Self-Supervised Object Detection and Retrieval Using Unlabeled Videos

Elad Amrani[1,2], Rami Ben-Ari[1], Inbar Shapira[1], Tal Hakim[1], Alex Bronstein[2]
[1]IBM Research AI
[2]Technion
elad.amrani@ibm.com
{ramib, inbar_shapira, thakim}@il.ibm.com
bron@cs.technion.ac.il

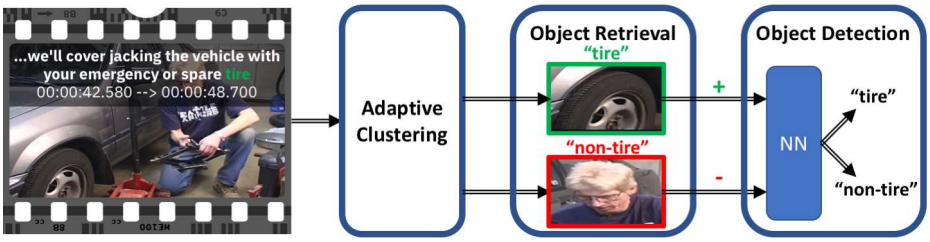


Figure 1: **Self-supervised object detection and retrieval**. Using unlabeled videos, and based on the natural correlation of speech and vision, our model learns the appearance **and** names of objects without any manual labeling involved. NN: Neural-Network

Abstract

Learning an object detection or retrieval system requires a large data set with manual annotations. Such data are expensive and time-consuming to create and therefore difficult to obtain on a large scale. In this work, we propose using the natural correlation in narrations and the visual presence of objects in video to learn an object detector and retriever without any manual labeling involved. We pose the problem as weakly supervised learning with noisy labels, and propose a novel object detection and retrieval paradigm under these constraints. We handle the background rejection by using contrastive samples and confront the high level of label noise with a new clustering score. Our evaluation is based on a set of ten objects with manual ground truth annotation in almost 5000 *frames extracted from instructional videos from the web. We demonstrate superior results compared to state-of-the-art weakly-supervised approaches and report a strongly-labeled upper bound as well. While the focus of the paper is object detection and retrieval, the proposed methodology can be applied to a broader range of noisy weakly-supervised problems.*

1. Introduction

Existing machine learning techniques still lag far behind human ability to learn from minimal supervision, and often require a tremendous amount of labeled data. Although huge progress has been made in the field, with the recent weakly-supervised training methods [3, 8, 23, 30, 25] coming close to results achieved in fully-supervised approaches [17, 10], the size, quality, and availability of labeled data are currently becoming a major bottleneck. One possible approach to break past this limitation is the self-supervised learning paradigm. Examples of self-supervised learning include reinforcement [13] and pretext-based learning such as Jigsaw [15] and Colorization [28]. Yet these methods require a reward function, or introduce rare tasks.

The huge and increasing amount of online videos brings several opportunities for training deep neural networks in the self-supervised regime. Large-scale video data sets such as the YouTube-8M [1] and the How2 data set [16] can be leveraged for this purpose. In this paper, we explore a method for self-supervised learning, tackling the challenging tasks of visual object detection and retrieval. To this end, we exploit the How2 data set by taking advantage of the multi-modal information it provides (video and automatic closed captions). Fig. 1 describes the targeted problem in this paper.

Given unlabeled training videos, the audio channel can be used as a "free" source of weak labels, allowing a convolutional network to learn objects and scenes. For instance, by seeing and hearing many frames where the word "guitar" is mentioned, it should be possible to detect the guitar due to its shared characteristics over different frames. Yet, self-supervised learning from the videos themselves is quite hard when performed in the wild, as the audio and the visual contents may often appear completely unrelated. Nevertheless, the results in Section 4 show that our approach is able to fairly successfully reduce the level of this source of noise, detecting frames that contain a desired object, and localizing the objects in the relevant frames – all this in hard scenarios of large variation in object appearance, typical motion blur in video frames and in the presence of strong label noise.

We pose the self-supervised object detection problem as a noisy, weakly-labeled binary learning task. To ground a certain category of object, we consider the large corpus of How2 [16] instructional videos covering a wide variety of topics across 13,000 clips (about 300 hours total duration), with word-level time-aligned subtitles. We extract candidate frames from time intervals corresponding to the subtitle, containing the object's name (synced with the speech mentioning the name of the object). This set of frames comprises our positive set, yet it is noisy labeled as the object might not appear in all of the selected frames. Creating a negative set (that most likely lacks the object of interest) allows discriminating between the object and background, essentially solving the object detection problem. The task is similar to weakly supervised object detection (WSOD), yet is distinct from it in two ways: 1) the high level of label noise and 2) its binary nature (i.e., each model is trained to detect a single object). We show that the performance of state-of-the-art weakly supervised models is degraded when handling noisy labeled data. Furthermore, most of the time these models fail to converge at all, when trained for single-class detection.

