Introspective Semantic Segmentation

Gautam Singh Jana Košecká
George Mason University
Fairfax, VA
{gsinghc,kosecka}@cs.gmu.edu

1. Introduction

The problem of semantic segmentation requires simultaneous segmentation of an image into regions and categorization of all the image pixels. Traditional approaches for semantic segmentation work in a supervised setting assuming a fixed number of categories and require large training sets. The performance is then reported in terms of the average per class accuracy and the global accuracy of the final labeling on the test split of a dataset. When applying the learned models in practical settings on unlabeled images possibly containing previously unseen categories, it is important to quantify the confidences of the image region classifiers. In this work, we propose to do so in the context of a non-parametric weighted k-nearest neighbor (k-NN) framework for semantic segmentation by using the **strangeness** measure as shown in Figure 1. The proposed measure is evaluated by introducing confidence based image ranking and showing its feasibility for a dataset containing large number of previously unseen categories.

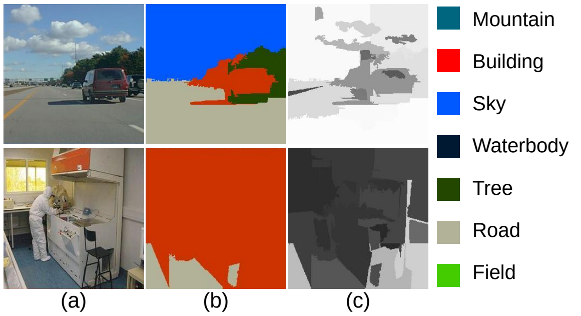

Figure 1. Example images from SUN09 labeled using SiftFlow as source dataset (best viewed in color). (a) Input image (b) Predicted labeling. (c) Strangeness based uncertainty. Darker intensity implies higher uncertainty of labeling. First row image has low uncertainty for the majority except for the road divider and vehicle which are unfamiliar categories. Second row is indoor scene with most of the image regions associated with high uncertainty.

2. Semantic Segmentation Approach

The semantic labeling is formulated on SLIC superpixels characterized using geometric and appearance features described in [4]. Unlike the non-parametric approach of [5] which treats all feature channels equally, for the labeling of the superpixels, we use a weighted k-NN classifier [1] to better exploit the contribution of the different feature channels to the variations between different semantic categories. The weights are estimated at test time by considering the local neighborhoods around a test point as described in [1].

3. Strangeness Measure

In this work, we are interested in associating an uncertainty with the predicted labeling. For this purpose, we extend our k-NN method to help identify image regions which can be characterized as *unfamiliar*. Given a source set T_s of labeled images, we segment the images and compute their features yielding the dataset of superpixel features and labels $G = \{(\mathbf{a_i}, y_i)\}$ where y_i is the segment label. For each segment s_i in this dataset, an individual strangeness [3] measure α_i is computed: $\alpha_i = \frac{\sum_{r=1}^{K} d_{ir}^{c}}{\sum_{r=1}^{K} d_{ir}^{\bar{c}}}$ where $c = y_i$ is the semantic label for segment s_i, d_{ir}^{c} is the r-th shortest distance between s_i and an instance of class c, $d_{ir}^{\bar{c}}$ is the r-th shortest distance between s_i and an instance not belonging to class c and K is the number of nearest neighbors considered for each sum. The neighbors are computed using the weighted k-NN method of the previous section. The strangeness measures how "strange" an instance is with respect to its semantic category as an example closer to own class instances in comparison to other class instances has higher strangeness and vice versa. After computing the strangeness for the instance, we count the number of examples of the category in the dataset which have a larger strangeness value and compute the p-value statistic proposed by [3]: $t_i = \sum_{\substack{r=1 \\ y_i=y_r=c}}^{|c|} 1\{\alpha_i > \alpha_r\}/|c|$ where $|c|$ is the number of instances in G with the label c and $1\{.\}$ is the indicator function. The value t_i can be viewed as a measure of the probability of having instances in the class with strangeness greater than or equal to that of s_i. The strangeness and corresponding p-values are computed for all the instances in G.

Given an image to label, while performing its semantic labeling, we also wish to discover regions which do not belong to any of the categories from G and this is done by utilizing the *strangeness* measure. When computing the strangeness for instances in G, we already know the instance's category and the strangeness computation is straightforward. However, for the regions in an input image, the category is unknown. But our goal is determining if the region belongs to *any* of the known semantic categories or not. Hence, we compute the strangeness and p-value for an input image superpixel assuming its putative label to be each of the known categories $\{1, 2, \ldots, L\}$ one by one i.e. given a segment s_i, we compute α_i^c and t_i^c $\forall$ $c \in \{1, 2, \ldots, L\}$. The uncertainty for the region to belong to the known semantic categories is defined as: $u_i = \min_{c=1}^{L}(1 - t_i^c)$. We compute the uncertainty as a complement of t_i^c as a lower t_i^c corresponds to higher uncertainty for s_i with respect to label c and the subsequent minimum function selects the least *strange* label for s_i. Examples of strangeness based uncertainty are shown in Figure 1 with more qualitative results in [4].

