Semantic Segmentation based on Stacked Discriminative Autoencoders and Context-Constrained Weakly Supervised Learning

Xiwen Yao, Junwei Han*, Gong Cheng, Lei Guo
School of Automation
Northwestern Polytechnical University
127 West Youyi Road, Xi'an Shaanxi, 710072, P. R. China
{yaoxiwen517, junweihan2010, chenggong1119}@gmail.com, lguo@nwpu.edu.cn
* Corresponding author

ABSTRACT
In this paper, we focus on tacking the problem of weakly supervised semantic segmentation. The aim is to predict the class label of image regions under weakly supervised settings, where training images are only provided with image-level labels indicating the classes they contain. The main difficulty of weakly supervised semantic segmentation arises from the complex diversity of visual classes and the lack of supervision information for learning a multi-classes classifier. To conquer the challenge, we propose a novel discriminative deep feature learning framework based on stacked autoencoders (SAE) by integrating pairwise constraints to serve as a discriminative term. Furthermore, to mine effective supervision information, global context about co-occurrence of visual classes as well as local context around each image region is exploited as constraints for training a multi-class classifier. Finally, the classifier training is formulated as an ultimate optimization problem, which can be solved efficiently by an alternate iterative optimization method. Comprehensive experiments on the MSRC 21 dataset demonstrate the superior performance compared with several state-of-the-art weakly supervised image segmentation methods.

Categories and Subject Descriptors
I.4.6 [**Image Processing and Computer Vision**]: Segmentation –*pixel classification;* I.4.6 [**Image Processing and Computer Vision**]: Feature Measurement–*feature representation;*

General Terms
Algorithms, Experimentation, Performance.

Keywords
Semantic segmentation; Stacked autoencoders; Discriminative feature learning; Weakly supervised learning.

MM '15, October 26-30, 2015, Brisbane, Australia
 ISBN 978-1-4503-3459-4/15/10…$15.00
DOI: http://dx.doi.org/10.1145/2733373.2806319

1. INTRODUCTION
Weakly supervised semantic segmentation is a fundamental yet challenging task to segment an image into several regions of homogeneous texture or color and simultaneously recognize their associated semantic categories under weakly supervised settings, in which only image-level labels are provided for each training image, specifying the classes present in the image. Recently, many efforts have been contributed to this study [1-6].

How to describe the intrinsic representation of regions is the key to the success of semantic segmentation. HoG and SIFT features are exploited in [4, 6, 7]. However, such low-level features often fail to offer sufficient discriminative power. We argue that the usage of more semantically abstract features, such as high-level features learned by recent developed deep learning methods, may offer a promising venue to empower the model to handle intrinsically diverse and complex visual classes in natural images [8] and also remote sensing images [9], and hence to improve segmentation performance. In this paper, we propose a novel discriminative deep feature learning framework based on stacked autoencoders by integrating pairwise constraints to serve as a discriminative term. The pairwise constraints, including must-link constraints (indicating the semantic similarity between superpixels) and cannot-link constraints (indicating the semantic dissimilarity between superpixels), guarantee the ability of discriminative representation by keeping the superpixels connected by the must-link be close to each other in the learned feature space while ensuring the superpixels of cannot-link to be kept far away.

In contrast to full-supervised methods [10, 11], it is more challenging for weakly supervised segmentation methods to train region-level classifier with only image-level labels in the training set. Since there are no ground truth labels of superpixels in the training set, it is necessary to mine effective constraint information to supervise the classifier training. In [4, 12, 13], image-superpixel label inclusion information is exploited to directly infer the superpixel labels. However, only the label inclusion information does not effectively deal with regions that of similar appearances corresponding to distinct semantic concepts. For example, a smooth region in blue may be a part of sky or a part of water. It is even difficult for human observers to individually classify such regions without context. Contextual information plays a very important role to reduce semantic ambiguity. In this paper, we propose to incorporate both global and local semantic contextual constraint information, together

with label inclusion information in a unified framework to provide supervision for training the multi-class classifier. The global semantic context describes the co-occurrence relationships of different concepts while the local context encourages neighboring superpixels with similar appearances within the same image to share the same concept label.

