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

Judy Hoffman, Deepak Pathak, Trevor Darrell
UC Berkeley
{jhoffman, pathak, trevor}@eecs.berkeley.edu

Kate Saenko
UMass Lowell
saenko@cs.uml.edu

Abstract

We develop methods for detector learning which exploit joint training over both weak (image-level) and strong (bounding box) labels and which transfer learned perceptual representations from strongly-labeled auxiliary tasks. Previous methods for weak-label learning often learn detector models independently using latent variable optimization, but fail to share deep representation knowledge across classes and usually require strong initialization. Other previous methods transfer deep representations from domains with strong labels to those with only weak labels, but do not optimize over individual latent boxes, and thus may miss specific salient structures for a particular category. We propose a model that subsumes these previous approaches, and simultaneously trains a representation and detectors for categories with either weak or strong labels present. We provide a novel formulation of a joint multiple instance learning method that includes examples from classification-style data when available, and also performs domain transfer learning to improve the underlying detector representation. Our model outperforms known methods on ImageNet-200 detection with weak labels.

1. Introduction

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

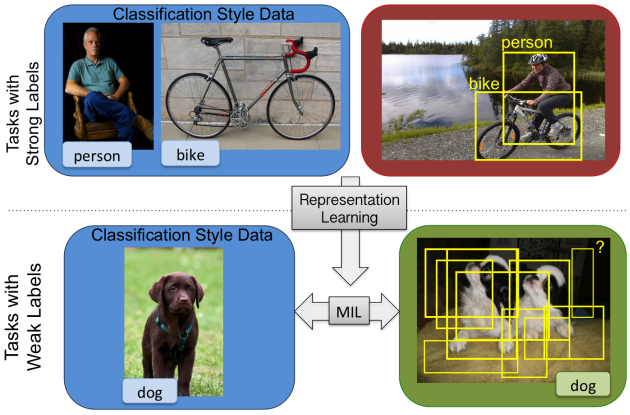

Figure 1: We learn detectors for categories with only weak labels (*bottom row*), by jointly transferring a representation from auxiliary categories with available strong annotations (*top row*) and solving an MIL problem on the weakly annotated data (green box).

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*. [10] 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 mod-

els [7], 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 [19], 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.

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 [27] by 200%, and outperforms the previous best domain-adaptation based approach [19] 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.

2. Related Work

CNNs for Visual Recognition Within the last few years, convolutional neural networks (CNNs) have emerged as the clear winners for many visual recognition tasks. A breakthrough was made when the positive performance demonstrated for digit recognition [25] began to translate to the ImageNet [27] classification challenge winner [22]. Shortly thereafter, the feature space learned through these architectures was shown to be generic and effective for a large variety of visual recognition tasks [12, 40]. These results were followed by state-of-the-art results for object detection [16, 29]. Most recently, it was shown that CNN architectures can be used to transfer generic information between the classification and detection tasks [19], improving detection performance for tasks which lack bounding box training data.

Training with Auxiliary Data Sources There has been a large amount of prior work on training models using auxiliary data sources. The problem of visual domain adaptation is precisely seeking to use data from a large auxiliary source domain to improve recognition performance on a target domain which has little or no labeled data available. Techniques to solve this problem consist of learning a new feature representation that minimizes the distance between source and target distributions [28, 23, 17, 15], regularizing the learning of a target classifier against the source model [37, 4, 9], or doing both simultaneously [20, 13].

Multiple Instance Learning Since its inception, the MIL [11] problem has been attempted in several frameworks including Noisy-OR [18], boosting [2, 41] etc. But most commonly, it was framed as a max-margin classification problem [3] with latent parameters optimized using alternating optimization [14, 38]. Overall, MIL is tackled in two stages: first finding better initialization, and then using better heuristics for optimization. A number of methods have been proposed for initialization which include using large image region excluding boundary [26], using candidate set which covers the training data space [33, 34], using unsupervised patch discovery [32, 30], learning generic objectness knowledge from auxiliary catgories [1, 10], learning latent categories from background to suppress it [36] or using class-specific similarity [31]. Approaches to better optimize the non-convex problem involve using multi-fold learning as a measure of regularizing overfitting [8], optimize Latent SVM for the area under the ROC curve (AUC) [6] and training with easy examples in beginning to avoid bad local optimization [5, 24]. Most of these approaches perform reasonably only when object covers most of the region of image, or when most of the candidate regions contain an object. The major challenge faced by MIL in general is that of fixed feature representation, and poor initialization particularly in non-object centric images. Our algorithm provides solutions to both of these issues.

