Semi- and Weakly- Supervised Semantic Segmentation with Deep Convolutional Neural Networks

Yuhang Wang, Jing Liu, Yong Li and Hanqing Lu
The National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
{yuhang.wang, jliu, yong.li, luhq}@nlpr.ia.ac.cn

ABSTRACT

Successful semantic segmentation methods typically rely on the training datasets containing a large number of pixel-wise labeled images. To alleviate the dependence on such a fully annotated training dataset, in this paper, we propose a semi- and weakly-supervised learning framework by exploring images most only with image-level labels and very few with pixel-level labels, in which two stages of Convolutional Neural Network (CNN) training are included. First, a pixel-level supervised CNN is trained on very few fully annotated images. Second, given a large number of images with only image-level labels available, a collaborative-supervised CNN is designed to jointly perform the pixel-level and image-level classification tasks, while the pixel-level labels are predicted by the fully-supervised network in the first stage. The collaborative-supervised network can remain the discriminative ability of the fully-supervised model learned with fully labeled images, and further enhance the performance by importing more weakly labeled data. Our experiments on two challenging datasets, i.e, PASCAL VOC 2007 and LabelMe LMO, demonstrate the satisfactory performance of our approach, nearly matching the results achieved when all training images have pixel-level labels.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: [Vision and Scene Understanding]; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

Keywords

Semantic Segmentation; Semi-Supervised Learning; CNN

1. INTRODUCTION

Image semantic segmentation is a core problem in computer vision, aiming at parsing images into several semantic regions and labeling them with their concepts. It demands more fine-granular recognition of images, since pixel-wise classification is required along with understanding of various image contents. The restrictions of limited fine-annotated images and diversity of objects brings more challenges into this task.

In the past decades, numerous efforts have been paid on the task of semantic segmentation. According to the different levels of supervision, they can be roughly divided into two categories: fully-supervised and weakly-supervised. In the fully-supervised setting, CRF (Conditional Random Field) models are used typically and have lots of effective extensions [3]. Besides, the techniques of deep learning are applied to solve the problem of semantic segmentation effectively. Long *et.al* [8] proposed a fully convolutional network constrained by strong supervision of pixel labels and refined its results with a hierarchical structure. Although satisfactory segmentation performance of the fully-supervised solutions can be achieved when given a large amount of training data, the high cost on the pixel-level annotations is bound to restrict its extensive applications.

To alleviate the dependence on the fine-grained labeled data, weakly-supervised methods with only image-level labels available have emerged and attracted much attention [4, 6, 7]. Liu *et.al* [6] built a graph propagation model considering consistency of superpixels and weak supervision information simultaneously. Liu *et.al* [7] formulated the problem as a weakly-supervised dual clustering task to cluster superpixels and assign a suitable label to each cluster. And Xie *et.al* [15, 16] further improved the graph construction method of the model. Despite of largely decreasing requirement on training data, the weakly-supervised methods suffer a poor performance in discovering the object structure. To achieve the balance of the dependence on training data and the model performance, we turn to a semi- and weakly-supervised method in this paper, using mainly the image-level labeled images as training data and very few pixel-level labeled images for supplementary.

In this paper, we propose a two-stage Convolutional Neural Network (CNN) based framework to accomplish a coarse-to-fine learning process for image semantic segmentation. In the first stage, we build a pixel-level supervised network (PS-CNN), using only a small number of images with their pixel-level labels. The network predicts labels for each pixel of the input images and thus carries out a fine-grained learning on various object details. The prediction can indicate the differences between objects or object parts from different classes to some extent, but is too crude to ensure a satisfactory segmentation performance because of too limited training data explored in the first stage. Then, a collaborative-supervised

MM'15, October 26–30, 2015, Brisbane, Australia.
© 2015 ACM. ISBN 978-1-4503-3459-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2733373.2806322.

