Learning Feature Hierarchies from Long-Range Temporal Associations in Videos

Panna Felsen Katerina Fragkiadaki Jitendra Malik Alexei Efros
Electrical Engineering and Computer Science
University of California, Berkeley
Berkeley, CA 94720
panna,katef,malik,eefros@berkeley.edu

Abstract

We propose learning a semantic visual feature representation by training a neural network supervised solely by point and object trajectories in video sequences. Currently, the predominant paradigm for learning visual features involves training deep convolutional networks on an image classification task using very large human-annotated datasets, e.g. ImageNet. Though effective as supervision, semantic image labels are costly to obtain. On the other hand, under high enough frame rates, frame-to-frame associations between the same 3D physical point or an object can be established automatically. By transitivity, such associations grouped into tracks can relate object/point appearance across large changes in pose, illumination and camera viewpoint, providing a rich source of invariance that can be used for training. We train a siamese network - we call it AssociationNet - to discriminate between correct and wrong associations between patches in different frames of a video sequence. We show that AssociationNet learns useful features when used as pretraining for object recognition in static images, and outperforms random weight initialization and alternative pretraining methods.

1 Introduction

Currently, the dominant paradigm for learning visual feature hierarchies is training deep convolutional neural networks to classify images into object or place categories, manually annotated by a large number of Amazon Mechanical Turkers [11, 4]. But are categorical annotations really necessary for learning good feature representations? Arguably, the labels themselves are not important: a random label permutation (that would name all cats chairs and the converse) would result in identical weights throughout the network apart from the very top layer, where high level feature activations, such as activation of the cat head neuron, would incorrectly give high probability to the chair class. However, such mistakes can be easily corrected using an additional small, correctly labelled training set, much smaller than Imagenet. Thus, it is easy to argue that the true value lies not in the category labels themselves but in the distances they induce on the images they describe: images that share the same label, e.g., both featuring cats, should have smaller feature distance than images featuring a cat and a dog respectively. In this paper, we propose mining image distance relationships from object and point trajectories in videos for building semantic hierarchies, bypassing category labels altogether. While category labels associate distinct, potentially visually dissimilar cats, video trajectories associate different transformations of the same cat in time, through pose, camera viewpoint, or illumination variability. We exploit such within-instance long temporal associations available in videos in lieu of human supervision needed for relating different instances of the same object category.

Given a video sequence, with arbitrary object and camera motion, we tracks objects and points. Object trajectories provide coarse, object-level correspondence, that persists across object rotations.
and object self-occlusions. Point trajectories provide fine-grain correspondences that terminate at occlusions. We train a siamese Convolutional Neural Network (CNN), which we call AssociationNet, to identify whether two image patches are projections of the same object or 3D physical point instance, i.e., whether the patches are part of the same object or point trajectory. We call this the instance re-identification task. Positive patch pairs are sampled on the same trajectory and negative patch pairs are hard-mined to have less than 50% overlap with the trajectory, yet non-zero overlap, as shown in Figure 1.

At test time, the weights from one stack of the AssociationNet are used as pre-training to aid recognition tasks with small number of training data, such as PASCAL object detection [6]. AssociationNet pretraining improves the mean Average Precision (AP) in PASCAL VOC object detection challenge by 3.9% over random weight initialization. It further outperforms alternative supervised pretraining methods that use smaller training sets than Imagenet, such as Caltech101. Our experiments show that learning within-instance associations delivers useful features for establishing cross-instance ones, suggesting that within-instance and cross-instance similarity is a continuum and long term video associations can effectively help towards effective unsupervised learning techniques.

2 Related work

Closest in spirit to our work is “slow feature learning”, a line of works that encourages learned features in videos to change slowly from frame to frame [20,21]. Slow feature analysis is inspired from psychophysical studies suggesting that brain activity in the higher levels of visual cortex can learn to become slow-changing, i.e., tolerant towards non-trivial transformations, by associating low-level features which appear in a coherent sequence [3,1,13]. The work of Mobahi et al. [16] combines temporal coherence with a supervised classification loss and shows improvements in detection performance in the COIL dataset [17], over a network that does not use temporal coherence. In COIL though, “classes” are images of the same instance from different viewpoints and the videos used in [16] are produced by rendering 3D meshes of objects against black background, with a camera rotating around the 3D object with constant velocity, conditions very different than real world videos. The proposed AssociationNet established feature distances between image patches not according to their frame distance, as SFL, but according to their spatial distance from trajectories. In this way, it is less susceptible to no motion or periodic motion, that can diminish the slowness assumption, yet can exploit really long temporal horizons, useful to associate dissimilarly looking image patches: the more dissimilar the associated instance appearances along a trajectory, the more useful the association.

Neural network models can be trained to model transformations between a pair of images, rather than the image content itself [14]. Work of Michalski et al. [15] trains a pyramid of bilinear gated autoencoders that reconstruct the next frame given an image sequence. Despite promising results on reconstruction, it has not been shown that the learned representations are useful for semantic tasks, such as object detection. This is the case with most unsupervised feature learning methods in the literature [5].
In this paper, we propose to instead learn feature representations through the discriminative task of instance re-identification in videos: instead of predicting the appearance of an image patch in time, we classify whether two patches are part or not of the same trajectory. This bypasses the problem of pixel reconstruction altogether.

