LaTeX Author Guidelines for CVPR Proceedings

Anonymous CVPR submission

Paper ID ****

Abstract

The ABSTRACT is to be in fully-justified italicized text, at the top of the left-hand column, below the author and affiliation information. Use the word "Abstract" as the title, in 12-point Times, boldface type, centered relative to the column, initially capitalized. The abstract is to be in 10-point, single-spaced type. Leave two blank lines after the Abstract, then begin the main text. Look at previous CVPR abstracts to get a feel for style and length.

1. Introduction

Aiming to assign each pixel in an image to one of predefined semantic categories, semantic segmentation is an attractive but challenging task in computer vision community. In the past few years, many different methods [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**] have been proposed for this task. Notwithstanding significant improvements they have achieved, most of them rely on full supervision: each pixel of the image for training is manually labeled by humans. Considering this kind of annotation is time-consuming and tedious, fully supervised methods cannot be widely applied in practice.

Recently, a few works have been proposed to address the semantic segmentation problem under the weakly supervised settings, where only the image-level annotations are available in the training process [**?**, **?**, **?**, **?**, **?**, **?**]. Comparing to the trandtional supervised semantic segmentation, such weakly supervised method is more flexible in real-world applications for the image-level annotated images are much easier to obtain. However, there are some extra constraints (*e.g*. labels must be precise and complete) for the inital image-level labels in weakly supervised semantic segmentation. Collecting the training images that satisfy all these constraints is still a labor-intensive task. Fortunately, owing to the collaborative image tagging system, *e.g*. Flickr, we can easily obtain a large mount of manually labeled images provided by Internet users, though these image-level labels might be noisy (incorrect or incomplete). Therefore, the main challenge lies in how to utilize the noisily labeled images for semantic segmentation (see Fig. **??** for an illustration).

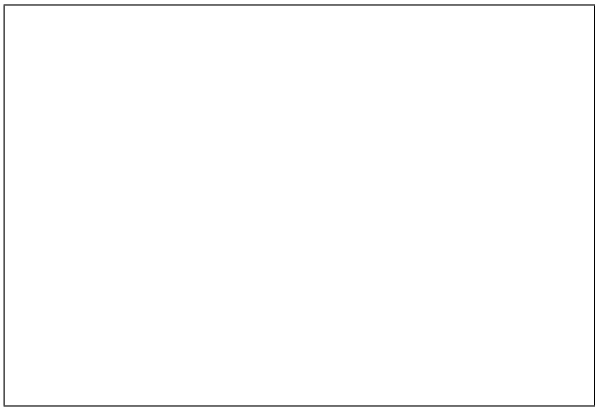

Figure 1. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

Moreover, most existing semantic segmentation methods, either fully or weakly supervised, depend on a single choice of image partitioning (quantization). The precise quantization of an image is of significance, and it is less likely to obtain a common optimal quantization (partitioning) level suitable for every object. To overcome this problem, [**?**, **?**, **?**, **?**, **?**] used multiple segmentations of the image and achieved good performances by heuristic strategies or enforcing label consistency with higher order potential.

In this paper,

2. Related Work

3. The Proposed Model

Assume we have a set of weakly labeled images and each image is oversegmented into several superpixels. We formulate this weakly supervised semantic segmentation as a joint learning problem which we factor into multi-class and binary CRFs.

...

Suppose we have a set of weakly labeled images $\mathcal{I} =$

$\{I^k\}_{k=1}^M$ and each image is oversegmented into m_k superpixels $\mathrm{X}^k = \{x_i^k\}_{p=1}^{m_k}$ by [**?**]. We describe each superpixel x_i^k by its appearance model and topic model (see Sec. **??** for details). Let $\boldsymbol{y}^k = (y_1^k, ..., y_L^k)^{\mathrm{T}}$ denote a vetoer of the L binary label variables, *i.e.* $y_i^k \in \{0,1\}$, where $y_i^k = 1$ indicates that category i is present in image k, and 0 otherwise. For each superpixel x_p^k, we define a random variable $h_p^k \in \{1, ..., L\}$ to represent its semantic category.