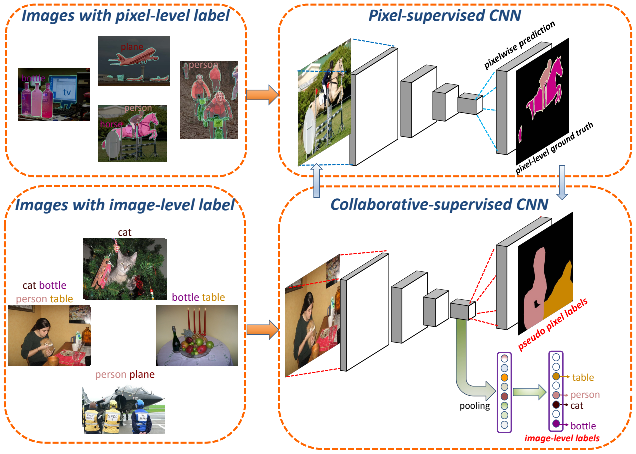

Figure 1: The overview of our approach. The upper and lower parts in the figure show the training process of the PS-CNN and CS-CNN respectively. During the training of CS-CNN, the response of its last convolutional layer is upsampled and averaged respectively and then supervised by the pixel-level and image-level labels collaboratively, which enhances the sensitivity of the model to both the object structures and the tiny objects.

network (CS-CNN) as the second stage is proposed to enhance the segmentation performance by making full use of two sides of information: one is the outputs of the first stage, the other is a large number of images associated with only image-level labels. Specifically, we build a two-head network to predict labels for each pixel and the entire image at the same time. Given a training image, the pseudo pixel labels obtained from the first stage and the image-level labels are used to supervise the two label prediction tasks respectively and the corresponding two loss functions are combined to contribute to updating the network. With such a collaborative-supervised network, we aim to refine the model by expanding the training scale with weakly labeled images, and better mask the object structure benefitting from the discriminative ability of the fully-supervised model learning with fully labeled images. Our experiments on two challenging datasets, i.e, PASCAL VOC 2007 dataset [1] and LabelMe LMO dataset [10], demonstrate the attractive performance of our approach, nearly matching the results achieved when all training images have pixel-level labels.

2. OUR APPROACH

Our approach consists of two stages of CNN training to progressively elaborate our model, of which one is a pixel-supervised CNN (PS-CNN) as a pre-trained model and the other is a collaborative-supervised CNN (CS-CNN) as a refined model. Both of the two networks are finetuned on the ImageNet-pretrained VGG 16-layer net [11], and the processes are implemented with Caffe [2]. An overview of our approach is shown in Figure 1.

2.1 Pixel-supervised CNN Model

In the first stage, we train a pixel-supervised CNN to learn the details of different objects, using few fully-supervised images. The model is built as a fully convolutional neural network referring to FCN [8]. The FCN model replaces the fully connected layers by convolutional ones with kernels covering their entire input regions, which further extends a convnet to adapt arbitrary-sized inputs and output a classification map. With this network structure, a feature map can be achieved from the last convolutional layer, of which each feature vector indicates the response of a receptive field in the input image. To connect the coarse output with the pixel-wise annotations, the feature map is then up-sampled with bilinear interpolation to the size of the input image . And classification is made on each pixel to make detailed segmentation.

The softmax loss function is used in this network for pixel-wise supervision and the averaged loss value of all the pixels in a training batch is used as the final loss. The loss function can be defined as follows,

$$Loss_1 = -\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{c} 1\{y_{ij} = k\} log(\hat{p}_{ij,k}^{pixel}) \qquad (1)$$

$$\hat{p}_{ij,k}^{pixel} = \frac{exp(x_{ij,k}^{pixel})}{\sum_{l=1}^{c} exp(x_{ij,l}^{pixel})} \qquad (2)$$

where n, m, c are the batchsize, pixel number in an image and the number of object classes, respectively. $\hat{p}_{ij,k}^{pixel}$ stands for the probability of the jth pixel in the ith image to be predicted to class k. x_{ij}^{pixel} stands for the feature vector of $pixel_{ij}$ given out by the last convolutional layer of the network and $x_{ij,k}^{pixel}$ is the value in its kth channel. y_{ij} is the groundtruth label of the $pixel_{ij}$ and $1\{y_{ij} = k\}$ is an indicator function judging whether the groundtruth label equals k.

