Summaries for "Weakly Supervised Semantic Segmentation with Convolutional Networks"

Weituo Hao 109241801
Stony Brook University
Computer Science Department
Advanced Computer Vision

Contribution

To solve the problem of inferring object segmentation by leveraging only object class information, and by considering only minimal priors on the object segmentation task, the author converts it to a kind of weakly supervised segmentation task which fits the Multiple Instance Learning framework. A Convolutional Neural Network-based model is proposed. This system is tested on a subset of the Imagenet dataset and the segmentation experiments are performed on the challenging Pascal VOC dataset. And the author claims their model beats the state of the art results in weakly supervised object segmentation task by a large margin.

Main idea

The architecture is a CNN, which is trained over a subset of Imagenet in an end-to-end manner, to produce pixel-level labels from image-level labels. The CNN is constrained during training to put more weights on pixels which are more important for classifying the image.

First a standard CNN architecture with 10 levels of convolutions and pooling is adopted and froze the first layer. Note that the writers make use of Overfeat architecture's first 6 convolution layers and 2 pooling layers which is a sort of pre-training. Then the author continues to add another four convolutional layers. Each of them is followed by a pointwise rectification non-linearity. The whole framework can be described as the following picture shows.

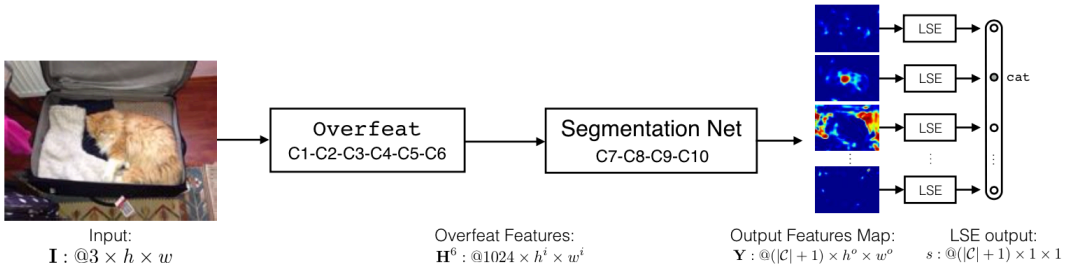

Figure 1 The CNN architecture of the proposed method

Second, the network above will produce one score for each pixel location (i,j) from the subsampled image I. And the image-level class scores are interpreted as class

conditional probabilities by applying softmax. $p(k|I,\theta) = \frac{e^{s^k}}{\sum_{c\in C} e^{s^c}}$, where θ means all the trainable parameters in the framework.

In addition, to reduce the false positive the author suggests to use image-level prior or smoothing prior which forces pixels with low probability of being part of an object to be labeled as background and guarantees local label consistency.

Experimental results

The Pascal VOC dataset is considered as a benchmark for segmentation. Comparison with fully supervised method can be seen in Table 1.

	bgnd	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Fully Sup.																						
O_2P	86.1	64.0	27.3	54.1	39.2	48.7	56.6	57.7	52.5	14.2	54.8	29.6	42.2	58.0	54.8	50.2	36.6	58.6	31.6	48.4	38.6	47.8
DivMBest	85.7	62.7	25.6	46.9	43.0	54.8	58.4	58.6	55.6	14.6	47.5	31.2	44.7	51.0	60.9	53.5	36.6	50.9	30.1	50.2	46.8	48.1
SDS	86.3	63.3	25.7	63.0	39.8	59.2	70.9	61.4	54.9	16.8	45.0	48.2	50.5	51.0	57.7	63.3	31.8	58.7	31.2	55.7	48.5	51.6
Weak. Sup.																						
Ours-sppxl	74.7	38.8	19.8	27.5	21.7	32.8	40.0	50.1	47.1	7.2	44.8	15.8	49.4	47.3	36.6	36.4	24.3	44.5	21.0	31.5	41.3	35.8
Ours-obj	76.2	42.8	20.9	29.6	25.9	38.5	40.6	51.7	49.0	9.1	43.5	16.2	50.1	46.0	35.8	38.0	22.1	44.5	22.4	30.8	43.0	37.0
Ours-seg	78.7	48.0	21.2	31.1	28.4	35.1	51.4	55.5	52.8	7.8	56.2	19.9	53.8	50.3	40.0	38.6	27.8	51.8	24.7	33.3	46.3	40.6

Table 1. Comparison with previous fully supervised methods

From this table the author claims that their model achieves significantly better results than the previous state- of-the-art weakly supervised algorithms, with an increase from 30% to 90% in average per-class accuracy.