We argue that a single-class detection model is essential for a scalable and robust multi-class object detection system for a number of reasons. Firstly, adding a new object does not require training from scratch on the entire data set. Secondly, a multi-class detection model of a very large set of objects (e.g., tens of thousands) is hard to train (convergence duration, data size, etc.). Thirdly, a single object detector should not require multi-class training data as is the case for WSOD models (cf. Table 2). Lastly, multiple single-class detectors applied in parallel act as an ensemble of models and are capable of obtaining better predictive performance than could be obtained from a single multi-class detector (Section 5.2).

Contribution. The key contribution of this paper is threefold:

- We introduce a methodology capable of detecting and retrieving objects **and** their names, learned from continuous videos without any manual labeling.
- We pose the object detection and retrieval problem as weakly-labeled binary learning task, but with noisy labels, and propose a novel noise robust model for accomplishing it.
- Our model incorporates a novel cluster scoring method that distills the detected regions separating them from the surrounding clutter, even with extreme weak-label noise and only binary weakly-labeled data.

2. Related Work

2.1. Self-Supervised multimodal learning

These approaches target learning from a large number of unlabeled videos [2, 4, 9, 18, 20, 27, 29, 21] capitalizing on the natural synchronization of the visual and audio modalities, to learn models that discover audio-visual correspondence. Particularly, [2, 9, 29, 21] use video and sound (not speech) to discover the relevant audio track for a certain region in a frame. In [2] the authors also suggest an object localization (in both modalities), yet by activation heat maps (in low resolution) and not as detection. The computer vision and NLP communities have begun to leverage deep learning to create multimodal models of images and text. Grounded language learning from video has been studied in [18, 20, 27] by learning joint representations of text sentences and visual cues from videos. For instance, in [27] words and objects are related using a trained object detector from an external source, and therefore is not self-supervised. Sun et al. [20] target a different problem of speech recognition and [18] studies correspondence between words and concepts in human actions. The work in [14] presents an unsupervised alignment method for natural language instructions in videos, with the specific goal of automatically align the video segments to the corresponding protocol sentences, and track hands in video and detect the blobs touched. Recently, [19, 12] proposed a self-supervised joint visual-linguistic model to learn high-level features from instructional videos. The closet work to our study is [7] that attempts to map word-like acoustic units in the continuous speech, to semantically relevant regions in the image. However this method is not self-supervised as the captions are manually created for the image samples.

2.2. Weakly-supervised object detection

Weakly supervised learning and particularly of object detection has attracted high interest as it's potential to reduce the annotation labour, involved in creating the weak

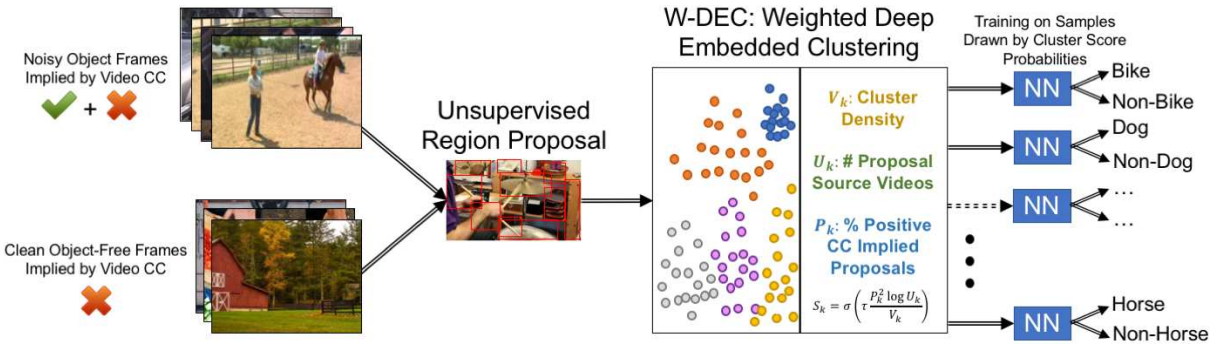


Figure 2: Schematic flow of the proposed self-supervised object detection or retrieval training scheme. We collect key frames based on speech-to-text transcript. A frame is labeled with object X if the noun X was mentioned by a speaker. For every object we also collect non-object frames from videos the noun X was not mentioned in. We cluster the data in the region proposal feature space and refine those clusters iteratively based on a newly suggested potential score function. Finally, we sample training data in a non-uniform manner based on the potential score, and thus reduce noise, to train a neural network (NN) region classifier.

labels. Particularly, weakly supervised object detection (WSOD) [3, 8, 23, 30, 25] uses only image-level annotations to train object detectors (and no bounding boxes around objects). While weakly supervised methods accept well curated and "clean" labels (correct labels for all the images in the train-set), our problem involves a "noisy" labeled data set, where many of our selected and self-labeled frames are falsely labeled. In this study we suggest a novel approach for weakly and noisily labeled object detection problem. We compare our method to two different state-of-the-art weakly supervised object detection models [23, 30], showing superior results for both.