We evaluate the proposed approach on MSRC-21 dataset [14]. Experimental results show that the proposed algorithm outperforms state-of-the-art weakly supervised image segmentation methods, and its performance is even comparable to those of the fully supervised segmentation models.

2. THE PROPOSED METHOD

We start by setting up the problem and notations. Let $\mathrm{X}=\{X_1,...X_m,...X_M\}$ be the training set with corresponding image-level labels $G=[G_1,...G_m,...G_M]$, where $G_m=[g_m^1,...g_m^c,...g_m^C]^{\mathrm{T}}$ is the label indicator vector with $g_m^c=1$ if the image containing class c and $g_m^c=0$ otherwise. We oversegment an image X_m into n_m superpixels by an over-segmentation algorithm [15] and obtain the low-level feature data matrix $X_m=[x_{m,1},...x_{m,i},...x_{m,n_m}]$, where $x_{m,i}$ is the feature of ith superpixel and the associated label is denoted as $y_{m,i}=[y_{m,i}^1,...y_{m,i}^c,...y_{m,i}^C]$ with $y_{m,i}^c=1$ if $x_{m,i}$ is assigned label c and $y_{m,i}^c=0$ otherwise. For brevity, we denote $X=[x_1,...x_n,...x_N]$ as the training set and $Y=[y_1,...y_n,...y_N]$ as the corresponding label matrix, where $N=\sum_{m=1}^{M}n_m$ is the total number of superpixels in the training set. The task of weakly supervised semantic segmentation is to recover the latent label Y and simultaneously learn a classifier, which later help to predict superpixel labels in the new test images.

2.1 Stacked Discriminative Autoencoders

At the encoding part of autoencoder [16], an input vector x is mapped to the hidden representation h by a linear deterministic mapping and a nonlinear activation function $h=f(W_1x+b_1)$, where W_1 is an encoding weight matrix, b_1 is encoding bias vector and $f(z)$ is the logistic sigmoid function. At the decoding part, hidden feature representation h is mapped back to a reconstruction $\hat{x}$ through $\hat{x}=f(W_2h+b_2)$, where W_2 is an decoding weight matrix, b_2 is decoding bias vector. Weight matrices W_1 and W_2 and bias vectors b_1 and b_2 are learned by minimizing the cost function:

$$J=\frac{1}{2}\sum_{i=1}^{N}\|x_i-\hat{x}_i\|_2^2+\frac{\lambda}{2}(\|W_1\|_2^2+\|W_2\|_2^2) \tag{1}$$

We constructed must-link constraints P by selecting two superpixels with similarity larger than ξ from different images sharing common labels in condition that they are the most similar to each other in their respective images. We constructed cannot-link constraints D by simply selecting two superpixels with similarity larger than ξ from different images sharing no common labels.

To introduce pairwise constraints into autoencoders as a discriminative term, the learned latent features (h_i,h_j) of P are encouraged to be close in the learned latent feature space and of D are ensured to be kept far away, which can be formulated as minimizing the cost function:

$$J_{DAE}=J+\frac{\beta}{2}\sum_{i,j=1}^{N}(\|h_i-h_j\|_2^2F_{ij}) \tag{2}$$

Parameter β controls the contribution of pairwise constraints, which are encoded in matrix F defined as:

$$F_{ij}=\begin{cases}+1, & \text{if}(x_i,x_j)\in \mathrm{P}\\ -1, & \text{if}(x_i,x_j)\in \mathrm{D}\\ 0, & \text{otherwise}\end{cases} \tag{3}$$

