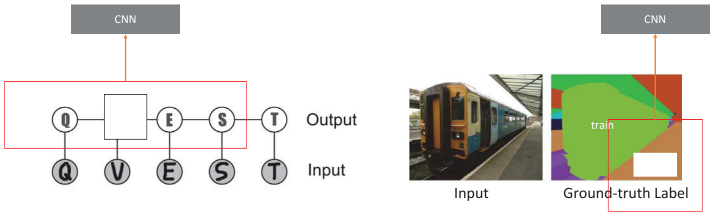
Top-down Learning for Structured Labeling with Convolutional Pseudoprior

Saining Xie* and Xun Huang* and Zhuowen Tu
{s9xie,ztu}@ucsd.edu, xh258@cornell.edu
University of California San Diego, Cornell University

Abstract



Current practice in convolutional neural networks (CNN) remains largely bottom-up and the role of top-down process in CNN for pattern analysis and visual inference is not very clear.In this paper, we propose a new method for structured labeling by developing convolutional pseudoprior (ConvPP) on the ground-truth labels. Our method has several interesting properties: (1) compared with classic machine learning algorithms like CRFs and Structural SVM, ConvPP automatically learns rich convolutional kernels to capture both short- and long- range contexts; (2) compared with cascade classifiers like Auto-Context, ConvPP avoids the iterative steps of learning a series of discriminative classifiers and automatically learns contextual configurations; (3) compared with recent efforts combining CNN models with CRFs and RNNs, ConvPP learns convolution in the labeling space with improved modeling capability and less manual specification; (4) compared with Bayesian models like MRFs, ConvPP capitalizes on the rich representation power of convolution by automatically learning priors built on convolutional filters. We accomplish our task using pseudo-likelihood approximation to the prior under a novel fixed-point network structure that facilitates an end-to-end learning process.

Background

- We approximate the appearance $p(\mathrm{X}|\mathrm{Y})$ with a discriminative model
  - Tu, et al. "Brain anatomical structure segmentation by hybrid discriminative/generative models." *Medical Imaging, IEEE Transactions on, 2008*
- And the prior $p(\mathrm{Y})$ with pseudo-prior
  - inspired from the pseudo-likelihood approximation: Besag, Julian. "On the statistical analysis of dirty pictures." *Journal of the Royal Statistical Society. Series B (Methodological)* (1986): 259-302.
  - Detailed approximation steps please refer to our paper

$$p(\mathbf{Y}|\mathbf{X}) \propto p(\mathbf{Y})p(\mathbf{X}|\mathbf{Y}) \dot{\propto} \underbrace{\prod_i p(y_i|\mathbf{Y}_{\mathcal{N}_i/i})}_{\text{Top-down convolutional pseudo-prior (ConvPP)}} \cdot \underbrace{\prod_i p(y_i|\mathbf{X})}_{\text{Bottom-up CNN}}$$

Method

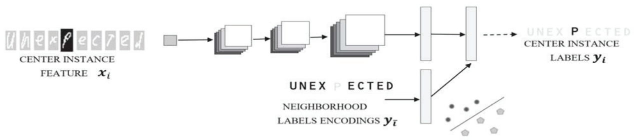
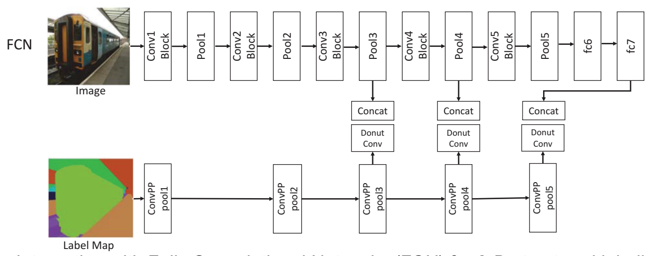
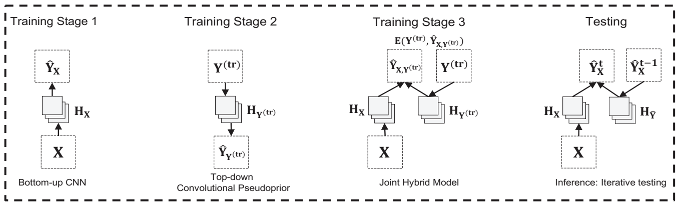