## 3 AssociationNet

Figure 1 illustrates the AssociationNet architecture. It is a siamese network where each stack has an architecture similar to the network of [12]: assume \( C(k, N, s) \) is a convolutional layer with kernel size \( k \times k \), \( N \) filters and a stride of \( s \), \( P(k, s) \) a max pooling layer of kernel size \( k \times k \) and stride \( s \), \( N \) a normalization layer, \( RL \) a rectified linear unit, and \( FC(N) \) a fully connected layer with \( N \) filters. The architecture of each stack of our network is: \( C(7, 96, 2) \) − \( RL \) − \( P(3, 2) \) − \( N \) − \( C(5, 384, 2) \) − \( RL \) − \( P(3, 2) \) − \( N \) − \( C(3, 512, 1) \) − \( RL \) − \( C(3, 512, 1) \) − \( RL \) − \( C(3, 384, 1) \) − \( RL \) − \( P(3, 2) \) − \( FC(4096) \) − \( RL \) − \( FC(4096) \) − \( RL \). Two images, one per stack, are input to the network, which is trained to classify whether or not the input images belong to the same object or point trajectory. The input to the binary classifier of the final layer of the network is formed by subtracting the concatenated fc6 and fc7 in the two stacks, as shown in Figure 1.

We consider two granularities of temporal associations:

1. **coarse-grained object associations.** Object trajectories are obtained by densifying in time object annotations in the video segmentation benchmarks of Moseg [2] and VSB100 [7].

2. **point associations.** We obtain point trajectories automatically by linking optical flow fields across frames. We reject spurious optical flow vectors via a forward-backward consistency check [18]. Resulting optical flow trajectories provide fine-grain associations that terminate at point occlusions. We consider square patches of size \( 100 \times 100 \) pixels around the points tracked.

## 4 Experiments

We use the weights of AssociationNet to initialize R-CNN [9], currently a well established paradigm for object detection in the PASCAL VOC challenge. Given a bounding box proposal, Region-CNNs classify it as belonging to one of the object classes or background, and refine the box coordinates in case it is an object. For our experiments, we use the fast R-CNN publicly available code of [8], which does not use box warping, but rather computes feature in a fully convolutional way for the whole image and pools them over selective search box proposals [19].

We experiment with three different training regimes for AssociationNet: 1) AssociationNet-obj, trained only on object trajectories (~2500); 2) AssociationNet-tr, trained only on point trajectories (~12500); and 3) AssociationNet-objtr, trained on both object and point trajectories. For VOC

<table>
<thead>
<tr>
<th></th>
<th>zero</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>dog</th>
<th>horse</th>
<th>mbike</th>
<th>person</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>RandomNet</td>
<td>44.2</td>
<td>43.4</td>
<td>21.0</td>
<td>20.3</td>
<td>9.5</td>
<td>42.9</td>
<td>50.5</td>
<td>38.0</td>
<td>12.7</td>
<td>27.7</td>
<td>39.2</td>
<td>24.7</td>
<td>51.8</td>
<td>48.5</td>
<td>34.1</td>
<td>12.4</td>
<td>25.6</td>
<td>28.3</td>
<td>51.9</td>
<td>39.4</td>
</tr>
<tr>
<td>Caltech101Net</td>
<td>47.2</td>
<td>44.7</td>
<td>24.0</td>
<td>21.6</td>
<td>9.5</td>
<td>46.1</td>
<td>52.5</td>
<td>40.5</td>
<td>14.0</td>
<td>33.3</td>
<td>40.0</td>
<td>29.6</td>
<td>55.4</td>
<td>48.8</td>
<td>32.7</td>
<td>13.9</td>
<td>24.2</td>
<td>33.4</td>
<td>53.2</td>
<td>41.6</td>
</tr>
<tr>
<td>SlowFast-obj</td>
<td>46.4</td>
<td>44.0</td>
<td>22.8</td>
<td>20.4</td>
<td>9.4</td>
<td>45.4</td>
<td>52.3</td>
<td>39.7</td>
<td>13.7</td>
<td>32.4</td>
<td>39.5</td>
<td>33.1</td>
<td>55.0</td>
<td>49.9</td>
<td>37.8</td>
<td>13.7</td>
<td>23.6</td>
<td>36.8</td>
<td>50.2</td>
<td>42.2</td>
</tr>
<tr>
<td>SlowFast-objtr</td>
<td>47.3</td>
<td>45.8</td>
<td>22.0</td>
<td>19.8</td>
<td>9.3</td>
<td>44.5</td>
<td>52.5</td>
<td>37.7</td>
<td>14.5</td>
<td>32.1</td>
<td>40.8</td>
<td>35.4</td>
<td>57.5</td>
<td>49.9</td>
<td>38.5</td>
<td>13.8</td>
<td>22.4</td>
<td>32.4</td>
<td>53.7</td>
<td>41.4</td>
</tr>
<tr>
<td>AssociationNet-obj</td>
<td>49.0</td>
<td>46.1</td>
<td>21.3</td>
<td>24.9</td>
<td>9.8</td>
<td>48.9</td>
<td>54.5</td>
<td>38.6</td>
<td>15.7</td>
<td>34.6</td>
<td>46.6</td>
<td>31.8</td>
<td>54.5</td>
<td>51.4</td>
<td>41.2</td>
<td>14.5</td>
<td>25.3</td>
<td>35.9</td>
<td>54.1</td>
<td>42.8</td>
</tr>
<tr>
<td>AssociationNet-tr</td>
<td>46.4</td>
<td>45.0</td>
<td>22.6</td>
<td>24.7</td>
<td>10.1</td>
<td>50.7</td>
<td>54.3</td>
<td>42.2</td>
<td>16.4</td>
<td>31.2</td>
<td>43.4</td>
<td>30.2</td>
<td>54.7</td>
<td>50.4</td>
<td>35.2</td>
<td>10.9</td>
<td>25.1</td>
<td>38.7</td>
<td>51.3</td>
<td>36.9</td>
</tr>
<tr>
<td>AssociationNet-objtr</td>
<td>42.6</td>
<td>49.9</td>
<td>21.6</td>
<td>26.1</td>
<td>9.9</td>
<td>51.3</td>
<td>56.7</td>
<td>41.4</td>
<td>18.0</td>
<td>27.4</td>
<td>38.6</td>
<td>29.9</td>
<td>54.2</td>
<td>50.9</td>
<td>35.9</td>
<td>11.1</td>
<td>22.5</td>
<td>39.9</td>
<td>52.4</td>
<td>41.1</td>
</tr>
</tbody>
</table>