3. Background: MI-SVM

We begin by briefly reviewing a standard solution to the multiple instance learning problem, Multiple Instance SVMs (MI-SVMs) [3] or Latent SVMs [14, 38]. In this setting, each weakly labeled image is considered a collection of regions which form a positive 'bag'. For a binary classification problem, the task is to maximize the bag margin which is defined by the instance with highest confidence. For each weakly labeled image $I \in \mathcal{W}$, we collect a set of regions of interest and define the index set of those regions as R_I. We next define a bag as $B_I = \{\boldsymbol{x}_i | i \in R_I\}$, with label Y_I, and let the i^{th} instance in the bag be $(\boldsymbol{x}_i, y_i) \in \mathcal{R}^p \times \{-1, +1\}$.

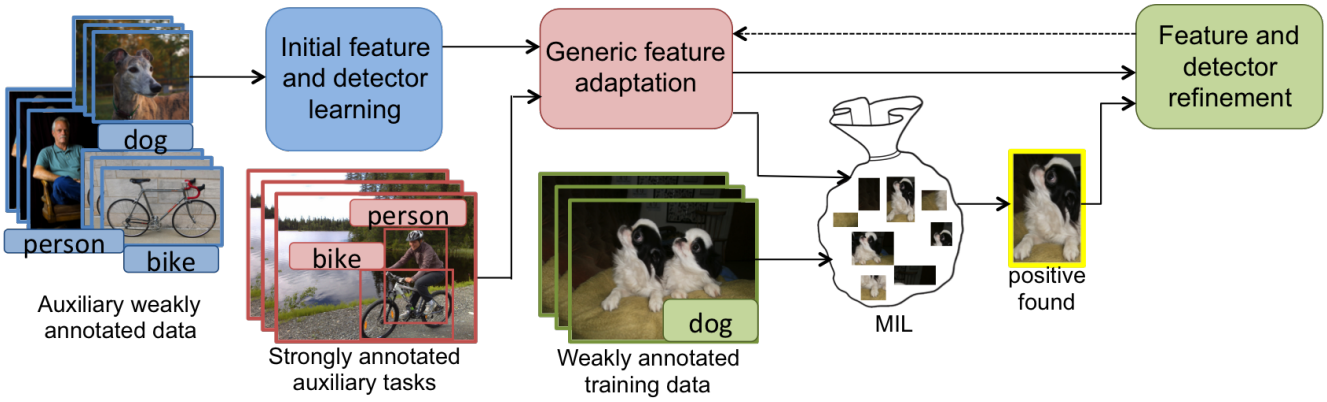

Figure 2: Our method jointly optimizes a representation and detectors for categories with only weakly annotated data. We first learn a feature representation conducive to MIL by initializing all parameters with classification style data. We then collectively refine the feature space with strongly annotated data from auxiliary tasks, and perform MIL in our detection feature space. The discovered positive patches are further used to refine the representation and detection weights.

For an image with a negative image-level label, $Y_I = -1$, we label all regions in the image as negative. For an image with a positive image-level label, $Y_I = 1$, we create a constraint that at least one positive instance occurs in the image bag.

In a typical detection scenario, R_I corresponds to the set of possible bounding boxes inside the image, and maximizing over R_I is equivalent to discovering the bounding box that contains the positive object. We define a representation $\phi(\boldsymbol{x}_i) \in \mathcal{R}^d$ for each instance, which is the feature descriptor for the corresponding bounding box, and formulate the MI-SVM objective as follows:

$$\min_{\boldsymbol{w} \in \mathcal{R}^d} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \alpha \sum_I \ell\Big(Y_I, \max_{i \in R_I} \boldsymbol{w}^T \phi(\boldsymbol{x}_i)\Big) \qquad (1)$$

where α is a hyper-parameter and $\ell(y, \hat{y})$ is the hinge loss. Interestingly, for negative bags i.e. $Y_I = -1$, the knowledge that all instances are negative allows us to unfold the max operation into a sum over each instance. Thus, Equation (1) reduces to a standard QP with respect to $\boldsymbol{w}$. For the case of positive bags, this formulation reduces to a standard SVM if maximum scoring instance is known.

