Expanding Object Detector's HORIZON: Incremental Learning Framework for Object Detection in Videos

Alina Kuznetsova[1,3], Sung Ju Hwang[2], Bodo Rosenhahn[1], Leonid Sigal[3]
[1]Leibniz University Hannover, [2]UNIST, [3]Disney Research Pittsburgh

Introduction and Motivation: Over the past several years it has been shown that there are significant biases among object detection datasets [2]. To handle such biases a number of domain adaptation methods have been proposed. While most domain adaptation techniques focus on applications where both training and test instances are images, a few address the problem in the context of image-to-video object detector adaptation [1, 5]. Image-to-video detector adaptation is conceptually appealing but at the same time very challenging. Web images tend to be of high resolution and are object-centric, where as videos often come at lower resolution, are not object centric and contain motion artifacts. Furthermore, while most of the domain adaptation techniques assume that data is separated into well defined discrete domains, many factors, such as changes in appearance and lighting, are inherently continuous and are better described as *expansions* of the original object domain (as opposed to instantiations of new domain).

We propose an incremental approach to expanding the object detector domain from images to unlabeled videos (see Figure 1). We start from a large-margin embedding (LME) classification model [6], which we adapt to a detection task. Given this LME formulation, we further propose a multi-prototype probabilistic extension of LME, which allows our model to perform intuitive confidence evaluation for test instances and supports incremental learning. Using this initial probabilistic LME model, objects in the arriving unlabeled videos are found, and tracks associated with most confident instances are extracted. If instances from these tracks form a cluster, they are used to adjust the complexity of the model, by adding additional prototypes; this procedure effectively expands domain of the detector. We show incremental domain expansion is effective in applying object detectors, trained with only ImageNet, to videos, improving performance by approximately 48% on Activities of Daily Living (ADL) dataset [3] and by 15% on the YouTube Objects dataset [4].

Initial model: Given a training image set $\{\mathbf{x}_i, y_i\}_{i=1}^{N_\mathcal{I}}$ where $\mathbf{x}_i \in \mathbb{R}^D$ is a feature descriptor of an image patch and $y_i \in \{1,\dots,C\}$ is the object label, LME learns a linear low-dimensional embedding defined by a projection matrix $\mathbf{W} \in \mathbb{R}^{d\times D}$, together with class prototypes $\mathbf{u}_c \in \mathbb{R}^d, c = \{1\dots C\}$, in the embedding space, such that a sample projected into this low dimensional space is closer to the correct class prototype than to all other prototypes:

$$\sum_{i,c:c\neq y_i} \xi_{ic}^+ + \lambda\|\mathbf{W}\|_{FRO}^2 + \gamma\|\mathbf{U}\|_{FRO}^2$$
$$d_\mathbf{W}(\mathbf{x}_i, \mathbf{u}_{y_i}) + \xi_{ic} \geq d_\mathbf{W}(\mathbf{x}_i, \mathbf{u}_c) + 1,$$
$$i = \{1\dots N\}, c = \{1\dots C\}, c \neq y_i,$$

where $d_\mathbf{W}(\mathbf{x}_i, \mathbf{u}_{y_i})$ defines similarity between a sample $\mathbf{x}_i$ and a prototype $\mathbf{u}_{y_i}$. To extend the model for the detection, we define a patch as not containing an object if it is sufficiently *dissimilar* to *all* known object class prototypes.