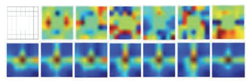
Illustration of ConvPP learning process in 1-D structured labeling tasks (OCR)



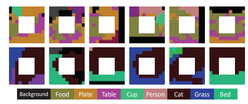
Integrating with Fully Convolutional Networks (FCN) for 2-D structured labeling tasks

Network Architecture

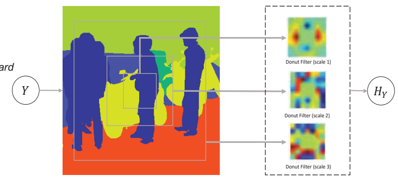




Learned donut filters vs trivial solution learned by standard convolutional kernels



Each row displays top 6 label patches (on pascal-context dataset) that produce highest activation for a donut filter



Multiple donut filter layers with the same kernel size are integrated into different depth of the joint network. Multi-scale context learning is naturally handled.

Results

- Structured Labeling: 1-D case

Methods	small	large
Linear-chain CRF (Do & Arti, 2010)	21.62%	14.20%
M^3N (Do & Arti, 2010)	21.13%	13.46%
SEARN (Daumé et al., 2009)	-	9.09%
SVM + CRF ((Hoefel & Elkan, 2008)	-	5.76%
Neural CRF (Do & Arti, 2010)	10.8%	4.44%
Hidden-unit CRF (van der Maaten et al., 2011)	18.36%	1.99%
Fixed-point (Li et al., 2013)	**2.13%**	0.89%
Online deep learning (Chen et al., 2014)	-	0.63%
ConvPP (ours)	6.49%	**0.57%**

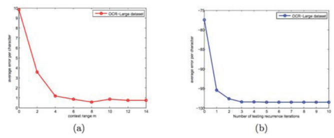
Performance on OCR dataset

Methods	accuracy
Linear SVM (Do & Arti, 2010)	9.87%
Linear CRF (van der Maaten et al., 2011)	6.54%
NeuroCRFs (Do & Arti, 2010)	6.05%
Hidden-unit CRF (van der Maaten et al., 2011)	4.43%
Online deep learning (Chen et al., 2014)	3.34%
ConvPP (ours)	**1.09%**

Performance on FAQ dataset

Training Data Percentage (%)	10	20	30	40	50	60	70	80	90
Generalization Error(%)	**6.49**	**3.28**	**2.09**	**1.67**	**1.55**	**1.03**	**0.92**	**0.72**	**0.57**

Performance on OCR dataset: varying training size



Performance on OCR dataset:
(a) Varying context window length
(b) Varying testing iterations

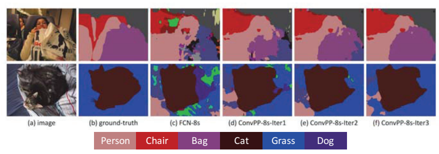
- Image Labeling: 2-D case

	mean IU
O_2P (Carreira et al., 2012)	18.1
CFM (VGG+SS) (Dai et al., 2015b)	31.5
CFM (VGG+MCG) (Dai et al., 2015b)	34.4
CRF-RNN (Zheng et al., 2015)	39.3
BoxSup (trained with additional COCO boxes) (Dai et al., 2015a)	40.5
FCN-32s (Long et al., 2015)	35.1
ConvPP-32s (ours)	37.1
FCN-16s (Long et al., 2015)	37.6
ConvPP-16s (ours)	40.3
FCN-8s (Long et al., 2015)	37.8
ConvPP-8s (ours)	**41.0**

Performance on Pascal-Context dataset

	mean IU
FCN-16s (Long et al., 2015)	39.1
ConvPP-16s (ours)	39.7
FCN-8s (Long et al., 2015)	39.5
ConvPP-8s (ours)	**40.7**

Performance on SIFT Flow dataset



Iterative update of labeling results during testing. Segmentation results are gradually refined.