Based on this idea, Equation (1) is optimized using a classic concave-convex procedure [39], which decreases the objective value monotonically with a guarantee to converge to a local minima or saddle point. Due to this reason, weakly trained MIL detectors are sensitive to the feature representation and initial detector weights (i.e. initialization in MIL)[8, 33]. With our algorithm we mitigate these sensitivities by learning a representation that works well for detection and by proposing an initialization technique for the weakly trained detectors which proves to avoid many of the pitfalls of prior MIL techniques (see Fig 3).

4. Large Scale Detection Learning

We propose a detection learning algorithm that uses a heterogeneous data source, containing only weak labels for some tasks, to produce strong detectors for all. Let the set of images with only weak labels be denoted as $\mathcal{W}$ and the set of images with strong labels (bounding box annotations) from auxiliary tasks be denoted as $\mathcal{S}$. We assume that the set of object categories that appear in the weakly labeled set, $\mathcal{C}_\mathcal{W}$, do not overlap with the set of object categories that appear in the strongly labeled set, $\mathcal{C}_\mathcal{S}$. For each image in the weakly labeled set, $I \in \mathcal{W}$, we have an image-level label per category, k: $Y_I^k \in \{1, -1\}$. For each image in the strongly labeled set, $I \in \mathcal{S}$, we have a label per category, k, per region in the image, $i \in R_I$: $y_i^k \in \{1, -1\}$. We seek to learn a representation, $\phi(\cdot)$ that can be used to train detectors for all object categories, $\mathcal{C} = \{\mathcal{C}_\mathcal{W} \cup \mathcal{C}_\mathcal{S}\}$. For a category $k \in \mathcal{C}$, we denote the category specific detection parameter as $\boldsymbol{w}_k$ and compute our final detection scores per region, $\boldsymbol{x}$, as $score_k(\boldsymbol{x}) = \boldsymbol{w}_k^T \phi(\boldsymbol{x})$.

We propose a joint optimization algorithm which learns a feature representation, $\phi(\cdot)$, and detectors, w_k, using the combination of strongly labeled detection data, $\mathcal{S}$, with weakly labeled data, $\mathcal{W}$. For a fixed representation, one can directly train detectors for all categories represented in the strongly labeled set, $k \in C_\mathcal{S}$. Additionally, for the same fixed representation, we reviewed in the previous section techniques to train detectors for the categories in the weakly labeled data set, $k \in \mathcal{C}_\mathcal{W}$. Our insight is that the knowledge from the strong label set can be used to help guide the optimization for the weak labeled set, and we can explicitly adapt our representation for the categories of interest and for the generic detection task.

Below, we state our overall objective:

$$\min_{\substack{\boldsymbol{w}_k,\phi \\ k\in\mathcal{C}}} \quad \sum_k \Gamma(\boldsymbol{w}_k) \qquad (2)$$
$$+\alpha \sum_{I\in\mathcal{W}} \sum_{p\in\mathcal{C}_\mathcal{W}} \ell(Y_I^p, \max_{i\in R_I} \boldsymbol{w}_p^T \phi(\boldsymbol{x}_i))$$
$$+\alpha \sum_{I\in\mathcal{S}} \sum_{i\in R_I} \sum_{q\in\mathcal{C}_\mathcal{S}} \ell(y_i^q, \boldsymbol{w}_q^T \phi(\boldsymbol{x}_i))$$

where α is a scalar hyper-parameter, $\ell(.)$ is the loss function and $\Gamma(.)$ is a regularization over the detector weights. This formulation is non-convex in nature due to the presence of instance level ambiguity. It is difficult to optimize directly, so we choose a specific alternating minimization approach (see Figure 2).

We begin by initializing a feature representation and initial CNN classification weights using auxiliary weakly labeled data (blue boxes Figure 2). These weights can be used to compute scores per region proposal to produce initial detection scores. We next use available strongly annotated data from auxiliary tasks to transfer category invariant information about the detection problem. We accomplish this through further optimizing our feature representation and learning a generic background detection weights (red boxes Figure 2). We then use the well tuned detection feature space to perform MIL on our weakly labeled data to find positive instances (yellow box Figure 2. Finally, we use our discovered positive instances together with the strongly annotated data from auxiliary tasks to jointly optimize all parameters corresponding to feature representation and detection weights.