2.3. Noisy labels

Many noisy labeled methods target the training of deep neural networks on large-scale weakly-supervised web images, which are crawled from the internet by using text queries, without any human annotation [5, 31]. Although our selected frames are expected to have fairly clean labels by construction, they still contain considerable noise, i.e. many extracted frames lack the object in the scene. Deep learning with noisy labels is practically challenging, as the capacity of deep models is so high that they can totally overfit the data with the noise [5, 6, 31]. Handling extreme noise levels for image *classification* has been shown in [6], (tested on MNIST and CIFAR) for up to 50% noise level. However, our problem addresses a more challenging problem of object detection with *weak* labels, that furthermore involves higher noise level of up to 68% .

3. Method

The proposed self-supervised scheme is described in Fig. 2 with the pipeline summarized in Algorithm 1. Our input

Algorithm 1: Self-Supervised Object Detection

Input: Unlabeled videos $\mathcal{X}$
Output: Trained object detector $\mathcal{D}$ for object $\mathcal{O}$

Extract transcript for all videos in $\mathcal{X}$. Section 3.1;
Extract positive & negative frames for $\mathcal{O}$. Section 3.1;
Extract N region proposals per frame. Section 3.1;
Compute feature representation. Section 3.1;
for *idx in Range(0,* MAX_EPOCHS*):* **do**
    /* W-DEC */
    if $idx == 0$ **then**
        Initialize cluster centers using uniform K-Means;
    else if $idx \;\%\% \; I == 0$ **then**
        Re-initialize cluster centers using Weighed K-Means and S_k (1). Section 3.3;
    Train clustering net for one epoch. Section 3.3;
    /* detection network */
    Compute S_k (1) per cluster. Section 3.2;
    Run DSD per frame per cluster. Section 3.4;
    Train region classifier for one epoch. Section 3.5;

comprises a large set of unlabeled videos with speech transcription from instructional video corpus of How2 [16]. In the following, we detail the key steps of our method.

3.1. Extraction of key frames

For a given object name, let's say a guitar, we extract a single key frame from each center of temporal period where the object was mentioned. While this is not ideal, it works fairly well for selecting frames that contain the object. We now dub these selected images as our (**noisy**) *positive* set, labeled as $Y_l = 1$. We construct also a balanced *negative* set, $Y_l = 0$, containing frames randomly selected from dis-

	Bike	Dog	Drum	Guitar	Gun	Horse	Pan	Plate	Scissors	Tire	Total/Average
No. of frames	419	1243	457	442	182	597	351	341	200	305	4537
Obj. instances	715	551	312	564	111	327	154	147	92	526	3499
Noise level %	28.4	59.5	53.4	35.1	61.5	47.1	68.4	67.7	58.5	22.6	55.4
SS recall %	85	97	98	94	95	99	94	97	83	83	92.5

Table 1: Data set: Statistics of the selected frames. *Noise level* refers to the percentage of selected frames without the object present and *SS recall* to Selective Search [24] recall at $IoU \geq 0.5$.

	Bike	Dog	Drum	Guitar	Gun	Horse	Pan	Plate	Scissors	Tire	mAP
Supervised (single-class)	42.7	60.0	51.8	55.4	49.0	73.2	43.8	38.7	32.0	48.6	49.5
PCL (single-class) [23]	F	3.2	F	6.6	0.2	11.3	10.9	**13.7**	0.6	11.1	7.2*
SPN (single-class) [30]	2.5	F	0.0	0.0	F	F	F	F	F	0.0	0.8*
Ours (single-class)	**24.4**	**14.3**	**23.3**	**18.3**	**9.6**	**27.8**	**20.4**	12.0	**8.2**	**18.3**	**17.7**
SPN (multi-class) [30]	13.6	12.1	**29.8**	12.6	0.7	22.9	6.1	7.3	0.0	13.6	11.9
Ours (multi-class)	**22.8**	**14.1**	19.7	**17.5**	**9.4**	**27.5**	**17.4**	**8.3**	**6.8**	**16.1**	**16.0**

Table 2: Evaluation on **HowTo10 test set: Average precision over 3-folds at $\mathbf{IoU} \geq \mathbf{0.5}$**. PCL [23], SPN [30] : Weakly Supervised. F: Failed to converge, *: Without the failed objects. single-class: a binary model is trained & tested separately for each object to distinguish between object and background. multi-class: standard multi-class detection. Best results comparing to Weakly Supervised are in bold.

parate videos, that the object was not mentioned in. These frames will most likely be without the object of interest, but will include elements contained in the surroundings of our object instances in the positive frames, such as faces, hands, tables, chairs, etc. For each image (positive and negative) we extract N region proposals using Selective Search [24]. Regions are labeled as positive or negative according to the corresponding frame label, $y_{li} = Y_l$ (similar to the *bag* and *instance* labels in multi-instance learning paradigm [11]). Using a pre-trained back-bone such as Inception-ResNet-v2 CNN [22], we map each candidate region to a feature space, represented by z_{li}.

