Incremental domain expansion for object detection in videos

Alina Kuznetsova

Joint with Sung Ju Hwang[1] and Leonid Sigal[2]

[1]Ulsan National Institute of Science and Technology
[2]Disney Research

Domain adaptation problem

‘Classical’ single-domain problem:

$$(x, y) \sim P(x, y)$$

Multiple domains:

Source domain: $(x, y) \sim P_S(x, y)$

Target domain: $(x, y) \sim P_T(x, y)$

$$P_S(x, y) \neq P_T(x, y)$$

Examples:

- Speech recognition (each subject - single domain)
- Natural language processing
- Computer vision

Domain adaptation problem: from images to videos

‘Classical’ scenario:

Train date

Test data

Domain adaptation problem: from images to videos

Real world:

Train date

Test data

Related work

B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. CVPR, 2012

M. Long, J. Wang, G. Ding, J. Sun, P. S. Yu., Transfer joint matching for unsupervised domain adaptation. CVPR 2014

K. Tang, V. Ramanathan, L. Fei-Fei, and D. Koller. Shifting weights: Adapting object detectors from image to video. NIPS, 2012

J. Donahue, J. Hoffman, E. Rodner, K. Saenko, and T. Darrell. Semi-supervised domain adaptation with instance constraints. CVPR, 2013