2.2 Collaborative-supervised CNN Model

In this stage, more weakly-supervised images with only image-level labels are added to expand training scale and refine our model. To utilize them, we design a collaborative-supervised CNN model which takes both of the pixel-level and image-level information into consideration.

For the pixel-level loss term, we employ the same network structure as in the first stage and use pseudo pixel-level labels for the weakly-supervised images as supervision. The pseudo labels are pixel-wise prediction generated by the PS-CNN model achieved above. It is coarse because of the limited training data but already competent to distinguish most objects or their parts. So we use it here to maintain the discriminating ability of the network on different object details.

But as the pseudo pixel labels contain much noise and may be insensitive to tiny objects, we add an image-level loss term to alleviate the problem. We average the response of each receptive field as the response of the entire image and then make the image-level classification. As there are always more than one object in the image, we use KL-Divergence to formulate the loss function and assume that the prediction probability should distribute evenly on all the object classes existing in the image. That is, for an image containing q kinds of objects, we set $\frac{1}{q}$ as the groundtruth probability for the q classes while 0 for the others. In this way, we actually demand that every existing object should be given equal notice by the network. Therefore, some of the tiny objects or unconspicuous object parts will be forced to produce higher response to ensure that they are not ignored or mislabeled,

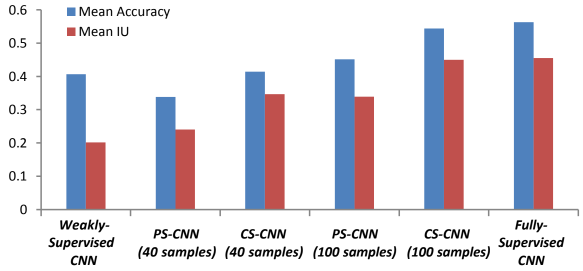

Figure 2: Performance of different CNN models on the PASCAL VOC 2007 dataset

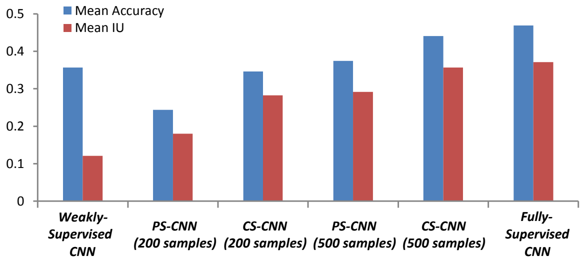

Figure 3: Performance of different CNN models on the LabelMe LMO dataset

which helps neutralizing the noise brought by the pseudo pixel labels. The loss function is as follows,

$$Loss_2 = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c} p_{i,k}^{image} log(\hat{p}_{i,k}^{image}) \tag{3}$$

$$\hat{p}_{i,k}^{image} = \frac{exp(x_{i,k}^{image})}{\sum_{l=1}^{c} exp(x_{i,l}^{image})} \tag{4}$$

where $\hat{p}_{i,k}^{image}$ denotes the probability of $image_i$ to be predicted to class k. And $p_{i,k}^{image}$ denotes the groundtruth probability. x_i^{image} is the averaged feature vector of $image_i$.