3.2. Potential score

The purpose of our learning approach is to find a common theme across positive regions that is less likely to exist in negative counterparts. To this end, we cluster the regions in the embedded space. Clusters with dense population of positive regions are likely to contain the object of interest. We therefore associate a *positive ratio* score to each cluster, defined as the ratio between the positive and the total number of samples in the cluster (note that regions are labeled according to their corresponding frame). Yet, high positive-ratio clusters are noisy, so that real object clusters are not always distinguishable. Specifically, we search for a target cluster, satisfying the following properties: (1) High positive ratio ; (2) Low cluster variance, for tendency to include a single object type; and (3) Cluster members that come from a wide variety of videos, since we expect the object to have a common characteristics among various videos. The latter property also copes with the high temporal correlation in a single video, that may create dense clusters. We formalize these constraints using the following softmax function S_k, to which we refer as the *potential score*, i.e. score of cluster k containing the object:

$$S_k = \sigma\left(\tau \frac{P_k^2 \cdot \log U_k}{V_k}\right) \qquad k \in \{0..K-1\}. \quad (1)$$

Here, $\sigma(\cdot)$ is the softmax function, K denotes the total number of clusters, $\tau \in \mathbb{R}$ is the softmax temperature, P_k is the positive-ratio (according to the raw weak labels, since the ground truth labels are not accessible), V_k is the cluster distance variance, and U_k denotes the number of unique videos. All parameters are normalized to unit sum. Our observations showed the following importance order in the potential score components: positive-ratio P_k, the cluster variance V_k, and lastly the number of unique videos U_k. For this reason P_k is squared and we take the $\log$ of U_k.

3.3. Weighed Deep Embedded Clustering

Following feature extraction, we cluster our region proposals using a weighed variant of Deep Embedded Clustering (DEC) [26] we call W-DEC. The original DEC is a method that simultaneously learns feature representations and cluster assignments using deep neural networks by building a mapping from the data space to a lower-