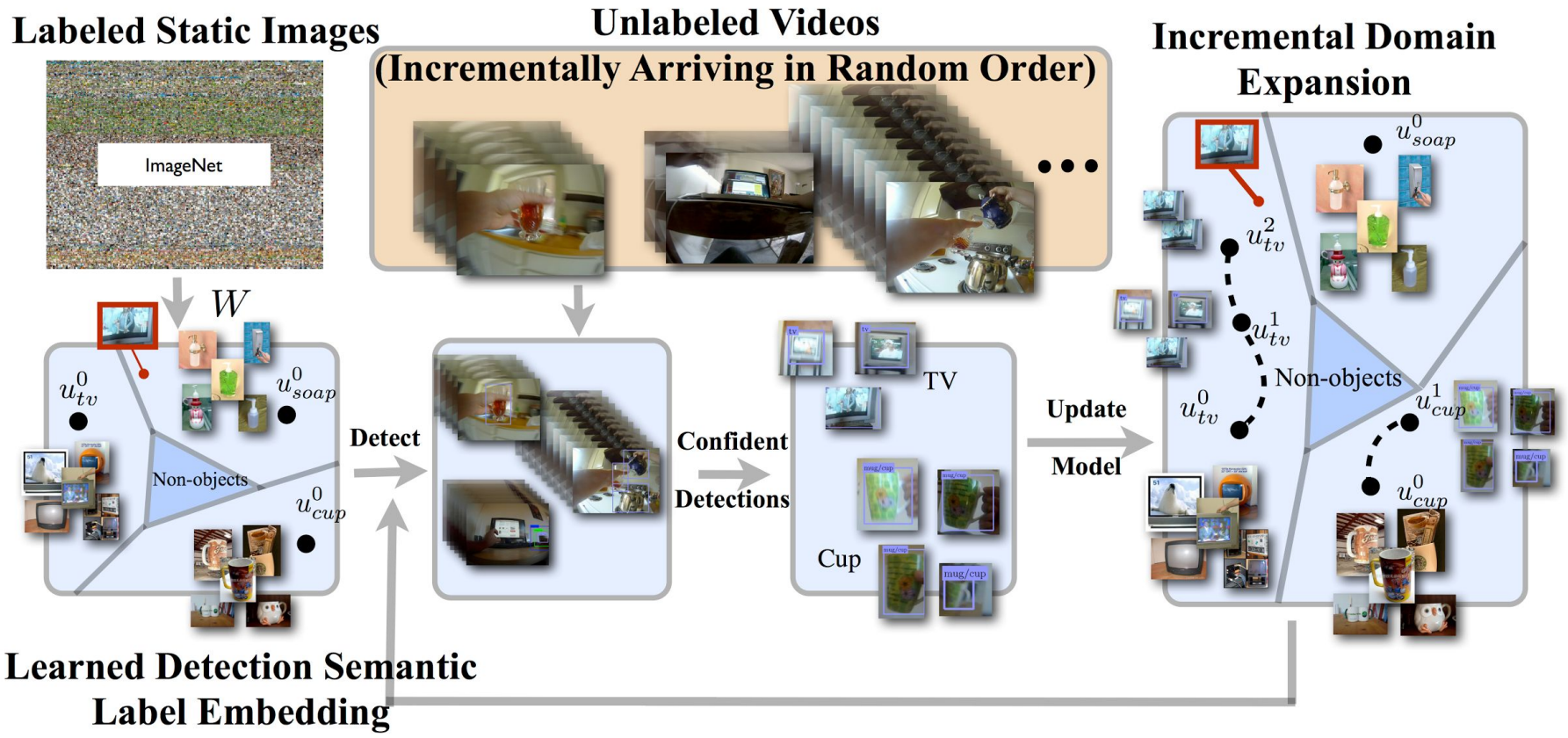
Goals

- Being able to adjust to multiple domains simultaneously without losing performance on previously seen domains [unsupervised].
- Improve performance on the similar but unseen domains.

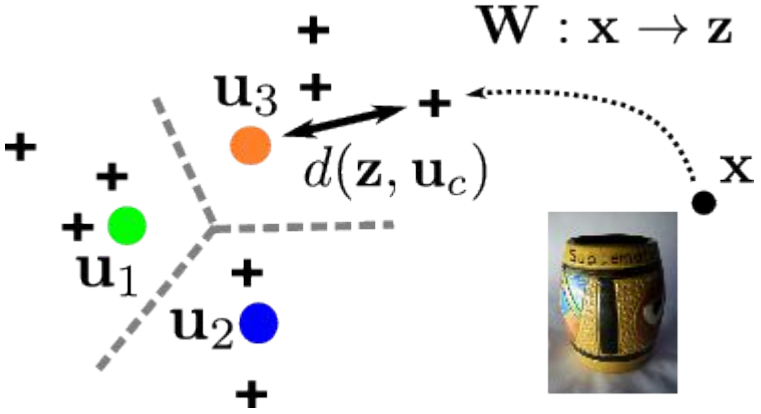
Two key ideas:

- Dynamically adjust number of model parameters as more data is seen by the model.
- Use temporal consistency to extract new informative samples.

Framework overview



Large Margin Embedding (LME-D)[1]



Similarity:

$$d_{\boldsymbol{W}}(\boldsymbol{x}, \boldsymbol{u}_c) = (\boldsymbol{W}\boldsymbol{x})^T \boldsymbol{u}_c$$

Minimize:

$$\sum_{i,c:c\neq y_i} \xi_{ic}^{+} + \sum_j \xi_{j0}^{+} + \lambda \|\mathbf{W}\|_{FRO}^2 + \gamma \|\mathbf{U}\|_{FRO}^2$$

Inter-class constraints:

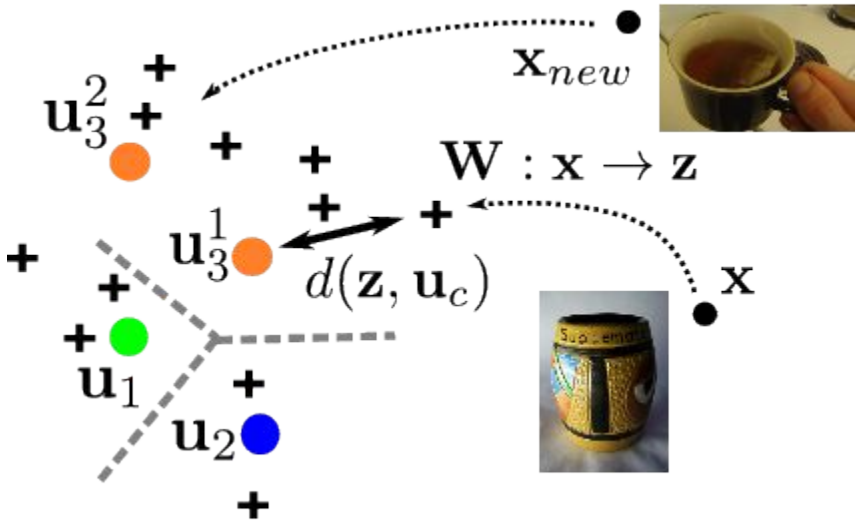
$$d_{\mathbf{W}}(\mathbf{x}_i, \mathbf{u}_{y_i}) + \xi_{ic} \geq d_{\mathbf{W}}(\mathbf{x}_i, \mathbf{u}_c) + 1$$

