CRF learning with CNN features for hyperspectral image segmentation

Author

Alam, Fahim Irfan, Zhou, Jun, Liew, Alan Wee-Chung, Jia, Xiuping

Published

2016

Conference Title

2016 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM (IGARSS)

Version

Accepted Manuscript (AM)

DOI

10.1109/IGARSS.2016.7730798

CRF LEARNING WITH CNN FEATURES FOR HYPERSPECTRAL IMAGE SEGMENTATION

Fahim Irfan Alam[1], *Jun Zhou*[1], *Alan Wee-Chung Liew*[1], *Xiuping Jia*[2]

[1] School of Information and Communication Technology, Griffith University
Nathan, QLD 4111, Australia
[2] School of Engineering and Information Technology, University of New South Wales
Canberra, ACT, Australia

ABSTRACT

This paper proposes a method that uses both spectral and spatial information to segment remote sensing hyperspectral images. After a hyperspectral image is over-segmented into superpixels, a deep Convolutional Neural Network (CNN) is used to perform superpixel-level labelling. To further delineate objects from a hyperspectral scene, this paper attempts to combine the properties of CNN and Conditional Random Field (CRF). A mean-field approximation algorithm for CRF inference is used and formulated with Gaussian pairwise potentials as Recurrent Neural Network. This combined network is then plugged into the CNN which leads to a deep network that has robust characteristics of both CNN and CRF. Preliminary results suggest the usefulness of this framework to a promising extent.

Index Terms— Image Segmentation, Superpixel, Deep Learning, Conditional Random Field, Convolutional Neural Network.

1. INTRODUCTION

Hyperspectral imaging is concerned with measurement and analysis of images acquired in contiguous spectral bands over a given spectral range [1]. It has become a valuable tool for a wide range of remote sensing applications such as agriculture, mineralogy, surveillance and environmental sciences [2]. A common event in these applications is the segmentation of image. Because hyperspectral images provide ample spectral information to identify and distinguish between spectrally similar materials, it brings more accurate performance in segmentation when compared to the traditional segmentation algorithms that face restrictions in separating similar looking different fragments of an object. Consequently, hyperspectral imagery provide the capability for more accurate and detailed information extraction than is possible with other types of remotely sensed data. Along with spectral information, we can also utilize the spatial relationships among various spectra in a neighbourhood, which allows further elaborate spectral-spatial models for accurate segmentation and classification of the image.

Recent advances in training multilayer neural networks have contributed much in a wide variety of machine learning problems including classification or regression tasks. Machine learning systems with multiple layers, often referred to as "deep" architecture can extract more abstract, invariant features of data and thus are believed to have the ability of producing higher classification accuracy than the traditional classifiers [3]. Classification methods based on spectral and spatial information using deep learning models were presented in [4][5]. Furthermore, Convolutional Neural Network (CNN) has been used in many image processing tasks for pixel-level labelling problems. With this model, we can learn a good representation of features which allows to perform an end-to-end labelling task.

However, there are a number of issues associated with CNN-based segmentation. For example, convolutional filters with large receptive fields produce coarse outputs. Presence of max-pooling layers reduces the chance of obtaining a fine segmentation result, and the absence of smoothing constraints can limit the performance of deep models. In this regard, probabilistic graphical models such as Markov Random Field (MRF) and Conditional Random Field (CRF) have already been used to refine weak and coarse segmentation outputs as a post-processing step. This is not ideal because this step is entirely disconnected from the training of the deep network and does not fully utilize the usefulness of CRFs. In this paper, we propose to integrate CNN with CRF to improve the segmentation performance. Our work is inspired by [6] where CRF was used as a part of the CNN in order to perform segmentation of color images. Instead of using pixel-level labelling, we perform superpixel-level labelling on hyperspectral images by deep learning and then learn the CRF parameters accordingly.

In our method, we first generate superpixels based on both spectral and spatial features of each pixel along the bands of a hyperspectral image. Then we use CNN to obtain labels for those superpixels. Given the labels, we introduce CRF into our framework whose pairwise potentials are modeled as weighted Gaussians which require an intensive approximation for inference. This process is reformulated as multiple layers of a CNN. This combination of CRF and CNN are later

added as a part of the resulting CNN to perform segmentation.