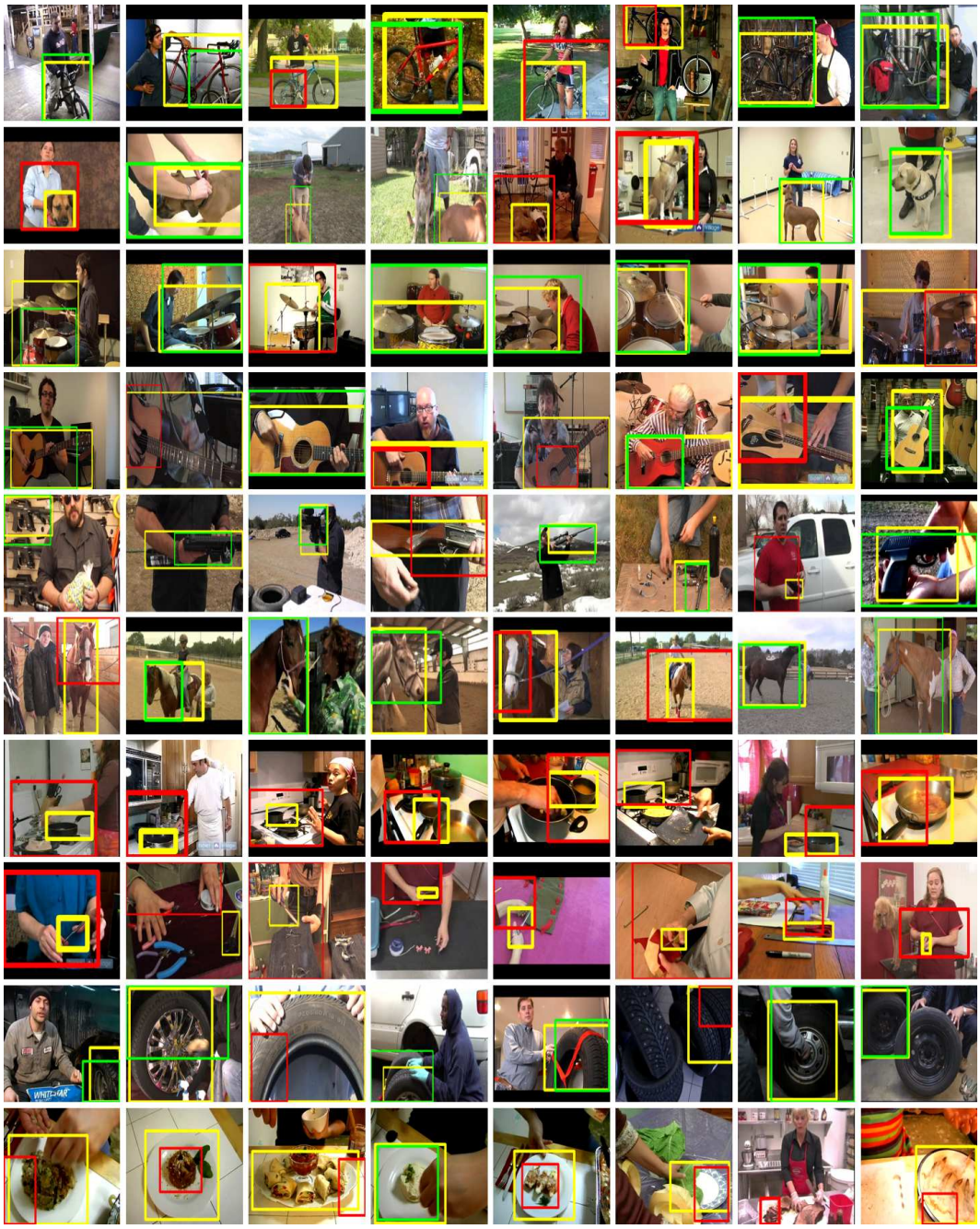


Figure 3: **Self-supervised retrieval results**. Our model is able to learn and retrieve the appearance **and** names of objects without any manual labeling involved. This is done simply by "watching and listening" to unconstrained unlabeled videos from the How2 data set [16] that includes 300 hours of instructional videos. Above, are the retrieval results of ten objects (eight top instances per object from left to right). Top to bottom: Bike, Dog, Drum, Guitar, Gun, Horse, Pan, Scissors, Tire and Plate. **Green** and **red** boxes represent success and failure, respectively, with regards to IoU=0.5. **Yellow** box represents ground truth of the predicted instance. For better object instance variety, objects from unique videos are presented.

dimensional feature space in which it iteratively optimizes a clustering objective. While DEC is a general clustering model, W-DEC drives the clustering toward true positive samples (with respect to the ground truth) without any

ground truth labels. It does so by applying the following modifications: Firstly, while DEC initializes cluster centers once using K-means, W-DEC re-initializes cluster centers using *weighed* K-means every I epochs (e.g., $I = 3$), with the weights set by the potential score function S_k (1) normalized by the number of positive samples in the cluster. This re-initialization is an important step since it breaks down candidate object clusters to smaller, yet higher quality clusters. Secondly, following [26], we further suggest the *weighed* Student's t-distribution as a similarity measure:

$$q_{i,j} = \frac{(1 + ||z_i - \mu_j||^2)^{-1} \cdot w(i,j)}{\sum_{j'} (1 + ||z_i - \mu_{j'}||^2)^{-1} \cdot w(i,j')} \quad (2)$$

with indices i and j referring to the sample and the cluster, respectively, z_i denoting the region embedding of sample i, and μ_j being the centroid of cluster j. Here and in the sequel, we drop the frame index l for simplicity. We set the weights according to the region label $y_i \in \{0, 1\}$ as

$$w(i,j) = \begin{cases} 0.5, & \text{if } y_i = 0 \\ 1, & \text{otherwise} \end{cases} \quad (3)$$

The newly added $w(i,j)$ simply weighs weakly-labeled positive samples in candidate object clusters higher. This simple weighing factor further refines those clusters while not completely ignoring the negative data set distribution. We use the new measure in (2) to drive the clustering to the target distribution $p_{ij} = \frac{q_{ij}^2 / f_j}{\sum_{j'} q_{ij'}^2 / f_{j'}}$ with $f_j = \sum_i q_{ij}$, using the Kullback-Leibler divergence loss (see [26] for more details). In practice, we apply the weighing to clusters with the positive ratio above a certain threshold (we refer to these clusters as candidate object clusters). In our implementation, only cluster centers were optimized while keeping the embeddings fixed.

3.4. Dense Subgraph Discovery

A frequent shortcoming of weakly-supervised approaches is their inability to distinguish between candidates with high and low object overlap. However, for training a high-performance object detector, regions with tight spatial coverage of the object are required, i.e., high Intersection-over-Union (IoU). To address this issue, we use the Dense Subgraph Discovery (DSD) algorithm [8] on top of W-DEC. This model defines an undirected unweighed graph for a set of region proposals in a given image. The nodes correspond to region proposals and the edges are formed by connecting each proposal (node) to its multiple neighbors, which have mutual IoU larger than a pre-defined threshold. For our use case, we found that simply extracting the top 10% of the most connected nodes works well. Unlike [8], we further make use of the remaining regions as "hard negative" examples, which we found to be beneficial.

