Deep Learning Markov Random Fieldfor Semantic Segmentation

Zhang, Liqiang

May 11,2018

1 Markov Random Field

Since pixels in natural images or videos generally exhibit strong correlation, jointly modeling label distribution in all locations is desirable. To capture these contextual information, Markov random field and conditional random field are commonly used as classic frameworks for semantic segmentation. [1]They model the joint distribution of labels by defining both unary term and pairwise terms. Unary term reflects the per-pixel confidence of assigning labels while pairwise terms capture the inter-pixel constraints.

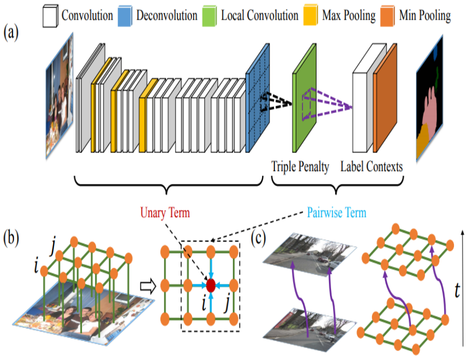

Figure 1: (a) The network architecture of a deep parsing network (DPN). (b) DPN extends a contemporary CNN architecture to model unary terms and additional layers are carefully devised to approximate the mean field algorithm (MF) for pairwise terms. (c) DPN enables dynamic linking of nodes in Markov Random Field (MRF) by incorporating domain knowledge.

2 The Approach

Markov Random Field, MRF is an undirected graph where each node represents a voxel in a video, **I**, and each edge represents relation between voxels, as shown in Fig. 1(b). Each node is associated with a binary latent variable, $y_i^u \in \{0,1\}$. indicating whether a voxel $\mathbf{i}$=[it_i]has label u. Here, i indicates a voxel' s spatial index with respect to an image, and t_i is its temporal index with respect to a sequence. We have $\forall u \in L = \{1, 2..., l\}$, representing a set of l labels. [3] The energy function of MRF is written as

$$E(y) = \sum_{\forall i \in V} \Phi(y_i^u) + \sum_{\forall i,j\varepsilon \in V} \Psi(y_i^u, y_j^u), \tag{1}$$

where y, V, and ε denote a set of latent variables, nodes, and edges, respectively, respectively.

Dynamic Node Linking,Traditional approaches usually define the edges ε on rectangular grid in 3-D space. However, when large motion exists, the actual temporal trajectory for certain pixel will not reside inside a rigid cube, which means the rectangular grid assumption does not hold.Specifically, we keep the 2-D structure in the spatial domain, as illustrated in Fig. 1(c).

$$(i, j) \in \varepsilon_t \Longleftrightarrow j = i + \triangle_{i \to j}, \tag{2}$$

Unary and Pairwise Terms, Intuitively, the unary terms represent per-voxel classifications, while the pairwise terms represent a set of smoothness constraints. The unary term in Eqn. 1 is typically defined as

$$\Phi(y_i^u) = -lnp(y_i^u = 1 \mid \mathbf{I}), \tag{3}$$

where $p(y_i^u = 1 \mid \mathbf{I})$indicates the probability of the presence of label u at voxel i, modeling by $VGG_1 6$. To simplify discussions, we abbreviate it as p_i^u.The smoothness term can be formulated as

$$\Psi(y_i^u, y_j^u) = \mu(u, v)d(\mathbf{i,j}), \tag{4}$$

where the first term learns the penalty of global co-occurrence between any pair of labels. For example, the output value of $\mu(u, v)$ is large if u and v should not coexist.

3 Experiments

name	training	validation	testing	video data
VOC12	10582	1449	1456	no
Cityscapes	2975	500	1525	no
CamVid	367	-	233	yes

Table 1: Evaluation of the self-paced regularizers on DAVIS.

Dataset. [2] The author choose those two benchmarks to evaluate the original DPN. On the other hand, CamVid dataset is composed of several video sequences, which is suitable for the evaluation of spatial-temporal DPN. They summarize the information of all datasets they used in Table. 1.

References

[1] Isola*et al.*[3]. Scene collaging: Analysis and synthesis of natural images with semantic layers. In *IEEE International Conference on Computer Vision*, pages 3048–3055, 2013.

[2] Jordan*et al.*[10]. Semantic image segmentation with deep convolutional nets and fully connected crfs. *Computer Science*, (4):357–361, 2014.

[3] Jordan*et al.*[7]. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.