The gradient of the objective function J_{DAE} with respect to $\{W_1,W_2,b_1,b_2\}$ can be computed as follows:

$$\begin{aligned}\frac{\partial J_{DAE}}{\partial W_1}&=\frac{1}{N}\sum_{i=1}^{N}\Delta_i^{(1)}x_i^{\mathrm{T}}+\lambda W_1\\&+\beta\sum_{i,j=1}^{N}(((h_i-h_j)\circ f'(z_i^{(1)}))x_i^{\mathrm{T}}-((h_i-h_j)\circ f'(z_j^{(1)}))x_j^{\mathrm{T}})F_{ij}\end{aligned} \tag{4}$$

$$\frac{\partial J_{DAE}}{\partial W_2}=\frac{1}{N}\sum_{i=1}^{N}\Delta_i^{(2)}h_i^{\mathrm{T}}+\lambda W_2 \tag{5}$$

$$\frac{\partial J_{DAE}}{\partial b_1}=\frac{1}{N}\sum_{i=1}^{N}\Delta_i^{(1)}+\beta\sum_{i,j=1}^{N}((h_i-h_j)\circ(f'(z_i^{(1)})-f'(z_j^{(1)})))F_{ij} \tag{6}$$

$$\frac{\partial J_{DAE}}{\partial b_2}=\frac{1}{N}\sum_{i=1}^{N}\Delta_i^{(2)} \tag{7}$$

where the operation $\circ$ denotes the element-wise multiplication, and $\Delta_i^{(1)}$ and $\Delta_i^{(2)}$ are defined as:

$$\Delta_i^{(1)}=W_2^{\mathrm{T}}\Delta_i^{(2)}\circ f'(z_i^{(1)}),\quad \Delta_i^{(2)}=-(x_i-\hat{x}_i)\circ f'(z_i^{(2)}) \tag{8}$$

Then the parameters $\{W_1,W_2,b_1,b_2\}$ can be learned by using the gradient descent method with learning rate μ.

However, due to the simple shallow structural characteristic, the representational power of a single layer discriminative autoencoder is limited. In this paper, we propose a stacked discriminative autoencoders (SDAE), in which DAE is used as a building block to learn discriminative high-level features from low-level ones as described in Section 3.1. The training of SDAE is performed in a greedy layer-wise learning manner introduced by Hinton *et al.*[17].

2.2 Context-constrained Weakly Supervised Learning

2.2.1 Constraint information

To perform segmentation, a neural network classifier is constructed by adding an additional classification layer on the top layer of SDAE described in the above subsection. Since there are not pixel-level labels, the following constraints information is taken into consideration to train the classifier.

2.2.1.1 Label inclusion constraint

The label inclusion constraint guarantees that there are no superpixels supporting an invalid label. That is, given an image X_m with image-level label $G_m=[g_m^1,...g_m^c,...g_m^C]^{\mathrm{T}}$ and its ith

superpixel label $y_{m,i}=[y_{m,i}^1,...y_{m,i}^c,...y_{m,i}^C]$, $max\, y_{m,i}^c$ should be close to zero, which is equal to minimize the objective function: $C_1=\sum_m\sum_c(1-g_m^c)\max_{x_{m,i}\in X_m} y_{m,i}^c$. We further give its matrix form $C_1=\sum_m\sum_c(1-g_m^c)h_c^{\mathrm{T}}Y^{\mathrm{T}}q_m$, where $h_c\in \mathrm{R}^C$ is an indicator vector with its all elements expect for the cth element are zeros. $q_m\in \mathrm{R}^N$ is a vector with its all elements except for those elements corresponding to the mth image are zeros.

2.2.1.2 Context constraints

We use the pairwise co-occurrence of concepts to capture the global context constraint. The element of reference co-occurrence matrix $A=\{a_{c_1c_2}\}\in \mathrm{R}^{C\times C}$ is defined as the conditional probability of coincidence of concepts. To integrate the global contextual information into formulation, we introduce the following objective function:

$$C_{gc}=\frac{1}{2}\sum_{c_1,c_2}^{C}(A_{c_1c_2}\,\|\,y_{c_1}^{\mathrm{T}}-y_{c_2}^{\mathrm{T}}\,\|_2^2)=Tr(YL_{gc}Y^{\mathrm{T}}) \tag{9}$$

where Laplacian matrix $L_{gc}=D_{gc}-A$ and D_{gc} is a diagonal matrix whose diagonal elements are the sums of the row elements of co-occurrence matrix A.