3.5. Sampling and training of the detector

Each cluster is assigned a potential score as defined in (1). This score is likely to correlate with *cluster purity*, i.e., the ratio of regions in a cluster that contains instances of the object. We then train a detector fed by the following samples: for positive samples we consider the regions selected by W-DEC followed by DSD and sample regions with high potential score S_k. Our sampling distribution is the normalized score S_k. Note that sample scores are associated with their corresponding cluster k. This sampling strategy allows sampling from several clusters, since object regions may be distributed among multiple clusters. This sampling regime continuously reduces the noise level in the positive set, that is necessary to reach a high accuracy detector (a region classifier). Negative samples are sampled uniformly from the negative frames and are combined with the rejected regions from DSD, and are used as hard negatives. Our detector is a multilayer perceptron with three fully connected layers trained to separate between object and background, using cross-entropy loss.

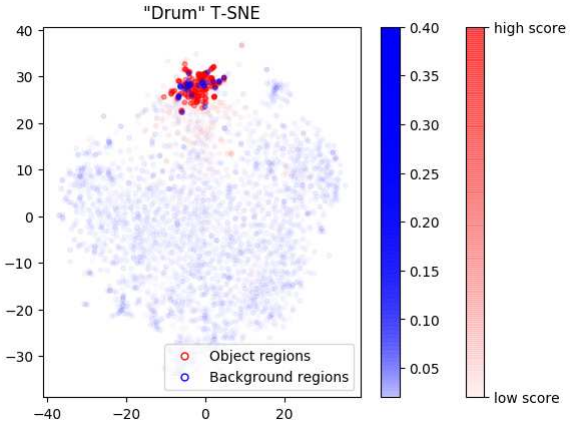
4. Experiments

4.1. Data set

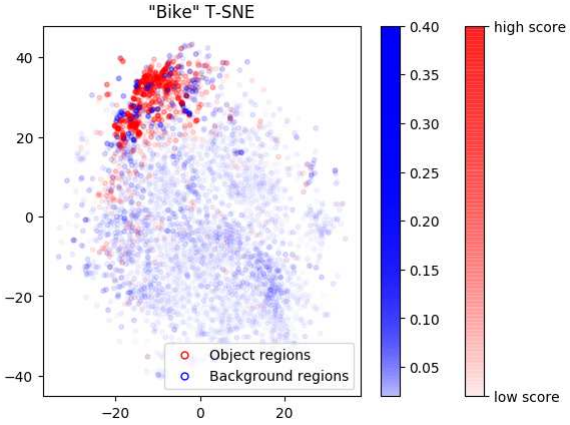
Our evaluation is based on the How2 data set [16] that includes 300 hours of instructional videos with synchronized closed captions in English. Processing caption text in the data set, we extract all object nouns and choose 10 top references to be manually annotated for testing purposes only. We call this data set (and its annotations) of 10 objects *HowTo10* and will release it publicly. We include in our corpus challenging nouns such as Bike that corresponds to both Bicycle and Motorbike, or Gun that refers to Pistol, Rifle and Glue Gun. This results a total of 4,537 frames from the videos, with an average of 453 frames per-object. Our transcript-based frame selection introduces 55% noisily labeled frames on average (i.e., only 45% of object frames are correctly labeled and contain the object of interest on average). The statistics of our data set and the recall rate for the Selected Search region proposal, are shown in Table 1.

4.2. Comparison to SoTA models and upper bound

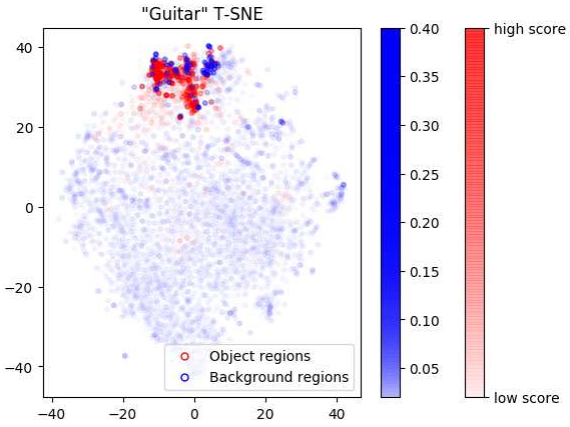
We compare our results to two different state-of-the-art weakly-supervised models - PCL [23] and SPN [30] in a setting of a single-class detection, as well as multi-class detection. For each model we use the code from the authors' GitHub. For SPN code we found that during test time the authors discard predictions that don't match the image-level labels. For a fair evaluation and comparison, we do not use image-level labels during test time. For the single-class test, to mimic the self-supervised scenario, we fed both models with noisy positive samples as well as clean negative samples. For each object category, the models were trained