Finally, we combine the two loss terms with a weighting parameter α and use them to constrain our model collaboratively. The final loss function is as follows,

$$Loss = Loss_1 + \alpha Loss_2 \tag{5}$$

3. EXPERIMENT

Extensive experiments are performed on PASCAL VOC 2007 dataset [1] and LabelMe Outdoor dataset [10], which are focused on object and scene segmentation, respectively. Both mean class accuracy and mean IU (intersection over union) are used as metrics to evaluate our model. Our approach is compared with several fully- and weakly- supervised methods and outperforms the state-of-the-art methods on both two datasets, which indicates the effectiveness of our approach. The results on different kinds of segmentation tasks also confirm the universality of our approach.

3.1 PASCAL VOC 2007 Dataset

The PASCAL VOC 2007 dataset consists of 422 training-validation images and 210 testing images belonging to 20 object classes. During the training process, we randomly sample a small percentage of images to train the PS-CNN with their pixel-level labels and the rest images are further added to train the CS-CNN with only image-level labels. The impact of the pixel-level labeled images is further evaluated by changing their numbers from 40 ($< 10\%$) to 100 ($< 25\%$). And the evaluations when the whole training set is pixel-level labeled (fully-supervised CNN, using only the pixel-level supervision) or image-level labeled (weakly-supervised CNN, using only the image-level supervision) are also performed for comparison.

Comparison between different settings are shown in Figure 2. It shows that, with both sampling rates, the proposed CS-CNN model outperforms the PS-CNN by about 10 percent on both the two evaluation metrics. It indicates the effectiveness of our semi-supervised approach which enhances the model with weakly-labeled images. Moreover, comparison between the results achieved with different sampling rates indicates the influence of the fully-supervised training data. Our results get an obvious improvement when the number of pixel-level labeled samples rises from 40 to 100. While compared with the weakly-supervised and fully-supervised CNN, the CS-CNN model always shows better results than the weakly-supervised CNN, by the reason of the few fully-supervised images. And it should be noticed that our CS-CNN model trained with only 100 ($< 25\%$) pixel-level labeled images already achieves very similar performance with the fully-supervised CNN using the whole training set with pixel-level annotations, which further comfirms the effectiveness of our approach.

Finally, results of our approach (semi-supervised, indicated as SS) and other fully-supervised (FS) or weakly-supervised (WS) methods are compared in Table 1. Our results are achieved using the CS-CNN model trained with 100 pixel-level labeled images. It can be seen that our approach shows better accuracy and robustness, which wins the highest accuracy in 8 classes among all of these methods and outperforms the state-of-the-art method [16] by 7 percent on mean accuracy.

3.2 LabelMe Outdoor Dataset

This dataset is a subset of LabelMe dataset [10] provided by [5]. It contains 2688 fully annotated images of 33 categories, of which 2488 images are for training and the rest for testing. Most of the images are outdoor scenes including sky, buildings and mountain. We continue to use the same settings as above and conduct our experiments with 200 ($< 10\%$) and 500 ($< 25\%$) pixel-level samples, respectively. The results are shown in Figure 3, which further verifies the above conclusion that our semi-supervised approach effectively enhances the model by exploiting more weakly-supervised images and is able to achieve similar performance with the fully-supervised CNN with only few pixel-level samples. Comparison with the state-of-the-art methods are shown in Table 2. Our result outperforms even the fully-supervised methods and exceeds the state-of-the-art method [9] by 12 percent on mean accuracy.

4. CONCLUSION

In this paper, we propose a semi- and weakly- supervised learning method for image semantic segmentation, which consists of two stages of CNN training. In the first stage, we train a pixel-supervised CNN to learn the discrimination between objects with very few pixel-level labeled images. In

Table 1: Semantic segmentation results on PASCAL VOC 2007 dataset.