Incremental learning: To allow incremental learning, we further extend the model to have multiple, prototypes per class c:

$$S_\mathbf{W}^\alpha(\mathbf{x}_i, \mathbf{U}_c) = \frac{\sum_{k=1}^{K_c} d_\mathbf{W}(\mathbf{x}_i, \mathbf{u}_c^k) e^{\alpha d_\mathbf{W}(\mathbf{x}_i, \mathbf{u}_c^k)}}{\sum_{j=1}^{K_c} e^{\alpha d_\mathbf{W}(\mathbf{x}_i, \mathbf{u}_c^j)}}.$$

Suppose we want to expand the prototype-based representation for the class c_n. Then, the newly added prototype $\mathbf{u}_{c_n}$ should satisfy two properties: (i) the new prototype should be representative and discriminative for its class; (ii) it should not cause misclassification of samples from other classes. Let $\tilde{\mathbf{U}}_{c_n} = [\mathbf{U}_{c_n}, \mathbf{u}_{c_n}]$, then, more formally, to update the model we optimize the following objective:

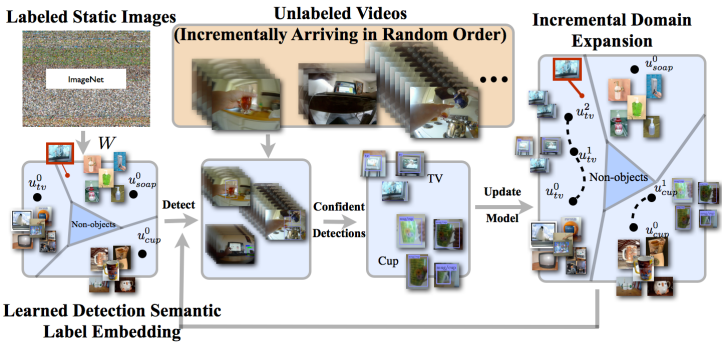

Figure 1: Illustration of the the proposed learning framework. Note that while a *TV* test sample (in red) may be too far in appearance from the original ImageNet trained model and hence misclassified, new prototypes, added based on tracks from videos, help to bridge the gap leading to correct classification.

minimize:

$$\sum_{\substack{i,c:y_i=c_n,\\ c\neq c_n}} \xi_{ic}^+ + \sum_{i:y_i\neq c_n} \zeta_i^+ + \sum_j \xi_{j0}^+ + \nu\|\mathbf{u}_{c_n} - \mathbf{u}_0\|^2 + \eta\|\mathbf{W} - \mathbf{W}_0\|^2,$$

subject to:

$$S_\mathbf{W}^\alpha(\mathbf{x}_i, \tilde{\mathbf{U}}_{c_n}) + \xi_{ic} \geq S_\mathbf{W}^\alpha(\mathbf{x}_i, \mathbf{U}_c) + 1, y_i = c_n$$
$$S_\mathbf{W}^\alpha(\mathbf{x}_i, \mathbf{U}_{y_i}) + \zeta_i \geq S_\mathbf{W}^\alpha(\mathbf{x}_i, \tilde{\mathbf{U}}_{c_n}) + 1, y_i \neq c_n$$
$$S_\mathbf{W}^\alpha(\mathbf{x}_j^0, \tilde{\mathbf{U}}_{c_n}) \leq 1 + \xi_{j0},$$

The above incremental update is especially beneficial, when not all data is available and the newly arriving data has evolving feature distribution.

Conclusion: We have developed a novel online multi-prototype large margin embedding model with detection constraints, which incrementally increase in number as self-paced learning algorithm selects confident samples from the incoming unlabeled videos to add. Our incremental domain expansion could serve as a lifelong learning system for object detection—as the model expands to encompass continuous stream of unlabeled new videos.

[1] Adrien Gaidon, Gloria Zen, and Jose A. Rodriguez-Serrano. Self-learning camera: Autonomous adaptation of object detectors to unlabeled video streams. In *Arxiv*, 2014.

[2] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A. Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *ECCV*, 2012.

[3] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012.

[4] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, 2012.

[5] Kevin Tang, Vignesh Ramanathan, Li Fei-Fei, and Daphne Koller. Shifting weights: Adapting object detectors from image to video. In *NIPS*, 2012.

[6] Kilian Q. Weinberger and Olivier Chapelle. Large margin taxonomy embedding for document categorization. In *NIPS*. 2009.

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