(a) Drum



(b) Bike



(c) Guitar

Figure 4: **Potential score**. Object & Background samples are denoted by red and blue, respectively. Opacity represents score (darker points have higher score). Object samples (red) are more likely to be darker and, thus sampled as positive regions, reducing the label noise.

on two classes: the positive class with noisy labels, and the negative class labeled as background. For the multi-class comparison, both models were trained and tested with multi-class labeled data. As an upper bound reference, we report the performance of the fully-supervised binary detection version of our method, where the detector, namely object region classifier, is trained with ground truth labels. These comparisons emphasize the challenge of learning object detection from our unlabeled and unconstrained videos, manifesting motion blur, extreme views (far-field and close-ups), and large occlusions. Qualitative retrieval examples are shown in Fig. 3.

4.3. Evaluation

We randomly split our data into 80%-20% train-test sets containing mutually exclusive frames, and evaluate the performance on 3 random folds. The training and test sets contain on average 362 and 91 frames per object, respectively. Since our task is self-supervised, we allowed frames from the same video to participate in both train and test sets. For quantitative evaluation, we manually annotated the bounding boxes of 10 object categories. Note that annotations were used only for the testing and were not available at training, to keep the method annotation-free. As the evaluation criterion, we use the standard detection mean average precision (mAP) at intersection-over-union (IoU) of 0.5. For the case of single-class detection, training and testing were performed for each object category separately, according to the "pseudo-labels" acquired from the transcript and on the selected frames of the current object (i.e., the noisy positive set). For a fair comparison to WSOD multi-class models, we tested each of our single-class models on all frames. Note that since the positive set is noisy (see Table 1), the evaluation was in fact also applied to frames without the objects.

4.4. Detection results

To the best of our knowledge, this is the first self-supervised object detection method trained and evaluated using standard evaluation practices. The results are summarized in Table 2, for single-class and multi-class detection. For single-class detection, the performance of standard WSOD methods suffer greatly. In fact most of the times it fails to converge at all. However, since our model does not explicitly rely on discriminative parts, it is able to perform well even with single-class noisy data. Our single-class model attains a mAP of 17.7%, while PCL [23] and SPN [30] reach mAP of 7.2% and 0.8% (computed only on converged objects), respectively. While SPN attains an mAP of 11.9% for the multi-class detection task, our model obtains 16.0%, mAP, showing ∼34% relative improvement. This is an outcome of lack of robustness of SPN to label noise. Interestingly, when SPN is trained only on *clean la-*

beled data (with correct weak labels), it only reaches 13.5% mAP. This indicates the difficulty involved in object detection in our HowTo10 data set (due to variety, size, video artifacts, etc.). Yet, this is still lower than the 16.0% mAP that our model attains, when trained on noisy data. Since the detection part of our model is basic, we attribute this to the fact that multiple single-class models optimized separately perform better (as ensemble), or as good as a single multi-class model that is optimized for all objects at once - a setting which SPN fails to converge with. As the upper bound, we present the results from our trained region classifier (see Fig. 2) when fed with true region labels (starting with the SS Recall as in Table 1). Although we are still far from this extreme labeling scenario, we believe that our results and labeled data can motivate others to tackle the challenging task of self-supervised detection.

4.5. Retrieval results

Our model is able to learn and retrieve the appearance **and** names of objects without any manual labeling involved, simply by "watching and listening" to unconstrained unlabeled videos from the How2 data set [16]. To retrieve the top n instances, for a given object, we choose the cluster with highest potential score (choosing other top scoring clusters is possible as well). Filtering out regions using DSD, we extract the n closest samples to the cluster's center. Qualitative retrieval results for 10 objects for $n = 8$ are shown in Fig. 3.

4.6. Implementation details

For clustering, we use $K = 50$, and set $\tau = 50$ in (1). We set the positive ratio threshold as $P_k \geq 0.6$. In our region classifier we use 3 FC layers (1024,1024,2) with a ReLU activation in layers 1-2 and a softmax activation for the output layer. Dropout is used for the two hidden layers with probability of 0.8. The classifier is trained with the cross-entropy loss function. We use ADAM for optimization with a learning rate of 10^{-4}. The learning rate is decreased by a factor of 0.6 every 6 epochs. We train our model for 35 epochs for all objects. All experiments were done on a Tesla K80 GPU. After initial feature extraction, a single epoch duration (W-DEC, DSD & detector training) is around 15 minutes, amounting to nearly 9 hours for an object.

5. Analysis

5.1. Potential score function

In this section we demonstrate the effectiveness of our potential score function (1) in producing semantically meaningful clusters. In the absence of ground truth labels during training, we use the potential score function as an approximation. An effective approximation allows us to ignore noisy images/regions during training. We visualize the correlation of the potential score function with ground truth labels using t-SNE for three different objects in Fig. 4.