4.1. Initialize Feature Representation and Detector Weights

We now discuss our procedure for initializing the feature representation and detection weights. We want to use a representation which makes it possible to separate objects of interest from background and makes it easy to distinguish different object categories. Convolutional neural networks (CNNs) have proved effective at providing the desired semantically discriminative feature representation [12, 16, 29]. We use the architecture which won the ILSVRC2012 classification challenge [22], since it is one of the best performing and most studied models. The network contains roughly 60 million parameters, and so must be pretrained on a large labeled corpus. Following the standard protocol, we use auxiliary weakly labeled data that was collected for training a classification task for this initial training of the network parameters (Figure 2: *blue boxes*). This data is usually object centric and is therefore effective for training a network that is able to discriminate between different categories. We remove the classification layer of the network and use the output of the fully connected layer, fc_7, as our initial feature representation, $\phi(\cdot)$.

We next learn initial values for all of the detection parameters, $\boldsymbol{w}_k,\ \forall k \in \mathcal{C}$. To solve this, we begin by solving the simplified learning problem of image-level classification. The image, $I \in \mathcal{S}$, is labeled as positive for a category k if any of the regions in the image are labeled as positive for k and is labeled as negative otherwise, we denote the image level label as in the weakly labeled case: Y_I^k. Now, we can optimize over all images to refine the representation and learn category specific parameters that can be used per region proposal to produce detection scores:

$$\min_{\substack{\boldsymbol{w}_k,\phi \\ k\in\mathcal{C}}} \sum_k \left[\Gamma(\boldsymbol{w}_k) + \alpha \sum_{I\in\{\mathcal{W}\cup\mathcal{S}\}} \ell(Y_I^k, \boldsymbol{w}_k^T \phi(I)) \right] \qquad (3)$$

We optimize Equation 3 through fine-tuning our CNN architecture with a new K-way last fully connected layer, where $K = |\mathcal{C}|$.

4.2. Optimize with Strong Labels From Auxiliary Tasks

Following the recent representation transfer result of Hoffman et al. [19] - LSDA, we learn to generically transform our classification feature representation into a detection representation by using the strongly labeled detection data to modify the representation (layers 1-7), $\phi(\cdot)$, through finetuning, as well as initializing a new "background" detector, $\boldsymbol{w}_b$. This detector explicitly attempts to recognize all data labeled as negative in our bags. Since we initialize this detector with the strongly annotated data, we know precisely which regions correspond to background. The intermediate objective is:

$$\min_{\substack{\boldsymbol{w}_q,\phi \\ q\in\{\mathcal{C}_\mathcal{S},b\}}} \sum_q \left[\Gamma(\boldsymbol{w}_q) + \alpha \sum_{I\in\mathcal{S}} \sum_{i\in R_I} \ell(y_i^q, \boldsymbol{w}_q^T \phi(\boldsymbol{x}_i)) \right] \qquad (4)$$

This is accomplished by fine-tuning our CNN architecture with the strongly labeled data, while keeping the detection weights for the categories with only weakly labeled data fixed. Note, we do not include the last layer adaptation part of LSDA [19], since it would not be easy to include in the joint optimization and was shown not to contribute significantly to performance. Instead, we propose using weakly supervised data explicitly, while learning to transform the classifier weights into detector weights. In addition, we use the weakly supervised data, along with the strongly supervised data, to jointly learn a fine-tuned representation that works well for detection all categories of interest at test time.

4.3. Jointly Optimize using All Data

With a representation that has now been directly tuned for detection, we fix the representation, $\phi(\cdot)$ and consider

solving for the regions of interest in each weak labeled image. This corresponds to solving the second term in Equation (2), i.e.:

$$\min_{\substack{\boldsymbol{w}_p \\ p\in\{\mathcal{C}_\mathcal{W},b\}}} \sum_p \Bigg[\Gamma(\boldsymbol{w}_p) + \alpha \sum_{I\in\mathcal{W}} \ell(Y_I^p, \max_{i\in R_I} \boldsymbol{w}_p^T \phi(\boldsymbol{x}_i))\Bigg] \tag{5}$$

Note, we can decouple this optimization problem and independently solve for each category in our weakly labeled data set, $p \in \mathcal{C}_\mathcal{W}$. Let's consider a single category p. Our goal is to minimize the loss for category p over images $I \in \mathcal{W}$. We will do this by considering two cases. First, if p is not in the weak label set of an image ($Y_I^p = -1$), then all regions in that image should be considered negative for category p. Second, if $Y_I^p = 1$, then we positively label a region $\boldsymbol{x}_i$ if it has the highest confidence of containing object and negatively label all other regions. We perform the discovery of this top region in two steps. At first, we narrow down the set of candidate bounding boxes using the score, $\boldsymbol{w}_p^T \phi(\boldsymbol{x}_i)$, from our fixed representation and detectors from the previous optimization step. This set is then refined to estimate the most region likely to contain the positive instance in a Latent SVM formulation. The implementation details are discussed section 5.2.