4. Experiments

Evaluation is carried out on the large scale SUN09 dataset composed of 8,662 images.

Cross Dataset Semantic Segmentation We first evaluate the efficacy of the weighted k-NN for cross dataset semantic labeling. We consider data from a source set and evaluate the learned models on another dataset. We perform labeling for the frequent common categories - *sky, building, tree, mountain, road, sea, field* using two smaller source sets: Stanford background and SiftFlow on the larger SUN09 dataset. As a baseline, we train a boosting classifier using decision trees as the weak learners. When using SiftFlow on SUN09, the per pixel (per class) accuracy is: boosting - 73.9 (61.7), weighted k-NN (WKNN) - 73.4 (61.9) and strange k-NN (which selects least strange label) - 74.2 (69.9). For Stanford, the accuracy is boosting - 68 (59.4), WKNN - 67.6 (58.9) and strange k-NN - 68.5 (60.9). The k-NN performance is similar to the boosting classifier and interestingly, using the least strange label outperforms the weighted k-NN output.

Confidence based Ranking The next evaluation focuses on the introspective capacity of the approach. Both Stanford and SiftFlow are composed of outdoor scenes while SUN09 consists of both outdoor and indoor scenes thereby providing an ample set of images on which confidence based ranking can be evaluated. We compare the strangeness based uncertainty to two baseline methods previously utilized for computing classifier confidence in active learning [2]: Normalized entropy (NEP) and Best versus Second Best probability (BvSB) which are computed using the probabilistic outputs of the decision tree based boosting classifier.

Uncertainty is computed for all superpixels of an image and an uncertainty score is computed for the image as a weighted sum of the uncertainties (based on superpixel size). Having obtained the image level uncertainty score, we sort the images of SUN09 in a descending order of the scores. For any metric which is being evaluated for confidence ranking, the goal is to obtain more images with higher content of unfamiliar categories in the higher ranks e.g. a beach scene should ideally have lower rank than a hospital scene when trained with dataset of outdoor scenes. The results for the different measures when using SiftFlow on SUN09 dataset is presented in Figure 2.

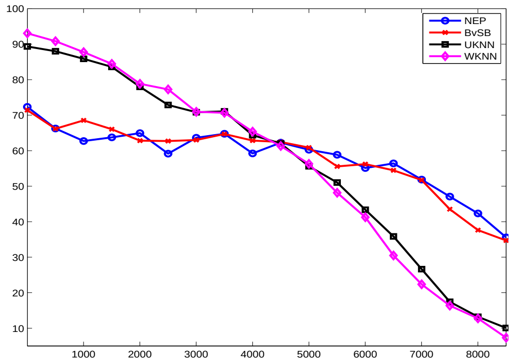

Figure 2. Comparison of uncertainty measures for confidence ranking of SUN09 images using SiftFlow dataset. UKNN refers to uniformly weighted k-NN. Y-axis denotes the percentage of unfamiliar category pixels in images of a particular ranking subset (size 500 images) using an uncertainty measure e.g. there are 93.1% unfamiliar pixels in images ranked 1-500 using WKNN strangeness while NEP has 72.3% unfamiliar pixels.

As can be observed, strangeness outperforms both NEP and BvSB as images with higher strangeness uncertainty scores include a higher percentage of unfamiliar category pixels. As the rankings drop, there is a decrease in the unfamiliar category pixels indicating that images with familiar categories are associated with lower uncertainty scores. The UKNN output outperforms both NEP and BvSB indicating the efficacy of using the transductive strangeness over the probabilistic output of the boosting classifier. When the strangeness is computed using a weighted k-NN, the performance improves highlighting the utility of feature relevance in a nearest neighbor framework.

Acknowledgements Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory, contract FA8650-12-C-7212. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

References

[1] C. Domeniconi, J. Peng, and D. Gunopulos. Locally adaptive metric nearest-neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1281–1285, 2002.

[2] A. J. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. In *CVPR*, pages 2372–2379, 2009.

[3] K. Proedrou, I. Nouretdinov, V. Vovk, and A. Gammerman. Transductive confidence machines for pattern recognition. In *ECML*, pages 381–390, 2002.

[4] G. Singh and J. Košecká. Introspective semantic segmentation. In *WACV*, 2014.

[5] J. Tighe and S. Lazebnik. SuperParsing: Scalable nonparametric image parsing with superpixels. *International Journal of Computer Vision*, 101(2):329–349, 2013.