Our goal is to find an optimal label configuration that ... To tackle this problem, we build a conditional random field (CRF) on the image-level label variables $\boldsymbol{y}$ and the superpixel variables $\boldsymbol{h}$. We connect each superpixel variables to its neighbors to encode a local smoothness constraint. Specifically, let $\mathcal{E}$ donate the superpixel neighborhood, we define an energy function E with five types of potential as follow:

$$\begin{aligned} E(\boldsymbol{y}, \boldsymbol{h}, I) = &\sum_{i=1}^{L} \psi_G(y_i, I) + \sum_{1 \le i,j \le L} \psi_R(y_i, y_j) \\ &+ \sum_{p=1}^{m} \psi_{at}(h_p, x_p) + \sum_{(p,q)\in\mathcal{E}} \psi_S(h_p, h_q) \\ &+ \psi_C(\boldsymbol{y}, \boldsymbol{h}) \end{aligned} \tag{1}$$

where ψ_G and ψ_{at} encode the unary potential of global and regional constraints respectively, ψ_R impose labels' correlation and co-occurrence, ψ_S are the spatial context constraints for each superpixel, and ψ_C ensure the consistency between global and regional labels. The details of each potential will be described in the following sections. The posterior distribution $P(\boldsymbol{y}, \boldsymbol{h}|I)$ of the CRF can be written as $P(\boldsymbol{y}, \boldsymbol{h}|I) = \frac{1}{Z(I)} \exp\{-E(\boldsymbol{y}, \boldsymbol{h}, I)\}$, where $Z(I)$ is the normalizing constant. Thus, the most probable labelling configuration $\boldsymbol{y}^\star, \boldsymbol{h}^\star$ of the random field can be defined as $\boldsymbol{y}^\star, \boldsymbol{h}^\star = \arg\min_{\boldsymbol{y},\boldsymbol{h}} E(\boldsymbol{y}, \boldsymbol{h}, I)$.

3.1. Label Prediction

We utilize a label classifier which leverages convolutional neural network in order to predict missing label at test time. In particular, we extract a 4296 dimensional feature vector for each image by concatenating appearance feature and topic distribution. The 4296 dimensional feature vector includes: the second to last layer of convolutional neural network pre-trained on ImageNet, as the appearance feature, and topic distribution which learned from pLSA, as topic distribution.

Unfortunately, the appearance model cannot be trained due to the fact that the assignment of superpixels to semantic labels is unknown, even at training time. As far as we concerned, fine-tuning a discriminatively pre-trained network is very effective in terms of task performance.

We model each images as a finite mixture of latent semantic concepts by probabilistic latent semantic analysis(pLSA), which can recover visual models of semantic labels in a completely unsupervised manner. Probabilistic latent semantic analysis (pLSA) is a probabilistic model that is well suited to weakly supervision. Each image has its own mixing proportions whereas the topics are shared by all images. In document analysis, the pLSA usually takes the histogram of occurrence frequency on words as input. Here we regard fully-connected layer as input when we consider each neuron as a visual word. We denote each visual word(neuron) as w_i, then the occurrence frequency of image x_j on w_i is the i-th dimension of x_j. In addition, there is a hidden semantic topic variable t_k associated with all the visual words. We treat each topic t_k as a latent category in a semantic label. The pLSA optimizes the joint probability $P(w_i, x_j, t_k)$. Marginalizing over the latent category t_k determines the conditional probability $P(w_i|t_k)$:

$$P(w_i|x_j) = \sum_{k=1}^{K} P(t_k|x_j) P(w_i|t_k) \tag{2}$$

where $P(t_k|x_j)$ is the probability of topic t_k occurring in image x_j.