The local context can yield more semantically consistent segmentation results by imposing neighboring superpixels with similar appearances within the same image to share the same concept label. To model the local semantic context, we have the following equation to be minimized:

$$C_{lc}=\sum_{i,j=1}^{N}V_{ij}\,\|\,y_i-y_j\,\|^2=Tr(Y^TL_{lc}Y) \tag{10}$$

where V encodes the interactions of superpixels in the adjacent set A, which is defined as:

$$V_{ij}=\begin{cases}\exp(-\|h_i-h_j\|^2/\sigma^2), & \text{if}(h_i,h_j)\in \mathrm{A}\\ 0, & \text{otherwise}\end{cases} \tag{11}$$

Laplacian matrix $L_{lc}=D_{lc}-V$ and D_{lc} is a diagonal matrix whose diagonal elements are the sums of the row elements of matrix V.

2.2.2 Weakly supervised learning

Jointly considering the above constraints, the classifier training is formulated as optimizing the following objective function:

$$\min_{W_{nn},Y} J_{NN}(Y,W_{nn})+C(Y)\;\; s.t.\, Ye_1=e_2, Y\ge 0 \tag{12}$$

where $J_{NN}(Y,W_{nn})=\|\,h_{W_{nn}}(X)-Y\,\|_{\mathrm{F}}^2+\frac{\lambda}{2}\|\,W_{nn}\,\|_2^2$ is the loss function of neural network classifier with output $h_{W_{nn}}(X)$.

$C(Y)=\gamma_{gc}Tr(YL_{gc}Y^{\mathrm{T}})+\gamma_{lc}Tr(Y^{\mathrm{T}}L_{lc}Y)+\alpha\sum_m\sum_c(1-g_{\mathrm{m}}^{\mathrm{c}})h_{\mathrm{c}}^{\mathrm{T}}Y^{\mathrm{T}}q$ is the constraint information. Constraint $Ye_1=e_2$ is introduced to ensure the sum of each row in Y is equal to 1 with $e_1=\mathbf{1}_{\mathrm{C}\times 1}$ and $e_2=\mathbf{1}_{\mathrm{N}\times 1}$.

The minimization problem of (12) can be solved in the following two alternate optimization steps:

$$W_{nn}^*=\underset{W_{nn}}{argmin}\,J_{NN}(Y,W_{nn})=\underset{W_{nn}}{argmin}\,\|\,h_{W_{nn}}(X)-Y^*\,\|_{\mathrm{F}}^2+\frac{\lambda}{2}\|\,W_{nn}\,\|_2^2 \tag{13}$$

$$Y^*=\underset{Y\ge 0}{argmin}\,\|\,h_{W_{nn}^*}(X)-Y\,\|_{\mathrm{F}}^2+C(Y)+\delta\,\|\,Ye_1-e_2\,\|_{\mathrm{F}}^2\;\; s.t.\, Y\ge 0 \tag{14}$$

The first subproblem is a standard neural network classifier training problem given Y^*, which can be solved by using back-propagation algorithm [18]. The second subproblem can be effectively solved by a nonnegative multiplicative updating procedure [19].

Therefore, considering the two alternate optimization steps together, after convergence, we can obtain superpixel labels Y and neural network classifier W_{nn}.

3. EXPERIMENT RESULTS

3.1 Experiment Setup

We evaluated our approach on the MSRC-21 dataset. This dataset contains 591 images of 21 visual classes with manually labeled object segmentation ground-truth [14]. Pixels on the boundaries of objects are usually labeled as background and are ignored during training and evaluation. For both datasets, we adopted the over-segmentation algorithm in [15] to obtain superpixels and described superpixel appearances with 980-dimimension low-level features including shape, location, histogram of texture, SIFT, and color. The proposed SDAE consists of 3 layers of discriminative autoencoders, which take the 980-dimimension low-level features as input and set the number of the first and hidden layer as 1960 and 1960, respectively and the number of the output layer is set to be 720. The performance was evaluated by average per-class accuracy, which measures the percentage of correctly classified pixels for an object class.

