Deep Convolutional Networks for Scene Parsing

David Grangier DAVID@GRANGIER.INFO
Léon Bottou LEON@BOTTOU.ORG
Ronan Collobert RONAN@COLLOBERT.COM
NEC Labs America, 4 Independence Way, Princeton, USA

Abstract

We propose a deep learning strategy for scene parsing, i.e. to asssign a class label to each pixel of an image. We investigate the use of deep convolutional network for modeling the complex scene label structures, relying on a *supervised* greedy learning strategy. Compared to standard approaches based on CRFs, our strategy does not need hand-crafted features, allows modeling more complex spatial dependencies and has a lower inference cost. Experiments over the MSRC benchmark and the LabelMe dataset show the effectiveness of our approach.

1. Introduction

Scene parsing aims at segmenting and recognizing the various objects appearing in an image. Formally, a parser maps each pixel of an image into one of several predefined classes. Applications of parsing includes robot navigation, image retrieval, or 3D scene reconstruction. The challenges of this task are two-fold: (i) like traditional computer vision systems such as detectors, a parsing model should be robust to changes in illumination, viewpoint, and inter-class variability, (ii) like a text parser, the model should also exploit the spatial dependencies between classes.

The parsing problem is mainly addressed with Conditional Random Fields (CRFs) (Schroff et al., 2008; Verbeek & Triggs, 2008). CRFs linearly combine input features (describing patches surrounding each pixel) along with contextual features (describing spatial interactions between labels). Such models are generally trained to maximize the likelihood of the correct classification given the input features. The limitations of CRFs are mainly: (i) their dependence on hand-designed features, i.e. linear combination or kernel-based models require providing strong features, (ii) their costly inference which practically deters from modeling complex label dependencies, i.e. inference requires searching over label configurations, e.g. through Gibbs sampling, and hence requires simple contextual features, (iii) their expensive training which need estimating the normalization factor.

This work proposes an alternative strategy which overcomes these limitations. It relies on deep neural networks (Bengio et al., 2007), a versatile, powerful familly of learning machines which allows to compactly model complex dependencies between features and labels. In our case, we rely on Convolutional Networks (CNNs), a type of neural network adapted to computer vision problem, thanks to convolutional layers performing 2D filtering operations (LeCun & Bengio, 1995). Deep models have two main advantages in the context of convolutional networks. First, like for any neural network, depth allows a compact representation of complex functions. Second, the succession of spatial convolutions allows modeling increasingly larger spatial dependencies.

The training of deep architecture represents a difficult optimization problem and has recently become an active research topic. Our work builds upon research suggesting greedy layer-wise learning strategies (Bengio et al., 2007). However, as opposed to this prior work, our layer-wise learning procedure relies on *supervised* learning, since our primary intend is to model the complex dependencies between the pixel labels of an image.

Compared to CRFs, our strategy has several advantages (i) it does not requires any hand-crafted features since deep networks can model arbitrary complex functions from the RGB input image, (ii) its inference does not involve searching the label space but simply requires the forward evaluation of a function (iii) discriminative training is performed efficiently through Stochastic Gradient Descent (SGD), without the need for estimating any normalization factor. Moreover, our experiments over the MSRC dataset demonstrate the generalization performance of our approach compared to CRF-based solutions.

In the following, Section 2 sketches our approach and Section 3 presents our experimental results.

2. Deep CNN for Scene Parsing

Our learning strategy starts with a simple model and iteratively add new layers to it. In the first step, cnn_1 consists of 4 layers, 3 layers performing a convolution followed by squashing function and a linear output layer. The next models $\text{cnn}_2, \text{cnn}_3$ and cnn_4 are built as follows: cnn_i contains all layers of cnn_{i-1} except the output layer, these layers are followed by a new convolution followed by squashing and a new linear output layer.

The models are learned sequentially: first, cnn_1 is trained to optimize the pixel-wise cross entropy through SGD. Then, cnn_2 is trained with the same approach. During this training, all parameters of cnn_2 are updated, including those shared with cnn_1. After cnn_2 training, cnn_3 and then cnn_4 are learned in the same manner. This sequential training has been selected after evaluating various learning strategies (not explained here for space consideration). In our experiments over LabelMe we have observed that adding depth always yields higher performance when using the presented learning algorithm. In comparison, when training cnn_2, cnn_3, cnn_4 from a random initialization, i.e. without the pretraining of the previous models, all models yield the same performance as cnn_1.

Conceptually, the penultimate layer of each model provides a hypothesis map from which a linear layer can predict the output class. The successive convolutions allow refining the hypothesis maps, thanks to added non linearities and a larger spatial context. In that sense, the deeper models produce higher level representations, which are increasingly complex functions of the inputs and correspond to less and less local descisions.

3. Experiments

Our experiments are performed over both MSRC[1] and LabelMe data[2]. The MSRC dataset contains 240 small images (320×213) and 9 classes, most pixels ($\sim 70\%$) have been labeled, each image contains only few classes (typically ≤ 3). Table 1 compares our approach to a CRF relying solely on hand-crafted features (Verbeek & Triggs, 2008), and to a CRF relying on a pretrained local classifier (a Random Forest) (Schroff et al., 2008). One can notice that our approach outperforms both CRFs, even on this small training setup that favors manually encoding prior vision knowledge.

Table 1. MSRC Pixel Error (%)

	CRF-features	CRF-classifier	cnn_4
9 classes	15.1	12.8	11.5

Table 2. LabelMe Pixel Error (%)

	cnn_1	cnn_2	cnn_3	cnn_4
5 classes	12.4	9.9	8.2	7.9
20 classes	30.0	25.1	21.8	N/A

The LabelMe version we use contains $2,700$ city scenes and represents a more challenging problem. In our setup, we used the 500×375 version of the images and we worked with the 5 and 20 most common classes, which corresponds to respectively 37% and 44% of the pixels. This task is more challenging than MSRC, with more classes per images (typically > 5), an incomplete, imprecise labeling and a very unbalanced class distribution. To our knowledge, no parsing results have yet been reported on this challenging set (so far only detection has been evaluated on this data). Our experiments hence intend to provide a baseline approach, and evaluate our model in a rich, challenging parsing environement[3]. Table 2 reports the improvement obtained by adding depth to the network.

4. Conclusions

This work proposes a supervised deep learning strategy for scene parsing that allows exploiting complex, non-local class dependencies. Compared to state-of-the-art CRF approaches, our proposal is shown to be advantageous, both in terms of generalization performance and inference speed. In the future, we would like to investigate whether unsupervised pre-training of the model might further help the learning process. We also would like to experiments data-driven approaches to incorporate prior knowledge. We further can envision extensions of this work to other structured output problems.

References

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. *Neural Information Processing Systems (NIPS).*

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time-series. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks.* MIT Press.

Schroff, F., Criminisi, A., & Zisserman, A. (2008). Object class segmentation using random forests. *British Machine Vision Conference (BMVC).*

Verbeek, J., & Triggs, B. (2008). Scene segmentation with crfs learned from partially labeled images. *Neural Information Processing Systems (NIPS).*

[1] http://research.microsoft.com/jump/51355

[2] http://people.csail.mit.edu/torralba/benchmarks/

[3] Parsing examples are available at `http://david.grangier.info/scene_parsing/`