Conclusion

The author proposes an innovative framework to segment objects with weakly supervision only and claims their approach surpasses by a large margin previous state-of-art models for weakly supervised segmentation as well as achieve competitive performance compared to state-of-the-art fully supervised segmentation systems.

Summaries for "Deep Convolutional Neural Fields for Depth Estimation from a Single Image"

Weituo Hao 109241801
Stony Brook University
Computer Science Department
Advanced Computer Vision

Contribution

The author proposes to consider the continuous characteristic of the depth values, depth estimations can be naturally formulated into a continuous conditional random field (CRF) learning problem. Therefore this paper presents a deep structured learning scheme which learns the unary and pairwise potentials of continuous CRF in a unified deep CNN framework. NYU v2 indoor scene reconstruction and Make3D outdoor scene reconstruction datasets are used to verify the method.

Main idea

Following the assumption that an image is composed of small homogeneous regions (superpixels), the writer let each superpixel portrayed by the depth of its centroid. Let $\mathbf{x}$ be an image and $\mathbf{y} = [y_1, \ldots, y_n]^T \in R^n$ be a vector of continuous depth values corresponding to all n superpixels in $\mathbf{x}$. So the whole image can be viewed as a graphical model with vertexes and edges. To predict the depth of a new image the problem can be summarized as following equation.

$$\Pr(\mathbf{y}|\mathbf{x}) = \frac{1}{\mathrm{Z}(\mathbf{x})} \exp(-E(\mathbf{y}, \mathbf{x}))$$

where

$$\mathrm{Z}(\mathbf{x}) = \int_{\mathbf{y}} \exp\{-E(\mathbf{y}, \mathbf{x})\} \mathrm{d}\mathbf{y}.$$

So we need to solve the following optimization equation

$$\mathbf{y}^\star = \underset{\mathbf{y}}{\operatorname{argmax}} \Pr(\mathbf{y}|\mathbf{x})$$

The original paper has demonstrated that the MAP inference problem listed above can be solved in a closed form. And the whole framework can be described as the following picture. In particular, the objective function can be divided into two parts. The first part is to deal with the nodes in the graphical model. It adopts the 5-layer convolution plus 4 fully connected layer. All the parameters are shared among different filters.