3.2 Parameters Analysis

In the implementation of SDAE, there are four parameters to be set: threshold ξ for pairwise constraints construction, weight decay cost parameter λ, discriminative parameter β and learning rate μ. For parameters λ and μ, we empirically set $\lambda=0.001$, $\mu=0.2$ based on the practical tricks introduced in book [20]. For parameters β and ξ, we varied the value of parameter ξ from 0.71 to 0.80 with a stride of 0.01 and set β with $\{0.001,\ 0.01,\ 0.1\}$. As shown in Fig.1 (a), we set $\xi=0.72$ and $\beta=0.001$ with the average accuracy reaching the peak point. In the classifier learning, parameters α, δ, γ_{gc} and γ_{lc} need to be set. Parameters α and δ are both set to be 1000 which is large enough to guarantee the label consistent constraint satisfied and to ensure the sum of each row in is equal to 1. Fig.1 (b) presents the accuracy with different values of γ_{gc} and γ_{lc}. Finally, we set $\gamma_{gc}=10$ and $\gamma_{lc}=100$.

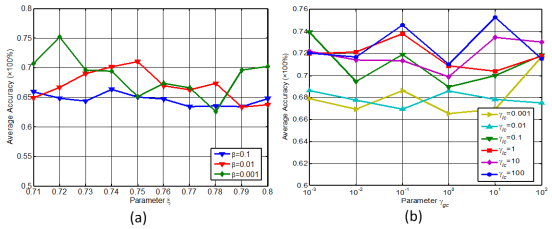

Fig.1 Average accuracy with different parameters.

3.3 Comparison with State-of-the-arts

Fig.2 shows some example results produced by our method in comparison with the ground-truth. Table 1 reports the average accuracy over all 21 object classes in comparison with state-of-the-art semantic segmentation methods, including both fully supervised (FS) [10, 11] and weakly supervised (WS) [4, 5]. As shown in Table 1, our method outperforms all other weakly supervised approaches and meanwhile is comparable to those of the fully supervised segmentation methods. Fig. 3 presents the per-class accuracy results on MSRC dataset. The class labels from 1 to 21 are assigned with building, grass, tree, cow, sheep, sky, aeroplane, water, face, car, bicycle, flower, sign, bird, book, chair, road, cat, dog, body and boat, respectively. Our method gets the best results on 10 out of 21 classes and especially works well on several confusing categories, such as aeroplane and boat, sky and water.

Table 1. Average accuracy (%) of our method and four comparison methods on MSRC-21 dataset.

	FS		WS		
Methods	[10]	[11]	[5]	[4]	Ours
Accuracy	75	76	69	71	**76**

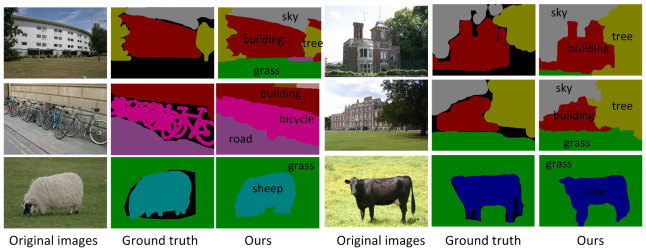

Fig. 2 Some results of our method on MSRC-21 dataset.