Table 1: **Object detection results on PASCAL VOC 2007 and 2012 using Fast R-CNN with two sets of hyper-parameters: one with a single scale and one that considers multiple image scales (* SA) during fine-tuning.**
Figure 2: **Conv1 filters.** Training using object trajectories results in many color filters. Training using point trajectories results in many checkerboard-like filters. The network trained using both object and point trajectories exhibits both types of filters.

2007, we experiment with two different Fast R-CNN hyper-parameter settings: 1) training images are shown to the network at a single scale, 2) training images are augmenting with multiple scales. Our VOC 2012 results were obtained with the single scale parameter setting.

We compare AssociationNet pretraining against the following baselines: 1) a network with random weights, termed RandomNet. 2) a network trained from scratch for image classification in Caltech 101 (Caltech101Net). Caltech101 has 9100 labeled examples and 101 object categories. 3) a stack of a triamese network trained with a ranking loss over triplets \(a, b, c\) of patches sampled on box and point trajectories: we request the Euclidean distance between fc7 features of \(a\) and \(b\) to be smaller by a certain margin than that of \(a\) and \(c\). This is an improved version of the contrastive slow feature learning loss of Mohabi et al. \[16\], applied to tracklets rather than whole frames. We consider two versions of this: one which samples triplets from only object trajectories (SlowFeat-obj) and one which samples triplets from both object and point trajectories (SlowFeat-objtr).

Table 1 reports the mean AP for AssociationNet pretraining and baselines. AssociationNet outperforms the baselines. Specifically, it outperforms Caltech101Net, showing category labels are not the solely way to good feature hierarchies. Further, in the single-scale setup, AssociationNet-obj and AssociationNet-tr outperform RandomNet by 3.9\% and 2.6\% mAP, respectively. In the multi-scale setup, AssociationNet-objtr and AssociationNet-tr each outperform RandomNet by roughly 2\%. In this setting, Caltech101Net and SlowFeat-obj are below by 0.9\% and 1.2\%. Given that scale augmentation effectively produces more training examples, it is sensible that the average performance gain of our method in the multi-scale regime is smaller than in the single-scale one. The more the training data, the less the importance of pretraining.

Object trajectories and point trajectories appear complementary. In the multi-scale parameter setting, AssociationNet-objtr boosts the mAP by 1.9\% above SlowFeat-objtr and by 3.7\% above RandomNet, which is almost the sum of the individual performance gains. The layer-1 filters, shown in Figure 2, are a mixture of the color and edge detectors from AssociationNet-obj and the bar and checkerboard filters from AssociationNet-tr. Unsurprisingly, AssociationNet-tr does not primarily learn color detectors: point trajectories are oftentimes surrounded by uniformly colored regions.

5 Conclusion

We have presented a method that learns from within-instance temporal associations of points and objects feature representations useful for object recognition. We believe this work opens exciting future research directions for large scale temporal association mining and feature representation learning. An important theoretical contribution of the work is the bridge between within-instance and across-instance similarity, that we treat as a continuum rather than as disjoint tasks. Pose, viewpoint, illumination variability of an object instance in time generalizes to correspondences across different instances of an object category. Interleaving such temporal based learning with sparse strong supervision in the form of object labels better matches the biological paradigm of a teacher sparsely revealing object labels to a moving observer, as opposed to fully supervised learning exclusively based on human annotations. Exploring such interleaving of temporal and semantic supervisions is the scope of our future work.
References