Our final optimization step is to use the discovered annotations from our weak data-set to refine our detectors and feature representation from the previous optimization step. This amounts to the subsequent step for alternating minimization of the joint objective described in Equation 2. We collectively utilize the strong annotations of images in $\mathcal{S}$ and estimated annotations for weakly labelled set, $\mathcal{W}$, to optimize for detector weights and feature representation, as follows:

$$\min_{\substack{\boldsymbol{w}_k,\phi \\ k\in\{\mathcal{C},b\}}} \sum_k \Bigg[\Gamma(\boldsymbol{w}_k) + \alpha \sum_{I\in\{\mathcal{W}\cup\mathcal{S}\}} \sum_{i\in R_I} \ell(y_i^k, \boldsymbol{w}_k^T \phi(\boldsymbol{x}_i))\Bigg] \tag{6}$$

This is achieved by re-finetuning the CNN architecture. Thus, the overall non-convex objective (2) is approximated through initialization in (3) and alternating optimization in (4)-(5).

The refined detector weights and representation can be used to mine the bounding box annotations for weakly labeled data again, and this process can be iterated over (see Figure 2). We discuss re-training strategies and evaluate the contribution of this final optimization step in Section 5.3.

Train	Num images	395905
	Num objects	345854
Val	Num images	20121
	Num objects	55502

Table 1: Statistics of the ILSVRC13 detection dataset. Training set has fewer objects per image than validation set.

5. Experiments

We now study the effectiveness of our algorithm by applying it to a standard detection task.

5.1. ILSVRC13 Detection Dataset & Setup

We use the ILSVRC13 detection dataset [27] for our experiments. This dataset provides bounding box annotations for 200 categories. The dataset is separated into three pieces: train, val, test (see Table 1). The training images have fewer objects per image on an average than validation set images, so they constitute classification style data [19]. Following prior work [16], we use the further separation of the validation set into val1 and val2. Overall, we use the train and val1 set for our training data source and evaluate our performance of the data in val2.

Specifically, we use $\sim$1000 randomly chosen images per class from the train set for initializing our CNN weights. For this data we consider only have weak labels for all categories and train with the classification objective. We use the train set for this purpose as it tends to have more object-centric images and is therefore better suited to initializing classification weights.

We have bounding box annotations for 100/200 of the categories in val1 ($\sim$5000 images with bounding boxes). Specifically, with the category names sorted alphabetically, categories 1-100 have strong annotations while 101-200 have only weak (image-level) annotations. Finally, we evaluate detection performance on the $\sim 10,000$ images in val2 across all 200 categories.

We use open source deep learning framework, Caffe [21], for the implementation, training and finetuning of our CNN architecture. For all experiments, the network is pre-trained using the ILSVRC2012 classification challenge dataset.

5.2. Analysis of Discovered Positive Boxes

One of the key components of our system is using strong annotations from auxiliary tasks to learn a representation where it's possible to discover patches that correspond to the objects of interest in our weakly labeled data source. We begin our analysis by studying the patch discovery that our feature space enables. We begin by using selective search [35] to produce candidate regions. We optimize the

Figure 3: 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 corresponding category. Examples with a green outline are categories for which our algorithm was able to correctly mine patches of the object, while the feature space with only weak label training was not able to produce correct patches. In yellow we highlight the specific example of "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.

patch discovery (Equation (5)) using a one vs all Latent SVM formulation and optimize the formulation for AUC criterion [6]. The feature descriptor used is the output of the fully connected layer, fc_7, of the CNN which is produced after fine-tuning the feature representation with strongly annotated data from auxiliary tasks. Following our alternating minimization approach, these discovered top boxes are then used to re-estimate the weights and feature representations of our CNN architecture.

