Using Human Knowledge to Judge Part Goodness: Interactive Part Selection

Ejaz Ahmed[1], Subhransu Maji[2], Gregory Shakhnarovich[2], and Larry S. Davis[1]
[1]University of Maryland, College Park and [2]Toyota Technological Institute at Chicago
{ejaz,lsd}@umiacs.umd.edu,{smaji,greg}@ttic.edu

Abstract

It is a common practice to model object detectors as collection of filters. For these detectors to be effective, it is important to select "good filters" covering most of the variation of the data. In order to achieve this, these methods invest majority of their time selecting a good subset of filters from a large pool. Good filters are less cluttered and their gradients are spatially correlated. Humans can differentiate between a good filter and a bad one by visualizing them. In addition humans bring with them the knowledge of diversity and effectiveness of filters, properties which are difficult to model. In this work, we show that humans can help build better detectors by including their knowledge of good filters. We show this by building an interactive framework for poselet selection. Our interactive framework improves the detection performance on the PASCAL VOC dataset and significantly improves the training time.

1. Introduction and Motivation

A common approach to modeling a visual category is to represent it as a composition of smaller fragments (parts) arranged in a variety of layouts or as library of exemplars, more generally, as collection of filters. Modeling a category as collection of filters helps in modeling a large amount of variation in data. In this work our focus is on an architecture which has two major steps (i) Candidate Generation and (ii) Selection. Examples of such architectures are poselets [2, 1], exemplar SVM [5] and discriminative patches [6]. In this paper we investigate poselets.

Candidate generation step involves training many HOG detectors [3] each representing a part of the object. This step involves training a HOG detector and mining for hard negatives, and is moderately expensive. The Selection step is most expensive and involves evaluating each generated part filter (large pool) on a large number of positive and negative examples. The selected parts should be (i) Discriminative - they should fire only at meaningful locations on the test image and (ii) Diverse - many candidate parts are highly similar to each other, there is no point in selecting very similar parts twice.

What is a good part? One way to determine the discriminativeness of filters is by evaluating them on a held-out set [1]. This is very expensive as there are many candidate parts. However, good discriminative parts have the property that they are less cluttered and their gradients are spatially correlated. Figure 1 shows examples of good and bad filters. It is easy to tell the difference between the two by visual inspection. For good filters, neighboring gradient orientation bins are active simultaneously and majority of them are entirely suppressed. This is due to the fact that the template has to account for small variations in local gradient directions in order to be robust. Also note that if a certain gradient orientation is encouraged, its orthogonal counterpart is often penalized. Also, dominant orientation bins of neighboring cells tend to coincide forming line segments or disagree by an angle to form curves and corners, or be parallel. This could be attributed to the fact that the template has to be robust to small spatial variations in alignment of training samples, which results in neighboring cell show similar patterns. Also, gradient based nature of HOG features tend to capture object outlines which are often smooth lines and curves. [4] exploited these properties in a generative framework to come up with structured prior for the SVM loss function. In addition to these there are certain properties which are difficult to model. For instance, a filter trained on the legs of a person (parallel lines) is although a good one, in the sense that gradients might be less cluttered. But in practice such a filter will not be effective since parallel lines are common in the real world and such a detector will fire all over the place.

In this work we show that it is possible to select filters with the help of human knowledge of good filters. Humans bring with them the knowledge about (i) good filters - by visualizing one can tell the difference between good and bad filters, (ii) diverse filters - humans know if two filters represent the same part and (iii) effective filters (leg example). We present an interactive framework to select discriminative and diverse set of filters with minimal effort.

	aplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	**mean**
Oracle	**32.37**	50.00	**12.82**	16.36	**31.57**	41.30	**56.00**	**20.84**	**19.20**	37.55	14.51	**17.04**	37.63	**35.91**	**36.65**	13.14	**31.87**	23.35	24.31	28.21	29.03
Interactive	29.84	**50.88**	12.57	**20.16**	31.48	**43.59**	55.82	19.85	18.29	**40.08**	**15.42**	16.66	**44.47**	35.08	35.56	**13.26**	31.55	**27.03**	**25.50**	**30.66**	**29.89**
Overlap	23	20	31	19	30	31	29	30	20	33	16	41	26	33	45	22	17	2	22	14	25.20

Table 1. Per Category Results on PASCAL VOC 2007. Best results are highlighted in bold. Best mean average precision of 29.86 is obtained using our interactive method. Note that this method resulted in significant speed up of the selection pipeline (5-10mins as compared against 15Hrs). Last row shows number of filters common to the two methods for each category.