Supervision	Method	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motorbike	person	plant	sheep	sofa	train	tv	bkgd	mean
WS	[6]	28	20	52	28	46	41	39	60	25	68	25	35	17	35	56	36	46	17	31	20	65	38
	[18]	48	20	26	25	3	7	23	13	**38**	19	15	39	17	18	25	47	9	**41**	17	33	-	24
	[17]	65	25	39	8	17	38	17	26	25	17	47	41	44	32	59	34	36	23	35	31	-	33
	[15]	**85**	55	**87**	45	42	31	34	57	21	**81**	23	16	6	11	42	31	**72**	24	49	40	41	42
	[16]	77	48	**87**	**50**	**56**	48	44	60	27	76	18	38	25	31	52	38	59	31	51	34	41	47
FS	[3]	27	33	44	11	14	36	30	31	27	6	50	28	24	38	52	29	28	12	45	46	-	30
	[14]	19	21	5	16	3	1	**78**	1	3	1	23	**69**	44	42	0	**65**	30	35	**89**	**71**	-	31
SS	Ours	53	**66**	11	40	46	**69**	46	**94**	15	41	**67**	51	**56**	**74**	**87**	19	62	26	84	43	**93**	**54**

Table 2: Semantic segmentation results on LabelMe LMO dataset.

Supervision	FS			WS		SS
Method	[5]	[12]	[9]	[13]	[7]	Ours
Accuracy	24	29	32	21	26	**44**

the second stage, a collaborative-supervised CNN is designed to jointly perform the pixel-level and image-level classification tasks, given a large number of images with only image-level labels available. For the collaborative-supervised model training, the pixel-level labels are pseudo ones predicted by the fully-supervised network in the first stage. Our results on two commonly used datasets have proved that our model can achieve a result close to or better than the fully-supervised methods with the help of only a small amount of pixel-level labeled images. Experiments on larger external datasets will be implemented in the future to further confirm the effectiveness and scalability of our method.

5. ACKNOWLEGMENTS

This work was supported by 973 Program (2012CB316304) and National Natural Science Foundation of China (61332016, 61272329, and 61472422).

6. REFERENCES

[1] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[2] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[3] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, pages 739–746, 2009.

[4] Y. Li, J. Liu, Y. Wang, H. Lu, and S. Ma. Weakly supervised rbm for semantic segmentation. In *IJCAI*, 2015.

[5] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *CVPR*, pages 1972–1979, 2009.

[6] S. Liu, S. Yan, T. Zhang, C. Xu, J. Liu, and H. Lu. Weakly supervised graph propagation towards collective image parsing. *TMM*, 14(2):361–373, 2012.

[7] Y. Liu, J. Liu, Z. Li, J. Tang, and H. Lu. Weakly-supervised dual clustering for image semantic segmentation. In *CVPR*, pages 2075–2082, 2013.

[8] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *arXiv preprint arXiv:1411.4038*, 2014.

[9] H. Myeong and K. M. Lee. Tensor-based high-order semantic relation transfer for semantic scene segmentation, 2013.

[10] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *IJCV*, 77(1-3):157–173, 2008.

[11] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[12] J. Tighe and S. Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. In *ECCV*, pages 352–365, 2010.

[13] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly supervised structured output learning for semantic segmentation. In *CVPR*, pages 845–852, 2012.

[14] V. Viitaniemi and J. Laaksonen. Evaluation of techniques for image classification, object detection and object segmentation (technical report tkk-ics-r2). http://www.cis.hut.fi/projects/cbir/, 2008.

[15] W. Xie, Y. Peng, and J. Xiao. Semantic graph construction for weakly-supervised image parsing. In *AAAI*, 2014.

[16] W. Xie, Y. Peng, and J. Xiao. Weakly-supervised image parsing via constructing semantic graphs and hypergraphs. In *MM*, pages 277–286, 2014.

[17] K. Zhang, W. Zhang, S. Zeng, and X. Xue. Semantic segmentation using multiple graphs with block-diagonal constraints. In *AAAI*, 2014.

[18] K. Zhang, W. Zhang, Y. Zheng, and X. Xue. Sparse reconstruction for weakly supervised semantic segmentation. In *IJCAI*, pages 1889–1895, 2013.