2. PROPOSED METHOD

In this paper, we propose a method that combines both CRF and CNN into a common framework by considering both spectral and spatial properties of the pixels in the hyperspectral image. Our method starts with formation of the superpixels and then gradually proceeds toward the combination of CRF and CNN. The individual steps of the method are briefly presented in the next discussions.

2.1. Forming Superpixels

In this work, we used Simple Linear Iterative Clustering (SLIC) algorithm [7] [8], a modified version of K-means algorithm, to produce superpixels by using both the spectral and spatial features of every individual pixel. The parameters of the algorithm control the size and the regularity of the superpixels with fast computation speed and good accuracy. It is also expected to generalize well to multiple spectral bands. The steps of the algorithm are as follows:

- Construct a feature vector for every pixel in the hyperspectral image based on both spectral and spatial information.
- Construct an initial set of cluster centers on a grid of size S. Each cluster center is moved to the lowest gradient position in an $n \times n$ neighborhood.
- Assign each pixel to the closest cluster center by measuring the Euclidean distances between the pixel and the set of initial cluster centers. This step is accelerated by simplifying the search within a $2S \times 2S$ neighborhood.
- Update the cluster centers based on the results given by the distance measure.
- Repeat the process iteratively until the distance between the successive cluster center updates is below a threshold.
- Finally, a postprocessing step enforces connectivity by reassigning disjoint segments to nearby cluster.

2.2. CNN for Superpixel-level Labelling

Convolutional neural networks (CNNs) are a class of deep learning models that integrate feature extraction, feature combination and classification with a single neural network that is trained end to end from raw pixel values to classifier outputs. A CNN consists of multiple layers of small neuron collections which look at small portions of the input image. The results of these collections are then tiled in order to obtain a better representation of the original image. This process is repeated for every such layer.

CNNs may include local or global pooling layers, which combine the outputs of the previous layer. They also consist of various combinations of convolutional and fully connected layers, with nonlinear activation functions like *tanh* applied at the end of each layer. A convolution operation on small regions of the input image is performed. Each layer applies some filters. One major advantage of a CNN is that it allows the use of shared weight in convolutional layers, which means that the same filter is used for each pixel in the layer. During the training phase, a CNN automatically learns the values of its filters. Each filter gradually proceeds toward forming lower-level features into higher-level representation. The basic working principle of a CNN is graphically represented in Figure 1.

To learn about huge information from a hyperspectral image, we need a model with large learning capacity. A promising feature of CNN is that their capacity can be controlled by varying their depth and breadth. Thus, compared to standard feedforward neural networks with similarly-sized layers, CNNs have much fewer connections and parameters and so they are easier to train. In our work, the CNN contains eight layers with weights; the first five are convolutional and the remaining three are fully-connected. The output of the last fully-connected layer is fed to an MLP which produces a distribution over the class labels of the superpixels.

2.3. A Mean-Field Iteration of CRF inference as a Stack of CNN Layers

In this section, we briefly explain how we used CRF for superpixel-wise labelling. The CRF models superpixel labels as random variables that form an MRF when conditioned upon a global observation which is considered to be the image [9]. In a fully connected pairwise CRF model, the energy of label assignment of a superpixel is given by the following equation:

$$E(x) = \sum_{i} \phi(x_i) + \sum_{i \neq j} \psi(x_i, x_j) \qquad (1)$$

where the first term represents the unary potential of the inverse likelihood of the superpixel i taking a particular label x_i and the second term is the pairwise potential between two superpixels i and j for assigning two labels x_i, x_j simultaneously.

In our model, the unary energies are obtained from the CNN which predicts the labels for the superpixels but it does not consider the smoothness and consistence of the label assignments. The pairwise energies contain a smoothness term that encourages assigning similar labels to superpixels with similar properties. As in [10], we model the pairwise energies as weighted Gaussians as follows:

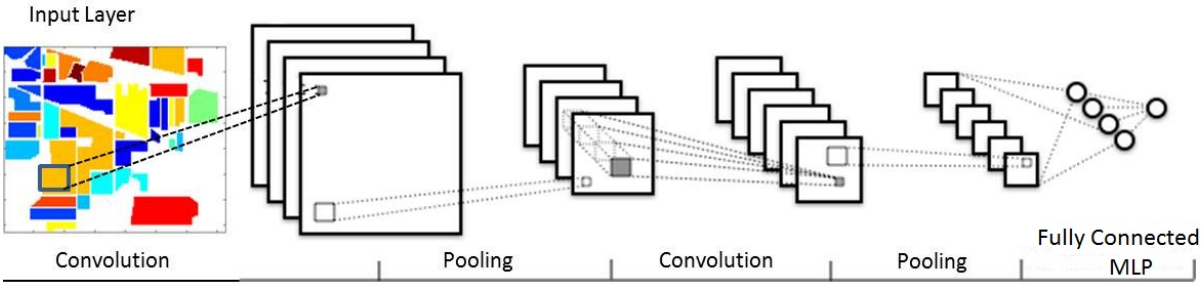

Fig. 1. A Convolutional Neural Network

$$\psi(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^{M} w_{(m)} k_{(m)}(v_i, v_j) \qquad (2)$$

where $\mu()$ is the label compatibility function which encodes a Potts model, i.e., $\mu(x_i, x_j) = 1_{x_i \neq x_j}$. The Potts model effectively penalises the case where two superpixels *i* and *j* are assigned different labels when $\sum_{m=1}^{M} w_{(m)} k_{(m)}$ is large. The efficient approximate inference requires the kernels $k_m(.,.)$ to be Gaussian kernels computed over elements of the feature vector v_i that describes pixel i with a scalar weight w_m. By minimizing the CRF energy, we obtain the most probable label assignment for the superpixels which is intractable. Therefore, we use the mean-field approximation algorithm to the CRF distribution for maximum posterior marginal inference. It does this by approximating the CRF distribution $P(X)$ by a simpler distribution $Q(X)$ which is expressed as the product of independent marginal distributions $Q(X) = \prod_i Q_i(X_i)$.

The inference algorithm works in an iterative manner. The first step is the initialization in which the following operation is performed:

$$Q_i(l) \leftarrow \frac{1}{Z_i} exp(U_i(l)) \qquad (3)$$

where $Z_i = \sum_l exp(U_i(l))$. Note that $U_i(l)$ denotes the negative of the unary energy. This operation is simply applying a soft-max function over the unary potential across all the labels at each superpixel.

The second step is the message passing which is applying Gaussian filters on the current estimation of the predictions of the superpixels. This reflect how strongly two superpixels are related to each other. By using back propagation, we calculate error derivatives on the filter responses. The next step is to take the weighted sum of the filter outputs for each label of the superpixels. When each label is considered, it can be reformulated as the usual convolution with filter with input channels and the output channel. The error can be calculated since both inputs and outputs are known during back-propagation. This allows an automatic learning of filter weights. Next, the compatibility transform step is performed followed by adding the original unary potential for each individual superpixel. Finally, the normalization step of the iteration can be expressed as another softmax operation.

The individual iteration of the mean-field algorithm can be formulated as a stack of CNN layers. Given an image I and the superpixel-wise unary potential values U and an estimation of marginal probabilities Q_{in} from the previous iteration, the next estimation of marginal distributions after one mean-field iteration is given by $f_\theta(U, Q_{in}, I)$. The vector θ represents the CRF parameters such as the weights, compatibility function and Gaussian kernel.

The multiple mean-field iterations can be implemented by repeating the above stack of layers in such a way that each iteration takes Q value estimates from the previous iteration and the unary potential in their original form. This is equivalent to treating the iterative mean-field inference as a Recurrent Neural Network (RNN). This complete system therefore combines the usefulness of both CNN and CRF and is trainable end-to-end by back propagation algorithm. In the forward pass of the network, when the execution enters the combined part of CNN and RNN, it falls into the loop of the RNN and perform number of iterations. In this stage, neither the CNN or the CRF need to perform any calculations because the update process is performed inside the loop of the RNN. After an output is obtained from the loop, the following stages of the deep network can continue the forward pass. In this paper, a softmax layer is used after the RNN to terminate the network. During backward pass of the network, the error differentials of the output go through same number of iterations inside the loop of the RNN before reaching RNN's input. After that it gradually backpropagates toward CNN. It is to be noted that the error differentials are calculated inside each iteration of the loop of the mean-field algorithm.