3.2. Label Consistency

We require that the superpixel labels be consistent with the image labels: if any superpixel x_p takes the label i, then image label indicator $y_i = 1$; otherwise $y_i = 0$. Such constraints can be encode by the following potential:

$$\psi_C(\boldsymbol{y}, \boldsymbol{h}) = C \cdot \sum_{i,p} I(y_i = 0 \text{ and } h_p = i) \tag{3}$$

where $I(\cdot)$ is the indicator function and C is a positive constant that penalizes any inconsistency between the global and local labels.

3.3. Appearance Model and Topic Model

We include both appearance and topic model as follow:

$$\begin{aligned} \psi_{at}(h_p, x_p) = -\log \{ &w_1 \phi_a(h_p, a_p, \theta_a) \\ &+ w_2 \phi_t(h_p, t_p, \theta_t) \} \end{aligned} \tag{4}$$

where a_p, t_p are the appearance and topic feature vectors extracted from the superpixels, θ_a, θ_t donate the parameters with repect to appearance model and topic model, $\{w_i\}|_{i=1}^2$ are the weighting coefficients for the unary terms. We define the appearance model $\phi_a(h_p, a_p, \theta_a) = f_{h_p}(a_p, \theta_a)$ and topic model $\phi_t(h_p, t_p, \theta_t) = g_{h_p}(t_p, \theta_t)$ measuring how well the local appearance a_p and topic t_p matches the semantic label h_p.

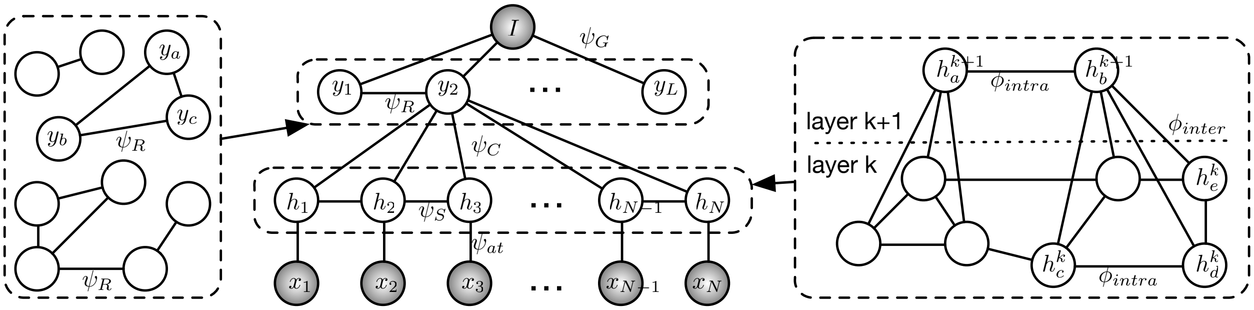

Figure 2. Example of a short caption, which should be centered.

3.4. Spatial Constraints and Hierarchical model

We

$$\psi_S(h_p,h_q)=\begin{cases}\phi_{inter}(h_p,h_q) & \text{if } |l_p-l_q|=1,\\ \phi_{intra}(h_p,h_q) & \text{if } l_p=l_q,\\ 0 & \text{otherwise}\end{cases}\qquad(5)$$

where l_p indicates the quantization level that the superpixel x_p belongs to. The inter-level energy cost ϕ_{inter} is defined as:

$$\phi_{inter}(h_p,h_q)=\gamma\cdot O(x_p,x_q)\cdot I(h_p\neq h_q)\qquad(6)$$

where $O(x_p,x_q)$ refers to the intersection (overlapping area) of two superpixels, $I(\cdot)$ is the indicator function and γ is the weighting coefficient. This formulation is based on the higher order constraints [**?**, **?**] that superpixels lying within the same clique are more likely to take the same label. And the intra-level energy cost ϕ_{intra} is defined as:

$$\phi_{intra}(h_p,h_q)=Sim(x_p,x_q)\cdot(1-R(h_p,h_q))\qquad(7)$$

where $Sim(x_p,x_q)\in[0,1]$ measures the visual similarity between superpixel x_p and x_q, $R(h_p,h_q)\in[0,1]$ is a learnt correlation between label h_p and h_q. Hence, we pay a high cost for the similar superpixels if they were assigned different labels and for the superpixels which were assigned an irrelevant label to the context.