To evaluate the quality of mined boxes, we do precision analysis with respect to their overlap with ground truth which is measured using the standard intersection over union (IOU) metric. Table 2 reports the precision for varying overlapping thresholds. Our optimization approach produces one positive patch per image with a weak label, and a discovered patch is considered a true positive if it overlaps sufficiently with the ground truth box that corresponds to that label. Since each patch, once discovered, is considered an equivalent positive (regardless of score) for the purpose of retraining, this simple precision metric is a good indication of the usefulness of our mined patches. It is interesting that a significant fraction of mined boxes have high overlap with the ground truth regions. For reference, we also computed the standard mean average precision over the discovered patches and report these results.

It is important to understand not only that our new feature space improves the quality of the resulting patches, but also what type of errors our method reduces. In Figure 3, we show the top 5 scoring discovered patches before and after modifying the feature space with strong annotations from auxiliary tasks. We find that in many cases the improvement comes from better localization. For example without auxiliary strong annotations we mostly discover the face of a lion rather than the body that we discover after our algorithm. Interestingly, there is also an issue with co-occurring classes. In the bottom row of Figure 3, we show the top 5 discovered patches for "tennis racket". Once we incorporate strong annotations from auxiliary tasks we begin to be able to distinguish the person playing tennis from the racket itself. Finally, there are some example mined patches

	Precision				mAP
	ov=0.3	ov=0.5	ov=0.7	ov=0.9	ov=0.5
Without auxiliary strong dataset	29.63	26.10	24.28	23.43	13.13
Ours	32.69	28.81	26.27	24.78	22.81

Table 2: Precision analysis and mAP performance of discovered patches in our weakly labeled training set (val1) of ILSVRC13 detection dataset. Comparison with varying amount of overlap with ground truth box. About 25% of our mined boxes have an overlap of at least 0.9. Our method is able to significantly improve the quality of mined boxes after incorporating strong annotations from auxiliary tasks.

Figure 4: Example mined boxes of the category "laptop" where using auxiliary strongly annotated data causes patch discovery to diverge. *Top row*: The mined boxes obtained after fine-tuning with images in classification settings only. *Bottom row*: The mined boxes obtained after fine-tuning with the auxiliary strongly annotated dataset that contains the category "computer keyboard". These patches were low scoring examples, but we show them here to demonstrate a potential failure case – specifically, when one of the strongly annotated classes is a part of one of the weakly labeled classes.

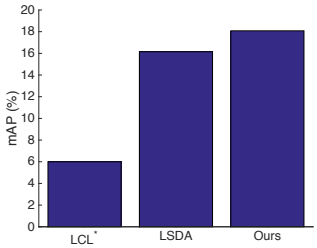

Figure 5: Comparison of mAP (%) for categories without any bounding box annotations (101-200 of val2) of ILSVRC13. Our method significantly outperforms both previous state-of-the-art algorithms: LCL [36] and LSDA [19]. *The value for LCL was computed across all 200 categories. Our experiments show this is an easier task resulting in higher numbers overall.

where we reduce quality after incorporating the strong annotations from auxiliary tasks. For example, one of our strongly annotated categories is "computer keyboard". Due to the strong training with keyboard images, some of our mined patches for "laptop" start to have higher scores on the keyboard rather than the whole laptop (see Figure 4).

5.3. Detection Performance

Now that we've analyzed the intermediate result of our algorithm, we next study the full performance of our system. Figure 5 shows the mean average precision (mAP) percentage computed over the categories in val2 of ILSVRC13 for which we only have weakly annotated training data (categories 101-200). We compare to two state-of-the-art methods for this scenario and show that our algorithm significantly outperforms both of the previous state-of-the-art techniques. The first, LCL [36], detects in the standard weakly supervised setting – having no bounding box annotations for any of the 200 categories. This method also only reports results across all 200 categories. Our experiments indicate that the first 100 categories are easier on average then the second 100 categories, therefore the 6.0% mAP may actually be an upper bound of the performance of this approach. The second algorithm we compare against is LSDA [19], which does utilize the bounding box information from the first 100 categories.

We next consider different re-training strategies for learning new features and detection weights after discovering the positive patches in the weakly labeled data. Table 3 reports the mean average precision (mAP) percentage for no re-training (directly using the feature space learned after incorporating the strong labels), re-training only the category detection parameters, and retraining feature representations jointly with detection weights. In our experiments the improved performance is due to the first iteration of the overall algorithm. We find that the best approach is to jointly learn to refine the feature representation and the detection weights. More specifically, we learn a new feature representation by fine-tuning all fully connected layers in the CNN architecture.