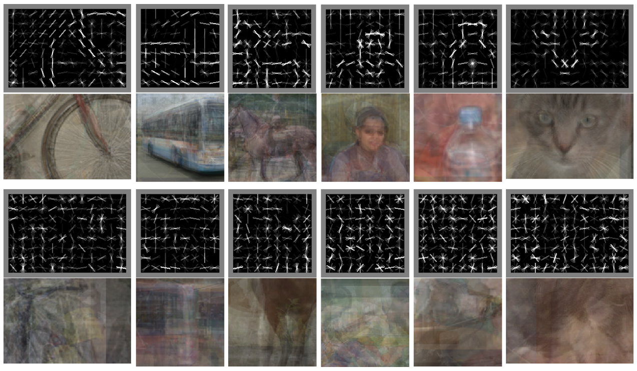

Figure 1. Good-Bad Filters and Seed Image(made by averaging top 10 seeds): Top 2 rows show good filters and bottom 2 rows show bad filters. Categories from left to right are bicycle, bus, horse, person, bottle and cat. Note the difference between gradient orientations for good and bad filters. Also seed images for the good filters are cleaner.

2. Original Poselet Selection

In the original paper [1] selection is performed by evaluating candidate poselets on the entire training set, and a subset is selected using a greed coverage algorithm that iteratively picks poselets that offer highest increase in detection accuracy at a fixed false positive rate. This is time consuming and takes 76% (15Hrs) of the total time (20Hrs) to train. Also note that this is just an approximation to discriminative and diversity criteria.

3. Interactive Selection

The idea is to display filters to a user and let it select good and diverse set of filters. To assist the user a simple heuristic of filter quality (norm) is used. As we know selected filters should be diverse too, hence we define a measure for similarity between a non selected filter and a set of selected filters. Initially, we display the filters sorted according to the descending value of normalized norm ($\frac{norm}{\#cells}$) of the filter. This acts as a heuristic of good filter and tends to push good filters higher in the visualization box. The user then browses through the filters and selects k = 5-10 filters. Then re-ranking of non-selected filters is done according to their similarity to the selected filters. For obtaining the similarity we compute the bounding box overlap of top $r = 3\%$ of the ordered list of training examples of poselets. Similarity of a non-selected filter with a set of selected filters is determined as k-th order value of similarity between candidate part and those already selected. This helps assist user to select good filters with a minimal effort and without browsing through all of the candidate filters. The process is then repeated until desired number of poselets are selected.

4. Experiments

We constructed a poselet model by selecting 100 poselets from a set of 800 poselets using original poselet selection method (Oracle) and using our interactive framework. For our interactive framework, initially user selects 10 poselets and then clicks "process of diversity". In the subsequent iterations user selects 5-10 poselets before clicking for "process of diversity". These steps are repeated until 100 poselets are selected. It takes about 5-10mins for selecting 100 poselets for a category. We evaluate these models as detectors on PASCAL VOC 2007 dataset. To isolate the effect of poselet selection, we use a simplified implementation that avoids some of the post-processing steps (Q-poselet models). The results are reported in table 1. We achieve an improvement in the detection performance, $\Delta(mAP) = +0.86$.

5. Conclusion

We have presented an interactive framework for selecting discriminative and diverse set of parts. With our method user knowledge of a good part enters the training pipeline. Our method significantly improves the training time, it takes about 5-10mins for a user to interactively select 100 poselets. As compared to 15Hrs (76% of 20Hrs) for computer to select them. Last but not the least, our method helps in constructing better detectors ($\Delta(mAP) = +0.86$). Moreover this method helps us to understand what a good detector is and gives direction for developing automatic methods for good part selection.

References

[1] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, 2010. 1, 2

[2] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009. 1

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1

[4] T. Gao, M. Stark, and D. Koller. What makes a good detector? structured priors for learning from few examples. In *ECCV*, 2012. 1

[5] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011. 1

[6] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012. 1