Detection constraints:

$$d_{\mathbf{W}}(\mathbf{x}_j^0, \mathbf{u}_c) \leq 1 + \xi_{j0}$$

[1] K. Q. Weinberger and O. Chapelle. Large margin taxonomy embedding for document categorization. NIPS 2009

LME extension for online updates



Smooth maximum approximation:

$$s(\{x_i\}_{i=1}^n)_\alpha = \frac{\sum_{i=1}^n x_i e^{\alpha x_i}}{\sum_{i=1}^n e^{\alpha x_i}}$$

Similarity:

$$S_{\boldsymbol{W}}^{\alpha}(\boldsymbol{x}, \boldsymbol{U}_c) = s_\alpha(\{d_{\boldsymbol{W}}(\boldsymbol{x}, \boldsymbol{u}_c^i)\}_i)$$

LME extension for online updates

Minimize w.r.t. $\boldsymbol{u}_{c_n}, \boldsymbol{W}$

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

Representativeness constraints:

$$S_{\mathbf{W}}^{\alpha}(\mathbf{x}_i, \tilde{\mathbf{U}}_{c_n}) + \xi_{ic} \geq S_{\mathbf{W}}^{\alpha}(\mathbf{x}_i, \mathbf{U}_c) + 1$$

Discriminativeness constraints:

$$S_{\mathbf{W}}^{\alpha}(\mathbf{x}_i, \mathbf{U}_{y_i}) + \zeta_i \geq S_{\mathbf{W}}^{\alpha}(\mathbf{x}_i, \tilde{\mathbf{U}}_{c_n}) + 1$$

Detection constraints: $S_{\mathbf{W}}^{\alpha}(\mathbf{x}_j^0, \tilde{\mathbf{U}}_{c_n}) \leq 1 + \xi_{j0}$

Samples selection

1. Select a confident detection:

2. Propagate in time using optical flow:

3. Score the accept/reject the tube.

4. Select samples based on confidence score

Evaluation

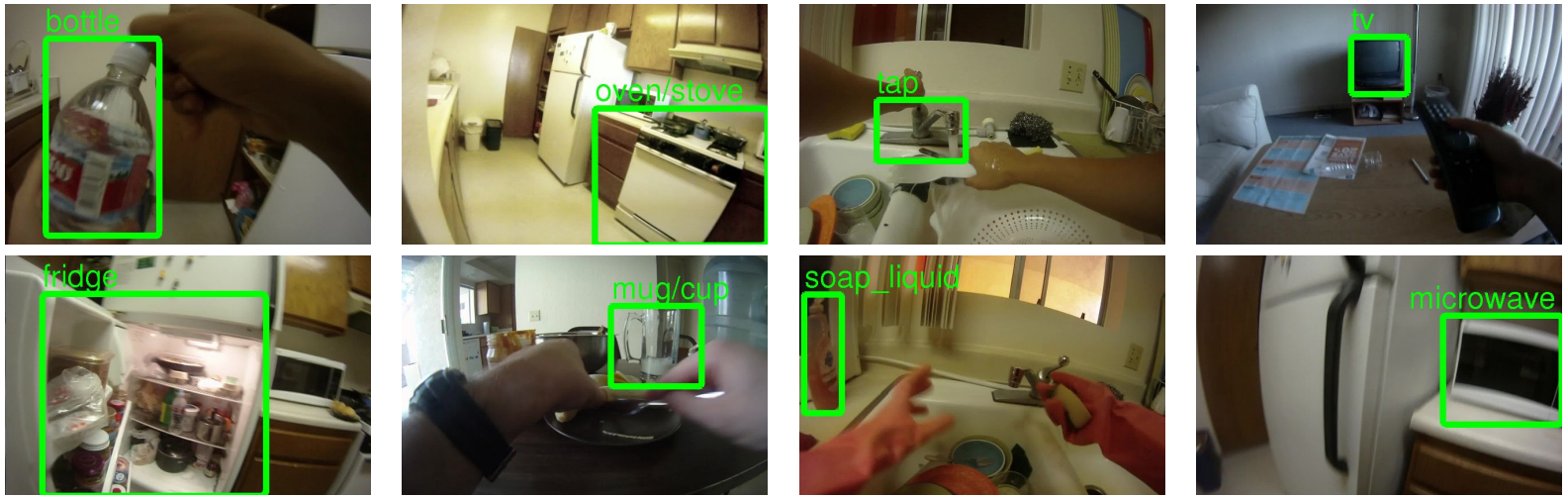
Activities of Daily Living Dataset[1]:

	mwave(*)	soap(**)	mAP ADL	ImageNet
DPM[1]	20.1	**2.5**	9.35	-
GK[2]	41.19	0.2	17.73	-
LME-A	40.30	0.37	18.36	76.96
LME-D	39.87	0.35	14.78	78.91
IDE-LME	**56.69**	0.25	**21.91**	**79.23**

Model comparison with the baseline in terms of mAP and accuracy on the training domain (ImageNet); (*) - best performing class, (**) -worst performing class.

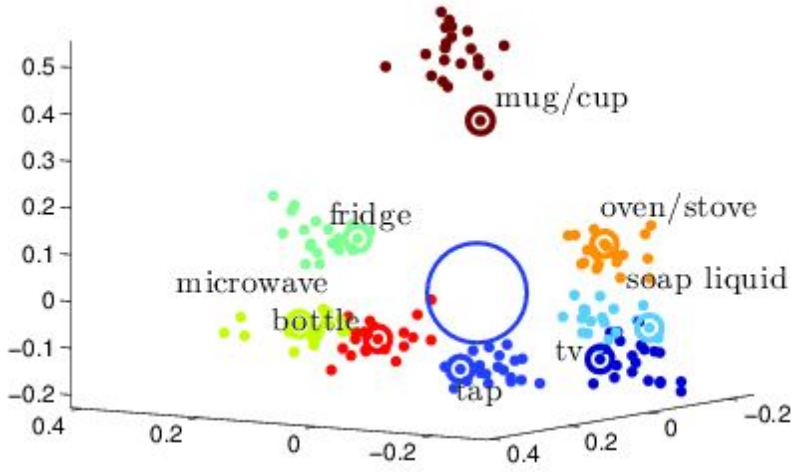
[1] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In CVPR 2012
[2] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In CVPR, 2012

Evaluation: detection examples



Examples of the correct detections of IDE-LME.

Projection of the learned embedding:



Evaluation

YouTube Objects Dataset[4] (test part):

	boat(*)	horse(**)	mAP YTO	ImageNet
DPM[3]	0.97	**26.78**	24.31	-
GK[2]	24.44	17.75	27.17	-
LME-A	35.20	25.86	**30.80**	**79.91**
LME-D	32.39	9.22	23.63	83.16
IDE-LME	**42.26**	11.83	**27.28**	**83.20**

Model comparison with the baseline in terms of CorLoc and accuracy on the training domain (ImageNet); (*) - best performing class, (**) -worst performing class

[3] V. Kalogeiton, V. Ferrari, and C. Schmid. Analysing domain shift factors between videos and images for object detection. Arxiv, 2015
[4] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In CVPR, June 2012

Discussion and future work

Takeaway: unlabeled data can be used to increase model performance in unsupervised manner.

Future research directions:

- Easily extendable to fully neural-network-based model.

Z. Li, D. Hoiem “Learning without Forgetting”, ECCV 2016

- Samples selection - how to select the most informative samples without introducing drift?

- How to select neighbouring domains?