5.2. Single-class detection and scalability

Scaling is an important aspect of any system. Specifically, detection systems must be scalable in two ways: 1) data size and 2) number of objects. We argue that a self supervised single-class detection model is in fact scalable in both ways. It is known that self-supervision allows learning from abundant unlabeled data extracted from the web. However, the availability of data is just the first part. A model must be capable of utilizing a big data set for a practical task. Scaling a *multi-class* object detection model for a large number of objects is hard, as it requires a big complex model that must be trained from scratch on the entire data set every time a new object is added. On the other hand, using multiple single-class models in parallel instead allows adding new objects more easily. Training is done with a smaller model and on a small portion of the data. Unfortunately, WSOD models, which are also capable of using (weakly-labeled) web data, struggle with single-class detection (as we show in Table 2). As opposed to standard WSOD models, the proposed self-supervised model is capable of training a single-class detector since it does not rely on discriminative regions explicitly. Therefore, it is better scalable both in the data size and in the number of objects.

6. Summary

We have presented a model for the challenging tasks of self-supervised object detection and retrieval using unlabeled videos. Considering a large corpus of instructional videos with closed captions, we select frames that correspond to the transcript where the object name is mentioned. We pose the problem as weakly binary and noisily-labeled supervised learning. Our object detection is based on a model that captures regions with a common theme across the selected frames, distinguished from frames from disparate videos. This new region-level and single-class approach shows promising results in the detection of objects with high appearance variability and multiple sub-classes arising from the language ambiguities. Additionally, being self-supervised and single-class, it is easily scalable with the number of objects and size of data set. We evaluate our method in terms of detection mean average precision for single-class, as well as multi-class detection. We report an upper bound performance and demonstrate superior results compared to top performing weakly-supervised approaches. Our model handles noisy labels in the weak setting, and is capable of detecting objects in challenging scenarios without any human labeling.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Apostol (Paul) Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. In *arXiv:1609.08675*, 2016.

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, 2017.

[3] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016.

[4] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *CVPR*, 2018.

[5] Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R. Scott, and Dinglong Huang. Curriculumnet: Weakly supervised learning from large-scale web images. In *ECCV*, 2018.

[6] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018.

[7] David Harwath and James Glass. Learning word-like units from joint audio-visual analysis. In *ACL*, 2017.

[8] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep self-taught learning for weakly supervised object localization. In *CVPR*, 2017.

[9] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *NeurIPS*, 2018.

[10] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[11] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. In *NIPS*, pages 570–576, 1997.

[12] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[13] Ofir Nachum, Shixiang Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. In *NeurIPS*, 2018.

[14] Iftekhar Naim, Young Chol Song, Qiguang Liu, Henry Kautz, Jiebo Luo, and Daniel Gildea. Unsupervised alignment of natural language instructions with video segments. In *AAAI*, 2014.

[15] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

[16] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: a large-scale dataset for multimodal language understanding. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS, 2018.

[17] Bharat Singh, Mahyar Najibi, and Larry S. Davis. SNIPER: efficient multi-scale training. In *NeurIPS*, 2018.

[18] Young Chol Song, Iftekhar Naim, Abdullah Al Mamun, Kaustubh Kulkarni, Parag Singla, Jiebo Luo, Daniel Gildea, and Henry Kautz. Unsupervised alignment of actions in video with text descriptions. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 2025–2031. AAAI Press, 2016.

[19] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[20] Felix Sun, David Harwath, and James Glass. Look, listen, and decode: Multimodal speech recognition with images. In *IEEE Spoken Language Technology Workshop*, 2016.

[21] Didac Surís, Amanda Duarte, Amaia Salvador, Jordi Torres, and Xavier Giró-i Nieto. Cross-modal embeddings for video and audio retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[22] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 4278–4284, 2017.

[23] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Loddon Yuille. PCL: Proposal cluster learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[24] J. R. Uijlings, K. E. Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *Int. J. Comput. Vision*, 104(2):154–171, Sept. 2013.

[25] Fang Wan, Pengxu Wei, Zhenjun Han, Jianbin Jiao, and Qixiang Ye. Min-entropy latent model for weakly supervised object detection. In *CVPR*, 2018.

[26] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 478–487, 2016.

[27] Haonan Yu and Jeffrey Mark Siskind. Grounded language learning from video described with sentences. In *ACL*, 2013.

[28] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[29] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, 2018.

[30] Yi Zhu, Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Soft proposal networks for weakly supervised object localization. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[31] Bohan Zhuang, Lingqiao Liu, Yao Li, Chunhua Shen, and Ian Reid. Attend in groups: a weakly-supervised deep learning framework for learning from web data. In *CVPR*, 2017.