3. EXPERIMENTAL RESULTS

In our experiments, we have used a hyperspectral data set Indian Pines in order to evaluate the usefulness of our proposed method. The Indian Pines dataset consists of 145×145 pixels and 224 spectral reflectance bands in the wavelength ranging from $0.4 - 2.5 \times 10^{-6}$ meters. This scene contains

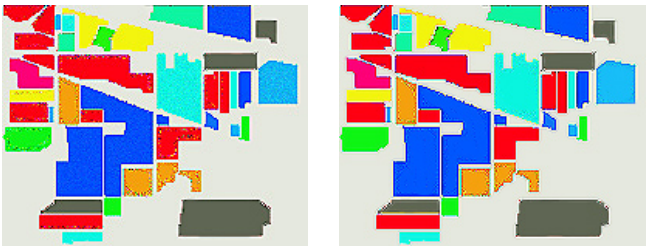

Fig. 2. Results after (a) classification and (b) segmentation.

two-thirds agriculture, and one-third forest or other natural perennial vegetation. There are two major dual lane highways, a rail line, as well as some low density housing, other built structures, and smaller roads. The ground truth of the original scene contains sixteen classes and are not all mutually exclusive.

There are several critical issues in our experiments that we were required to take into account. We extracted individual bands and carefully analysed each of them in order to be certain about their usability in our experiments. This enabled us in identifying some noisy bands which we later removed and selecting useful bands for the tests. Then we normalized the data from the selected bands and mapped them from 0 to 1. These critical preprocessing steps were performed in order to receive useful features from our hyperspectral data.

For measuring the performance, we calculated the average accuracy of the proposed method. In total, we achieved 93.4% as the classification accuracy and 95.6% as the segmentation accuracy. The classification accuracy denotes the status of our result that we obtained from the first step of our method, i.e., from the CNN which provided us the labels for the superpixels. Later after integrating the results from the CNN into the combined framework of CNN with RNN, we achieved our final segmentation result. The increase in the accuracy shows the potential of our proposed method in a sense that the integration of both functionalities of CNN and CRF can contribute greatly in the segmentation performance of hyperspectral images. We illustrate the the results obtained from the two individual steps: classification result after CNN performs unary potential which is the superpixel-wise labelling and final segmentation result in Figure 2.

4. CONCLUSION

In this paper, we attempted to perform hyperspectral image segmentation by combining the desirable properties of both CRF and CNN into a common framework. The initial preliminary results on a hyperspectral dataset is promising. We will try to apply our proposed framework on a more complex hyperspctral dataset. In this regard, we also plan to include gabor features in different wavelengths to form a more robust feature vector for the pixels in the hyperspectral dataset. In this way, we will be able to experiment on the hand crafted features in a deep learning framework.

5. REFERENCES

[1] D. Landgrebe, "Hyperspectral image data analysis," *Signal Processing Magazine, IEEE*, vol. 19, no. 1, pp. 17–28, Jan 2002.

[2] J. A. Richards, *Remote Sensing Digital Image Analysis: An Introduction*, Springer Berlin Heidelberg, NJ, USA, 5th edition, 2013.

[3] N. Kruger, P. Janssen, S. Kalkan, M. Lappe, A. Leonardis, J. Piater, A. J. Rodriguez-Sanchez, and L. Wiskott, "Deep hierarchies in the primate visual cortex: What can we learn for computer vision?," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1847–1871, Aug 2013.

[4] Y. Chen, X. Zhao, and X. Jia, "Spectral-spatial classification of hyperspectral data based on deep belief network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 6, pp. 2381–2392, 2015.

[5] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral-spatial classification of hyperspectral data using loopy belief propagation and active learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 2, pp. 844–856, 2013.

[6] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr, "Conditional random fields as recurrent neural networks," in *International Conference on Computer Vision (ICCV)*, 2015.

[7] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[8] X. Zhang, Selene E. Chew, Z. Xu, and N. D. Cahill, "SLIC superpixels for efficient graph-based dimensionality reduction of hyperspectral imagery," *Proc. SPIE*, vol. 9472, no. S2, 2015.

[9] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, USA, 2001, pp. 282–289.

[10] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in Neural Information Processing Systems*, pp. 109–117. 2011.