We finally analyze examples where our full algorithm outperforms the previous state-of-the-art, LSDA [19]. Figure 6 shows a sample of the types of errors our algorithm improves on. These include localization errors, confusion with other categories, and interestingly, confusion with co-

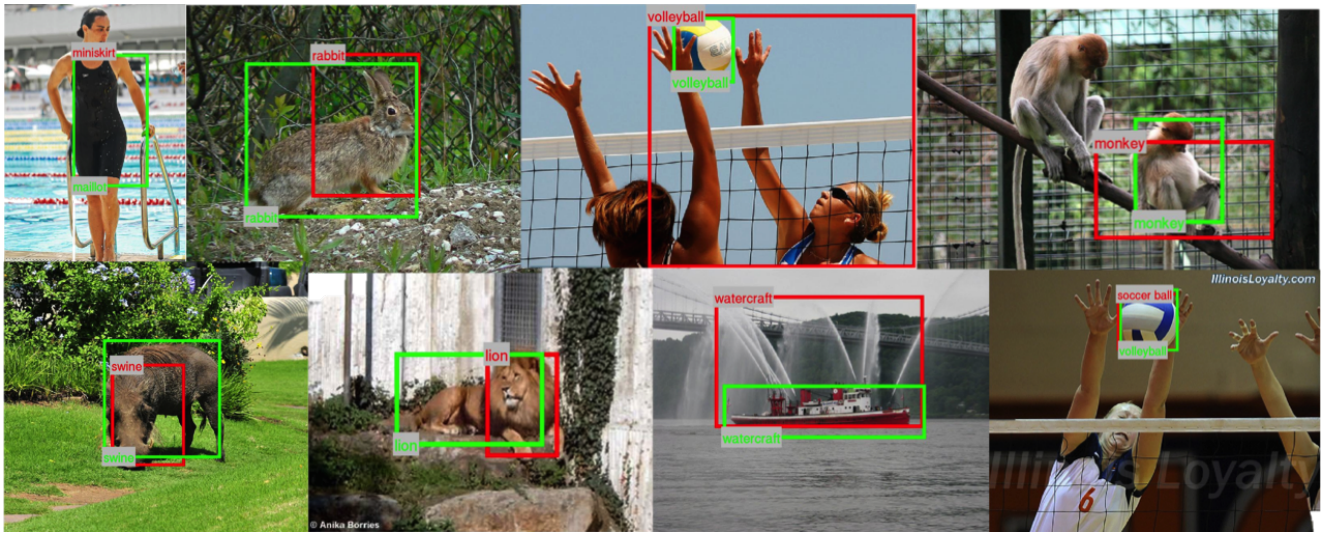

Figure 6: Examples where our algorithm outperforms the previous state-of-the-art. We show the top scoring detection from the baseline detector, LSDA [19], with a Red box and label, and the top scoring detection from our method, LSDL, as a Green box and label. Our algorithm improves localization (ex: rabbit, lion etc), confusion with other categories (ex: miniskirt vs maillot), and confusion with co-occurring classes (ex: volleyball vs volleyball player)

	Re-train Strategy	Category Set: Weakly Labeled	Category Set: Strongly Labeled	All
Ours	-	15.85	27.81	21.83
	detectors	17.01	27.85	22.43
	rep+detectors	**18.08**	27.40	22.74

Table 3: Comparison of different ways to re-train after discovery of positive patches. We show mAP on val2 set from ILSVRC13. We find that the most effective way to re-train with discovered windows is to modify the detectors and the feature representation.

occurring categories. In particular, our algorithm provides improvement when searching for a small object (ball or helmet) in a sports scene. Training only with weak labels causes the previous state-of-the-art to confuse the player and the object, resulting in a detection that includes both. Our algorithm is able to localize only the small object and recognize that the player is a separate object of interest.

6. Conclusion

We have presented a method which jointly trains a feature representation and detectors for categories with only weakly labeled data. We use the insight that strongly annotated detection data from auxiliary tasks can be used to train a feature representation that is conducive to discovering object patches in weakly labeled data. We demonstrate using a standard detection dataset (ImageNet-200 detection) that our method of incorporating the strongly annotated data from auxiliary tasks is very effective at improving the quality of the discovered patches. We then use all strong annotations along with our discovered object patches to further refine our feature representation and produce our final detectors. We show that our full detection algorithm significantly outperforms both the previous state-of-the-art methods which uses only weakly annotated data, as well as the algorithm which uses strongly annotated data from auxiliary tasks, but does not incorporate any MIL for the weak tasks.