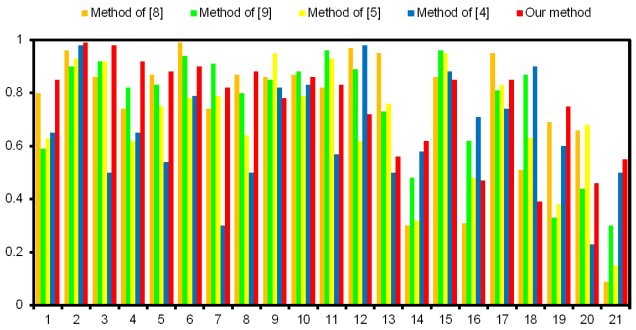

Fig.3 Per-class accuracy of our method and four comparison methods on MSRC-21 dataset.

4. CONCLUSIONS

In this paper, we proposed a coherent framework to perform semantic segmentation under weakly supervised settings. The framework includes two parts: a novel discriminative deep feature learning method based on stacked autoencoders with pairwise constraints and a context-constrained weakly supervised multi-class classifier learning method by employing both global and local semantic contextual information. Experiments on MSRC-21 dataset demonstrated the effectiveness of our proposed method, compared with several state-of-the-art full supervised and weakly supervised methods.

5. ACKNOWLEDGMENTS

This work was supported by the National Science Foundation of China under Grant 91120005 and 61473231, and Doctoral Fund of Ministry of Education of China under grant 20136102110037.

6. REFERENCES

[1] A. Vezhnevets, *et al.*, "Weakly supervised semantic segmentation with a multi-image model," in *ICCV*, 2011, pp. 643-650.
[2] A. Vezhnevets, *et al.*, "Active learning for semantic segmentation with expected change," in *CVPR*, 2012, pp. 3162-3169.
[3] A. Vezhnevets, *et al.*, "Weakly supervised structured output learning for semantic segmentation," in *CVPR*, 2012, pp. 845-852.
[4] Y. Liu, *et al.*, "Weakly-Supervised Dual Clustering for Image Semantic Segmentation," in *CVPR*, 2013, pp. 2075-2082.
[5] K. Zhang, *et al.*, "Sparse reconstruction for weakly supervised semantic segmentation," in *IJCAI*, 2013, pp. 1889-1895.
[6] Y. Liu, *et al.*, "Boosted MIML method for weakly-supervised image semantic segmentation," *Multimedia Tools and Applications,* pp. 1-17, 2014.
[7] L. Zhang, *et al.*, "Probabilistic graphlet cut: Exploiting spatial structure cue for weakly supervised image segmentation," in *CVPR*, 2013, pp. 1908-1915.
[8] R. Girshick, *et al.*, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.
[9] J. Han, *et al.*, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.,* vol. 53, pp. 3325-3337, 2015.
[10] L. Ladicky, *et al.*, "Associative hierarchical crfs for object class image segmentation," in *ICCV*, 2009, pp. 739-746.
[11] A. Lucchi, *et al.*, "Structured image segmentation using kernelized features," in *ECCV*, 2012, pp. 400-413.
[12] S. Liu, *et al.*, "Weakly supervised graph propagation towards collective image parsing," *IEEE Trans. Multimedia,* vol. 14, pp. 361-373, 2012.
[13] W. Xie, *et al.*, "Weakly-Supervised Image Parsing via Constructing Semantic Graphs and Hypergraphs," in *ACM MM*, 2014, pp. 277-286.
[14] J. W. Shotton, *et al.*, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *ECCV*, 2006.
[15] G. Mori, "Guiding model search using segmentation," in *ICCV*, 2005, pp. 1417-1423.
[16] H.-C. Shin, *et al.*, "Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data," *TPAMI,* vol. 35, pp. 1930-1943, 2013.
[17] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science,* vol. 313, pp. 504-507, 2006.
[18] M. T. Hagan and M. B. Menhaj, "Training feedforward networks with the Marquardt algorithm," *IEEE Trans. on Neural Networks,* vol. 5, pp. 989-993, 1994.
[19] X. Liu, *et al.*, "Unified solution to nonnegative data factorization problems," in *ICDM*, 2009, pp. 307-316.
[20] G. B. Orr and K.-R. Müller, *Neural networks: tricks of the trade*: Springer, 2003.