Learning visual groups from co-occurrences in space and time
(2015)

Phillip Isola, Daniel Zoran, Dilip Krishnan, Edward H. Adelson

Notes

Contributions

Based on the fact that the physical world is highly structured, adjacent locations are usually semantically related. The authors propose a new self-supervised training objective, setting up a simple binary classification problem in which the goal is to predict if two visual inputs occur in the same temporal context (two video frames are close together) or spatial context (two image patches are adjacent).

With this simple kind of grouping, the learned representations can uncover semantically meaningful groupings, such as using spatial proximity of the patches as a proxy for finding if two patches belong to the same object, giving results competitive with state of the art methods.

Method

Given two visual primitives, A and B, we train a classifier in a supervised manner to predict spatial or temporal proximity $\mathcal{C} \in \{0, 1\}$, where $\mathcal{Q} = 1$ iff A and B are nearby in space or in time, this proxy task can then be used to approximate another task where the training examples are hard to acquire (e.g. A and B belong to the same object) in the hopes that $\mathcal{C}$ will serve as a cheap proxy for $\mathcal{Q}$.

The proximity is modeled as the probability of the two primitives accuring within some context $\mathcal{C} = 1$ with symmetry enforced (A close to B and B close to A):

$$w(A, B) = \frac{P(\mathcal{C} = 1|A, B) + P(\mathcal{C} = 1|B, A)}{2}$$

In case A and B are independent when we sample *iid* manner, we'll endup with $P(\mathcal{C} = 1|A, B) \propto \frac{P(A,B|\mathcal{C}=1)}{P(A)P(B)}$ and in this case: $w(A, B) = P(A)P(B)$.

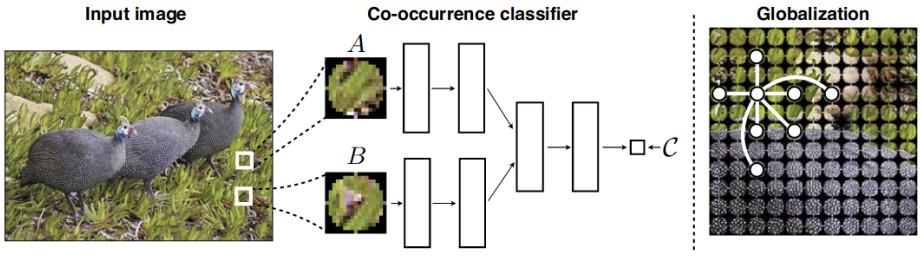

Figure 2: Overview of our approach to learning to group patches. We train a classifier to that takes two isolated patches, A and B, and predicts $\mathcal{C}$: whether or not they were taken from nearby locations in an image. We use the output of the classifier, $P(\mathcal{C} = 1|A, B)$, as an affinity measure for grouping. The rightmost panel shows our grouping strategy. We setup a graph in which nodes are image patches, and all nearby nodes are connected with an edge, weighted by the learned affinity (for clarity, only a subset of nodes and edges are shown). We then apply spectral clustering to partition this graph and thereby segment the image. The result on this image is shown in Figure 5.

A CNN with parameters θ is used to model $w(A, B)$, the network takes as input the two visual primitives A_i and B_i and outputs a probability C_i of spatial / temporal proximity between the two visual primtives, the network is then trained using a simple logistic regression loss function:

$$E(\mathbf{A}, \mathbf{B}, \mathcal{C}; \theta) = \frac{-1}{N} \sum_{1}^{N} \mathcal{C}_i \log\left(\sigma\left(f\left(\mathbf{A}_i, \mathbf{B}_i; \theta\right)\right)\right) + \left(1 - \mathcal{C}_i\right) \log\left(1 - \sigma\left(f\left(\mathbf{A}_i, \mathbf{B}_i; \theta\right)\right)\right)$$

The number of examples used for training is $N = 500.000$ with 50% negatives and 50% positives.

Results

Co-occurences in images In images, A and B are a set of 17×17 pixel patches with circular masks, $\mathcal{C} = 1$ for adjacent pixel with no overlap and $\mathcal{C} = 0$ for radomly sampled patches. To evaluate the performance of this task, we evaluate the model on the target task $\mathcal{Q}$, so after having 50% negatives and 50% positives of examples for the proxy task, we can train the network to predict the spatial proximity. We then sample 100% positives $\mathcal{Q}$ (i.e. patches belonging to the same object), and see how is the accuracy of our network on these pairs of patches. Giving us an average precision of 0.8, so our co-occurance objective is a good proxy for object detection.

	Patches		Frames		Photos	
Affinity measure	$\mathcal{C}$	$\mathcal{Q}$	$\mathcal{C}$	$\mathcal{Q}$	$\mathcal{C}$	$\mathcal{Q}$
Raw color	0.83	0.73	0.77	0.58	0.58	0.58
Mean color	0.87	0.74	0.82	0.63	0.56	0.57
Color histogram	0.95	**0.80**	0.90	0.64	0.63	0.62
HOG	0.67	0.67	0.77	0.61	0.63	0.75
Co-occurrence classifier	**0.96**	**0.80**	**0.95**	**0.67**	**0.70**	**0.79**

Forming visual objects Another usage of the co-occurences, given that this task is a good proxy for finding if two patches belong to the same object is to contruct visual groupd, go segmentation masks in an unsupervised manner, first we train our network to predict $\mathcal{C}$ using patches from Pascal Voc training set.

For a given image, to form visual groups, we first devide it into patches of 17 x 17 pixels, with strides of 8 and a circual mask to avoid any overlaping, all the patches that are spatially close together (less than 33 and more than 17) are connected by an edge, and this edge is weighted by the output of the netword $\mathbf{W}_{i,j} = w\left(\mathbf{A}_i, \mathbf{B}_i\right)^{\alpha}$ using two nodes A_i and B_i as inputs (with $\alpha = 20$), resulting in an affinity matrix $\mathbf{W}$. We apply spectral clustering to the matrix W to find the clusters (by finding the edges to remove), ending enup with good visual results:

Figure 4: Example object proposals. Out of 100 proposals per image, we show those that best overlap the ground truth object masks. Average best overlap (defined in Krahenbuhl & Koltun (2014)) and recall at a Jaccard index of 0.5 are superimposed over each result.