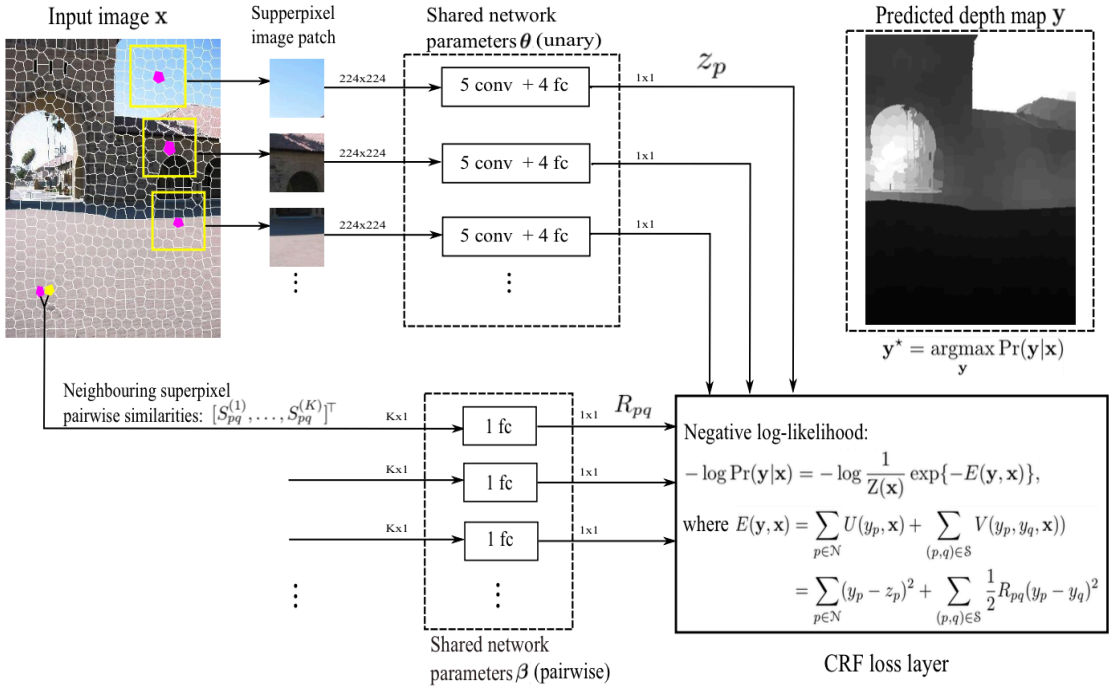

Figure 1 The main architecture of the proposed method

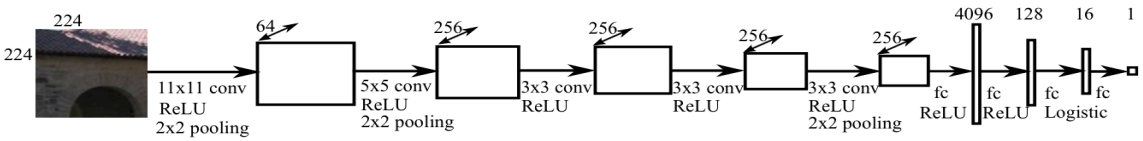

Figure 2 The CNN for the unary part

The second part (the bottom part in figure 1) means to compute the pairwise similarities neighbor superpixels of the original image. Three kinds of similarities are used: the color difference, color histogram difference, and texture disparity in terms of local binary patterns. All these similarities are calculated in L2 norm.

Experimental results

To evaluate the method, the author compares the result with previous work in the following terms

average relative error (rel): $\frac{1}{T}\sum_p \frac{|d_p^{gt} - d_p|}{d_p^{gt}}$;

root mean squared error (rms): $\sqrt{\frac{1}{T}\sum_p (d_p^{gt} - d_p)^2}$;

average $\log_{10}$ error (log10):
$\frac{1}{T}\sum_p |\log_{10} d_p^{gt} - \log_{10} d_p|$;

accuracy with threshold thr:
percentage (%) of d_p s.t.: $\max(\frac{d_p^{gt}}{d_p}, \frac{d_p}{d_p^{gt}}) = \delta < thr$;

The data are from two datasets: NYU v2 indoor scene reconstruction. The result is listed in the table 1

Method	Error (lower is better)			Accuracy (higher is better)		
	rel	log10	rms	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Make3d [15]	0.349	-	1.214	0.447	0.745	0.897
DepthTransfer [5]	0.35	0.131	1.2	-	-	-
Discrete-continuous CRF [16]	0.335	0.127	1.06	-	-	-
Ladicky *et al*. [8]	-	-	-	0.542	0.829	0.941
Eigen *et al*. [1]	**0.215**	-	0.907	0.611	**0.887**	**0.971**
Ours (pre-train)	0.257	0.101	0.843	0.588	0.868	0.961
Ours (fine-tune)	0.230	**0.095**	**0.824**	**0.614**	0.883	**0.971**

Table 1 Comparison with previous work on NYU dataset

And the second dataset is Make3D outdoor scene reconstruction

Method	Error (C1) (lower is better)			Error (C2) (lower is better)		
	rel	log10	rms	rel	log10	rms
Make3d [15]	-	-	-	0.370	0.187	-
Semantic Labelling [7]	-	-	-	0.379	0.148	-
DepthTransfer [5]	0.355	0.127	9.20	0.361	0.148	15.10
Discrete-continuous CRF [16]	0.335	0.137	9.49	0.338	0.134	**12.60**
Ours (pre-train)	0.331	0.127	8.82	0.324	0.134	13.29
Ours (fine-tune)	**0.314**	**0.119**	**8.60**	**0.307**	**0.125**	12.89

Table 1 Comparison with previous work on Make3D dataset

Conclusion

The author proposes to combine CNN and CRF method to estimate the depth from single image and manages to convert the problem into a MAP inference problem with a closed solution. The result shows that without any extra tuning or training data the methods can achieve the same level accuracy as previous work did.