Acknowledgements This work was supported in part by DARPA's MSEE and SMISC programs, NSF awards IIS-1427425, IIS-1212798, and IIS-1451244, and the Berkeley Vision and Learning Center.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *Proc. CVPR*, 2010.

[2] K. Ali and K. Saenko. Confidence-rated multiple instance boosting for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[3] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Proc. NIPS*, pages 561–568, 2002.

[4] Y. Aytar and A. Zisserman. Tabula rasa: Model transfer for object category detection. In *IEEE International Conference on Computer Vision*, 2011.

[5] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *In Proc. ICML*, 2009.

[6] H. Bilen, V. P. Namboodiri, and L. J. Van Gool. Object and action classification with latent window parameters. *IJCV*, 106(3):237–251, 2014.

[7] D. Borth, R. Ji, T. Chen, T. Breuel, and S. F. Chang. Large-scale visual sentiment ontology and detectors using adjective nown paiars. In *ACM Multimedia Conference*, 2013.

[8] R. G. Cinbis, J. Verbeek, C. Schmid, et al. Multi-fold mil training for weakly supervised object localization. In *CVPR*, 2014.

[9] H. Daumé III. Frustratingly easy domain adaptation. In *ACL*, 2007.

[10] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *IJCV*, 2012.

[11] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 1997.

[12] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *Proc. ICML*, 2014.

[13] L. Duan, D. Xu, and I. W. Tsang. Learning with augmented features for heterogeneous domain adaptation. In *Proc. ICML*, 2012.

[14] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Tran. PAMI*, 32(9):1627–1645, 2010.

[15] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proc. ICCV*, 2013.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *In Proc. CVPR*, 2014.

[17] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Proc. CVPR*, 2012.

[18] D. Heckerman. A tractable inference algorithm for diagnosing multiple diseases. *arXiv preprint arXiv:1304.1511*, 2013.

[19] J. Hoffman, S. Guadarrama, E. Tzeng, R. Hu, J. Donahue, R. Girshick, T. Darrell, and K. Saenko. LSDA: Large scale detection through adaptation. In *Neural Information Processing Systems (NIPS)*, 2014.

[20] J. Hoffman, E. Rodner, J. Donahue, K. Saenko, and T. Darrell. Efficient learning of domain-invariant image representations. In *Proc. ICLR*, 2013.

[21] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012.

[23] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Proc. CVPR*, 2011.

[24] M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *In Proc. NIPS*, 2010.

[25] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1989.

[26] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *Proc. ICCV*, 2011.

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. K. amd Michael Bernstein, A. C. Berg, and L. Fei-Fe. Imagenet large scale visual recognition challenge. arXiv:1409.0575, 2014.

[28] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *Proc. ECCV*, 2010.

[29] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013.

[30] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*. 2012.

[31] P. Siva, C. Russell, and T. Xiang. In defence of negative mining for annotating weakly labelled data. In *ECCV*. 2012.

[32] P. Siva, C. Russell, T. Xiang, and L. Agapito. Looking beyond the image: Unsupervised learning for object saliency and detection. In *Proc. CVPR*, 2013.

[33] H. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell. On learning to localize objects with minimal supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

[34] H. O. Song, Y. J. Lee, S. Jegelka, and T. Darrell. Weakly-supervised discovery of visual pattern configurations. 2014.

[35] J. U. K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *IJCV*, 2013.

[36] C. Wang, W. Ren, K. Huang, and T. Tan. Weakly supervised object localization with latet category learning. In *European Conference on Computer Vision (ECCV)*, 2014.

[37] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive svms. *ACM Multimedia*, 2007.

[38] C.-N. J. Yu and T. Joachims. Learning structural svms with latent variables. In *Proc. ICML*, pages 1169–1176, 2009.

[39] A. L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4):915–936, 2003.

[40] M. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. *ArXiv e-prints*, 2013.

[41] C. Zhang, J. C. Platt, and P. A. Viola. Multiple instance boosting for object detection. In *Advances in neural information processing systems*, 2005.