Multi-Objective Convolutional Learning for Face Labeling

Hongzhi Liu

May 17, 2018

1 A Novel Learning Method

It is known that deep convolutional neural networks (CNNs) have been applied to image labeling and parsing problem. As powerful end-to-end nonlinear classifiers, CNNs generate more discriminative representations compared with traditional methods based on hand-crafted features. But it will significantly increase computational cost and may not guarantee convergence when one wants to learn CNNs with structure loss. Professor Liu and his team present a multi-objective convolutional learning method which is developed for image labeling problems [1] by decomposing the structured loss of CRFs [2] into two distinct, nonstructured losses, and optimizing a single unified CNN model with weight sharing. Besides, they introduce a nonparametric facial prior to CNN training that significantly reduces the network size.

In the paper of Professor Liu, he proposes to learn a single unified CNN for both unary and pairwise classifiers. By sharing all the features within a single CNN, the two classifiers are able to enjoy better generalization ability and higher computational efficiency. They denote the parameters of the shared CNN network by w, and the feature response extracted from the topmost intermediate layer of CNN by $h_i = h(x_i, w)$. Thus, the output of the unary classifier is given by a softmax function as Equation (1) and the output of pairwise classifier is given by a logistic function as Equation (2):

$$P_u(y_i = \ell | h_i, w_u) = \frac{exp((w_u^\ell)^T) h_i}{\sum_{\ell=1}^{K} exp((w_u^\ell)^T) h_i} \quad (1)$$

$$P_b(z_{ij} = 1 | h_i, w_b) = \frac{1}{1 + exp(-w_b^T h_i)} \quad (2)$$

where w_u^ℓ represents the parameters for the ℓ-th class. Accordingly, the softmax loss for unary term as Equation (3):

$$L_u(y_i, x_i, w, w_u) = \log P_u(y_i = \ell | h_i, w_u) \quad (3)$$

Based on these two loss functions as Equation (1)and as Equation (2), Professor Liu trains the unified CNN through multi-objective optimization as Equation (4):

$$\min_w \{O_u(w, w_u), O_b(w, w_b)\},$$
$$\begin{cases} O_u(w, w_u) = \mathbb{E}(\sum_{i \in \nu} L_u(y_i, x_i, w, w_u)) + \Psi(w, w_u) \\ O_b(w, w_b) = \mathbb{E}(\sum_{i,j \in \epsilon} L_u(z_{ij}, x_{ij}, w, w_b)) + \Psi(w, w_b) \end{cases} \quad (4)$$

where $O_u(w, w_u)$ is the expected loss $\mathbb{E}(\cdot)$ for the unary classifier and $O_b(w, w_b)$ is the expected loss for the binary classifier over all the training samples. In addition, $\Psi(w, w_u)$ and $\Psi(w, w_b)$ are regularization terms. The network is updated through combining gradients of both the softmax and logistic loss functions for backpropagation.

This multi-objective CNN has two main advantages: First, the convolutional network generates expressive representations at lower levels that can be utilized for both unary and pairwise model regressions. Second, the unified network can be learned by backpropagating errors from both outputs jointly such that the network can learn features that are highly adaptive to both objectives.

2 CNN Architecture

Since CNNs usually operate on a patch level centered at each pixel, the labeling pipeline is based on a sliding window input [3] with overlapping patches, as shown in Figure 1.

The inputs are 72×72 single scale patches which are passed to two top consecutive convolutional units with a

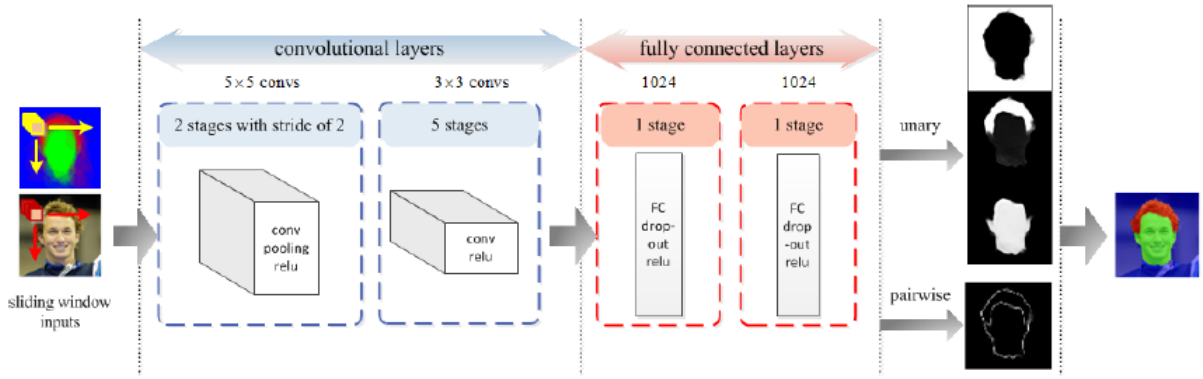

Figure 1: Proposed CNN classifier with sliding window based inputs.

filter of 5×5, where each convolutional layer is followed by one max pooling layer with a downsampling stride of 2. Following that is another stack of small convolutional units with a receptive field of 3×3 and no pooling layer.

3 Evaluation of the Method

Professor Liu evaluates the proposed algorithm on two different benchmark datasets with different face labeling tasks. They show that it applies to both tasks and performs favourably against state-of-the-art methods with the same framework and experimental settings. Specifically, the team demonstrate that both the multi-objective approach and the nonparametric prior improve the performance in all aspects compared to a per-pixel CNN classifier.

In the thesis, they use the LFW dataset which has been used by recent methods for face labeling. In table 1, we test a series of approaches with the channel numbers of the two FC layers as 4096 and 1024, and evaluate the results with respect to per-pixel accuracy and F-measure of each class. The first 3 rows show the approaches without nonparametric prior input, while the lower 3 rows are those using it. Overall, the nonparametric prior significantly improves the results when compared with all corresponding approaches.

Table 1: Overall per-pixel accuracy on the *LFW-PL* dataset with channel numbers of two FC layers setting as 4096 and 1024. The F-measure of skin (F-skin), hair (F-hair) and background (F-bg) are also presented.

Method	accuracy	F-skin	F-hair	F-bg
S-CNNs	92.92%	90.07%	73.73%	95.18%
MO-unary	93.45%	91.45%	78.03%	95.84%
MO-GC	93.77%	91.95%	79.06%	96.03%
S-CNNs with prior	94.25%	92.79%	77.18%	96.63%
MO-unary with prior	94.94%	93.64%	79.95%	97.02%
MO-GC with prior	95.12%	93.93%	80.70%	97.10%

References

[1] Sifei Liu, Jimei Yang, Chang Huang, and Ming Hsuan Yang. Multi-objective convolutional learning for face labeling. In *Computer Vision and Pattern Recognition*, pages 3451–3459, 2015.

[2] Ren Ranftl and Thomas Pock. A deep variational model for image segmentation. In *German Conference on Pattern Recognition*, pages 107–118, 2000.

[3] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann Lecun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *Eprint Arxiv*, 2013.