Detector Discovery in the Wild: Joint Multiple Instance and Representation Learning

Judy Hoffman[1], Deepak Pathak[1], Trevor Darrell[1], Kate Saenko[2]
[1]UC Berkeley. [2]UMass Lowell.

It is well known that contemporary visual models thrive on large amounts of training data, especially those that directly include labels for desired tasks. Many real world settings contain labels with varying specificity, e.g., “strong" bounding box detection labels, and “weak" labels indicating presence somewhere in the image. We tackle the problem of *joint detector and representation learning*, and develop models which cooperatively exploit heterogeneous sources of training data, where some classes have no “strong" annotations. Our model optimizes a latent variable multiple instance learning model over image regions while simultaneously transferring a shared representation from detection-domain models to classification-domain models. The latter provides a key source of automatic and accurate initialization for latent variable optimization, which has heretofore been unavailable in such methods.

Previous methods employ varying combinations of weak and strong labels of the same object category to learn a detector. Such methods seldom exploit available strong-labeled data of different, auxiliary categories, despite the fact that such data is very often available in many practical scenarios. Deselaers et.al. [2] uses auxiliary data to learn generic objectness information just as an initial step, but doesn’t optimize jointly for weakly labeled data.

We introduce a new model for large-scale learning of detectors that can jointly exploit weak and strong labels, perform inference over latent regions in weakly labeled training examples, and can transfer representations learned from related tasks (see Figure 1). In practical settings, such as learning visual detector models for all available ImageNet categories, or for learning detector versions of other defined categories such as Sentibank’s adjective-noun-phrase models [1], our model makes greater use of available data and labels than previous approaches. Our method takes advantage of such data by using the auxiliary strong labels to improve the feature representation for detection tasks, and uses the improved representation to learn a stronger detector from weak labels in a deep architecture.

To learn detectors, we exploit weakly labeled data for a concept, including both “easy" images (e.g., from ImageNet classification training data), and “hard" weakly labeled imagery (e.g., from PASCAL or ImageNet detection training data with bounding box metadata removed). We define a novel multiple instance learning (MIL) framework that includes bags defined on both types of data, and also jointly optimizes an underlying perceptual representation using strong detection labels from related categories. The latter takes advantage of the empirical results in [3], which demonstrated knowledge of what makes a good perceptual representation for detection tasks could be learned from a set of paired weak and strong labeled examples, and the resulting adaptation could be transferred to new categories, even those for which no strong labels were available. An example of discovered regions for the category “tennis racket” are depicted in Figure 2.

We evaluate our model empirically on the largest set of available ground-truth visual detection data, the ImageNet-200 category challenge. Our method outperforms the previous best MIL-based approaches for held-out detector learning on ImageNet-200 [4] by 200%, and outperforms the previous best domain-adaptation based approach [3] by 12%. Our model is directly applicable to learning improved “detectors in the wild", including categories in ImageNet but not in ImageNet-200, or categories defined ad-hoc for a particular user or task with just a few training examples to fine-tune a new classification model. Such models can be promoted to detectors with no (or few) labeled bounding boxes.

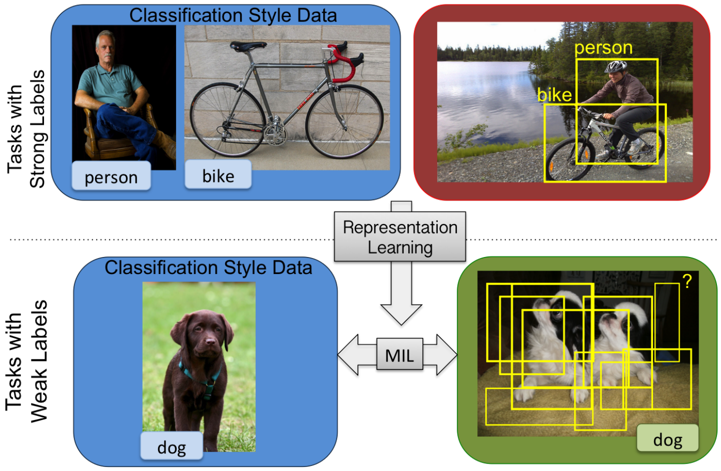

Figure 1: We learn detectors for categories with only weak labels (*bottom row*), by jointly transferring a representation from auxiliary categories with available strong annotations (*top row*) and solving an MIL problem on the weakly annotated data (green box).

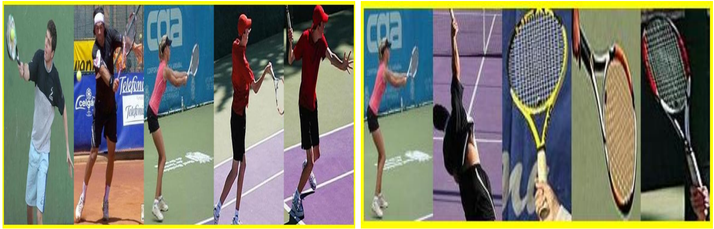

Figure 2: Example mined bounding boxes learned using our method. Left side shows the mined boxes after fine-tuning with images in classification settings only, and right side shows the mined boxes after fine-tuning with auxiliary strongly annotated dataset. We show top 5 mined boxes across the dataset for the category “tennis racket”. None of the discovered patches from the original feature space correctly located the tennis racket and instead included the person as well. After incorporating the strong annotations from auxiliary tasks, our method starts discovering tennis rackets, though still has some confusion with the person playing tennis.

[1] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih Fu Chang. Large-scale visual sentiment ontology and detectors using adjective nown paiars. In *ACM Multimedia Conference*, 2013.

[2] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Weakly supervised localization and learning with generic knowledge. *IJCV*, 2012.

[3] Judy Hoffman, Sergio Guadarrama, Eric Tzeng, Ronghang Hu, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. LSDA: Large scale detection through adaptation. In *Neural Information Processing Systems (NIPS)*, 2014.

[4] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla amd Michael Bernstein, and Alexander. Imagenet large scale visual recognition challenge. arXiv:1409.0575, 2014.

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