3.5. Label Correlation and Co-occurrence

Different from the supervised semantic segmentation, the pixel-level label is not given. It is challenging because large appearance variations in cluttered backgrounds. However, the context contain some useful latent information which can be learned for semantic label noise reduction. Due to the unknown label of superpixels, learning these latent information is an unsupervised learning problem.

Multiple labels do not appear independently but occur correlatively and usually interact with each other at semantic space. The inter-label correlation matrix is constructed to characterize the interdependency between semantic concepts and helps to model the inter-label co-occurrence among feature space of superpixels.

The key idea is to analyze the change in the classification scores when artificially blackout different regions of the image. We observe that blackout a region that contains an discriminative region causes a massive confusion in cluster condition. This produces for each image a set of sub-windows from segmentation that are deemed likely to contain the discriminative region for specific semantic label. After localizing discriminative region in cluster condition, we can construct the inter-label correlation matrix by using both the available image-level label and region-level overlap.

3.6. Joint Inference with Alternate Procedure

The energy minimization problem (**??**) can be solved in the following two alternate optimization steps:

$$\begin{aligned}\boldsymbol{y}^*=\arg\min_{\boldsymbol{y}}\sum_i\psi_G(y_i,I)+\frac{1}{2}\psi_C(\boldsymbol{y},\boldsymbol{h}^*)\\ +\sum_{1\leq i,j\leq L}\psi_R(y_i,y_j),\end{aligned}\qquad(8)$$

$$\begin{aligned}\boldsymbol{h}^*=\arg\min_{\boldsymbol{h}}\sum_p\psi_{at}(h_p,x_p)+\frac{1}{2}\psi_C(\boldsymbol{y}^*,\boldsymbol{h})\\ +\sum_{(p,q)\in\mathcal{E}}\psi_S(h_p,h_q).\end{aligned}\qquad(9)$$

As a standard binary CRF problem, the first subproblem in Eq. (**??**) has an explicit solution which utilizes min-cut/max-flow algorithms (*e.g*. the Dinic algorithm [**?**]) to obtain the global optimal label configuration. And the second subproblem in Eq. (**??**) reduces to an energy minimization for a multiclass CRF. Although finding the global optimum for this energy function has been proved to be a NP-hard problem, there are various approximate methods for fast inference, such as approximate *maximum a posteriori* (MAP) methods (*e.g*. graph-cuts [**?**]). In this paper, we adopt *move making* approach [**?**] that finds the optimal

α-expansion [**?**, **?**] by converting the problems into binary labeling problems which can be solved efficiently using graph cuts techniques. The energy obtain by α-expansion has been proved to be within a known factor of the global optimum [**?**]. Considering the two alternate optimization steps together, we summarize our XXXX in Algorithm **??**.

Algorithm 1 Energy minimization

```
123
```

4. Appearance and Topic Model Generation

We use Convolutional Neural Network (CNN) to encode the superpixels' appearance. CNN has made a significant breakthrough in object detection and semantic segmentation tasks [**?**]. As demonstrated in [**?**], the classification network trained on ImageNet [**?**] can generalize well to the detection task. We train a classification model on ILSVRC with the same setup to [**?**], which uses five convolutional layers and three fully-connected layers. We represent each superpixel by the *fc6* layer, which is the first fully-connected layer containing 4096 neurons. Therefore, the appearance representation of each superpixel is a feature vector with 4096 dimentions.

Moreover, we learn the latent category (known as topic model) from the superpixels.

5. Experiments

6. Conclusion