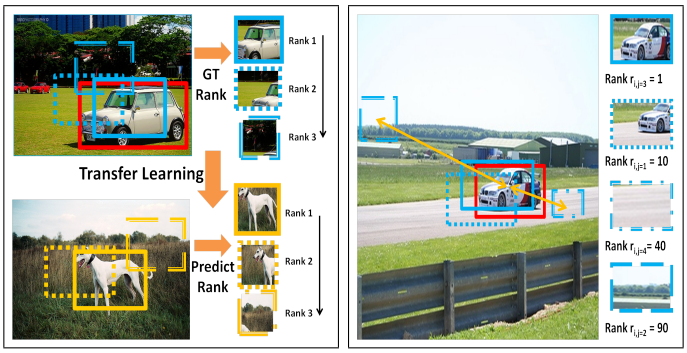

Figure 2: (a) The mapping relationship between the degree of overlap and the appearance similarity between salient regions and the ground truth location is transferred from the auxiliary data to the target data. (b)A higher rank is given to salient regions (blue) with higher overlap to the ground truth (red). In this image, a road region ($j = 4$) which contains more relevant context to the car, is ranked higher than a sky region ($j = 2$) according to their distances to the ground truth region (red).

More specifically, for each image $i \in \mathcal{A}$, we represent each of the N salient regions with an unnormalised BOW histogram $x_{i,j}$, where $j = 1 \ldots N$. To compute a feature from $x_{i,j}$ that is independent of the object category we define a difference vector $d_{i,j}$ as the feature of interest:

$$d_{i,j} = \left| \frac{x_{i,j}}{\|x_{i,j}\|_1} - \frac{g_i}{\|g_i\|_1} \right|, \quad (1)$$

All N salient regions are sorted by its overlap with the ground truth bounding box (Fig. 2(b)), where overlap is defined by [4] as the intersect area divided by union area. They will form enormous preference pairs for a moderate number of images and salient regions in each image (N). For efficient learning, we use the primal-based pairwise RankSVM algorithm proposed in [2] to minimise the objective function:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{(k,l)\in\mathcal{P}} \ell(\mathbf{w}^T \hat{d}_{i,k} - \mathbf{w}^T \hat{d}_{i,l}), \quad (2)$$

We show that our novel transfer learning model outperforms the state-of-the-art WSL approaches on the challenging PASCAL VOC 2007 dataset.

[1] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *CVPR'10*, 2010.

[2] O. Chapelle and S.S. Keerthi. Efficient algorithms for ranking with svms. *Inf. Retr.*, 13(3):201–215, June 2010.

[3] T. Deselaers, B. Alexe, and V. Ferrari. Localizing objects while learning their appearance. ECCV'10, 2010.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

[5] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 2010.

[6] P. Siva and T. Xiang. Weakly supervised object detector learning with model drift detection. In *ICCV'11*, 2011.