Integrating Parametric and Non-parametric Models for Scene Labeling

Bing Shuai, Zhen Zuo, Gang Wang, Bing Wang, Lifan Zhao

School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

We adopt Convolutional Neural Networks (CNN) as our parametric model to learn discriminative features and classifiers for local patch classification. They are able to produce satisfactory labeling results for visually dissimilar pixels. However, CNNs struggle in visually similar pixels due to using their limited context. As shown in Figure 1, the sand pixels are highly confused with road and sidewalk pixels in a local view.

We propose to utilize global scene semantics to eliminate ambiguity of local context, for example, the confusion between 'road' and 'sand' pixels in Figure 1 can be easily removed if the "coast" scene is revealed. The global scene constraint is achieved by adding a global potential to the energy function. The energy function is formally written as:

$$E(X,Y) = \sum_{i \in X} \Phi_I(X_i, Y_j) + \Phi_G(X,Y)$$
(1)

where $\Phi_I(X_i, Y_j) = -P_I(X_i, Y_j)$ is the unary potential function defined as negative likelihood of pixel X_i being labeled as Y_j by parametric CNN model; $\Phi_G(X, Y)$ is the global potential of image X taking labeling configuration Y. Since it's infeasible to model the huge labeling state of Y parametrically $(|L|^N)$, a non-parametric approach like [2] is adopted to model the global potential, which is defined as:

$$\Phi_G(X,Y) = -\sum_{i \in X} P_G^{\mathcal{S}(X)}(X_i,Y_j) \tag{2}$$

where S(X) is the similar exemplars of image X and $P_G^{S(x)}(X_i, Y_j)$ is global class likelihood of X_i labeled as Y_j . By rewriting the energy function, it gives us the following form:

$$E(X,Y) = -\sum_{i \in X} (P_I(X_i, Y_j) + P_G^{S(X)}(X_i, Y_j))$$
(3)

Therefore, the energy function can be interpreted as an integration of beliefs from two sources: (1), Local belief: $P_I(X_i, Y_j)$ measures the belief for local context centering on pixel X_i ; (2), Global belief: $P_G^{S(X)}(X_i, Y_j)$ denotes the belief for X_i from global scene view. The global belief is calculated in a weighted K-NN manner:

$$P_{G}^{\mathcal{S}(X)}(X_{i},Y_{j}) = \frac{\sum_{k} \phi(X_{i},X_{k}) \delta(Y(X_{k}) = Y_{j})}{\sum_{k} \phi(X_{i},X_{k})}$$

$$\phi(X_{i},X_{j}) = exp(-\alpha||x_{i} - x_{j}||)exp(-\gamma||z_{i} - z_{j}||)$$
(4)

where X_k is the *k*-th nearest neighbor of X_i among all the pixel features in S(X), $Y(X_k)$ is the ground truth label for pixel X_k ; $\delta(Y(X_k), Y_j)$ is an indicator function; $\phi(X_i, X_k)$ measures the similarity between X_i and X_k , which is defined over spatial and feature space; $x_i = F(X_i)$ denotes the CNN pixel feature for X_i , z_i is the normalized coordinate along the image height axis and α, γ controls the belief exponential falloff.

Furthermore, we replace the softmax layer of previous CNN (CNNsoftmax) with a fully connected layer parameterized by W and fix the biases to be zero, which serves as a Mahalanobis metric ($M = W^T W$). We call the new network CNN-metric. In details, the Mahalanobis metric $M = W^T W$ is learned by minimizing the loss function, which is formally written as:

$$L = \frac{\lambda}{2} ||W||^{2} + \frac{1}{2N} \sum_{i,j} g(x_{i}, x_{j})$$

$$g(x_{i}, x_{j}) = max(0, 1 - \ell_{i,j}(\tau - ||Wx_{i} - Wx_{j}||^{2}))$$
(5)

where $\ell_{i,j}$ indicates whether two features have the same semantic label or not, and $\ell_{i,j} = 1$ if X_i and X_j are from the same class, or $\ell_{i,j} = -1$ otherwise; $\tau(>1)$ is the margin and λ controls the effect of regularization; $x_i = F(X_i)$ is the feature representation for X_i and N is the number of features. The objective function would enforce the pixel features from the same semantic class to be close and stay within the ball with radius $1 - \tau$, and enforce data from different classes to be far away from each other by at least $1 + \tau$.





Figure 1: Motivation of our method: the parametric model can distinguish visually different pixels very well, but get confused for pixels that are visually similar in local context. However, the local features can be disambiguated from global scene semantics. A more consistent labeling result can be achieved by integrating their beliefs. The figure is best viewed in color.

	Stanford	Sift Flow
Multiscale convnet[1] [†]	78.8% (72.4 %)	-
Multiscale convnet[1] [‡]	-	67.9% (45.9 %)
Plain CNN (133×133)[3]	79.4% (69.5%)	76.5% (30.0%)
Recurrent CNN (67×67) [3]	76.2%(67.2%)	65.5%(20.8%)
RCNN (133×133)[3]	80.2% (69.9%)	77.7% (29.8%)
$[1]^{\dagger}$ + CRF	81.4% (76.0%)	78.5% (29.4%)
[1] [‡] + CRF	-	72.3% (50.8 %)
Ours CNN (65×65)	79.1% (70.1%)	74.6% (38.2%)
Ours Final(65×65)	80.3% (70.9%)	79.8 % (39.1%)
Ours Final(65×65, metric)	81.2 % (71.3%)	80.1 % (39.7%)

Table 1: Performance comparison with state-of-the-art methods. The numbers following the networks indicate the size of input context. The percentages given outside and inside of parenthesis denote overall pixel accuracy and average class accuracy respectively.

The quantitative results are presented in Table 1. In comparison with other CNNs that are fed with richer context input, our integration model is able to yield significantly better results that are comparable to state-of-theart. As evidenced by Table 1, our integration model is capable of significantly boosting the qualitative results (global pixel accuracy) of CNN local labeling by introducing global scene constraint: 2.1% and 5.0% global pixel accuracy improvement for Stanford Background and Sift Flow benchmark respectively. In the meantime, our method can also improve the average class accuracy.

- Clément Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1915–1929, 2013.
- [2] Pablo Márquez-Neila, Pushmeet Kohli, Carsten Rother, and Luis Baumela. Non-parametric higher-order random fields for image segmentation. In *ECCV 2014*, pages 269–284. Springer, 2014.
- [3] Pedro Pinheiro and Ronan Collobert. Recurrent convolutional neural networks for scene labeling. In *Proceedings of The 31st International Conference on Machine Learning*, pages 82–90, 2014.