Hierarchical Feature For Scene Parsing Using Fully Recurrent Network

N. Shanmugapriya¹
Research Scholar, Department of CSE,
P. A. College of Engineering and Technology,
Coimbatore, Tamil Nadu.

M. Gayathri²
Department of CSE
P. A. College of Engineering and Technology,
Coimbatore, Tamil Nadu.

T. Manigandan³
Principal
P. A. College of Engineering and Technology,
Coimbatore, Tamil Nadu.

D. Chitra⁴
Professor, Department of CSE
P. A. College of Engineering and Technology,
Coimbatore, Tamil Nadu.

Abstract—In scene parsing, the wide-range contextual information is not effectively encoded. Scene parsing provides segmentation and determines an scene into different regions associated with semantic categories. The main objective of scene parsing is to reduce semantic gap between humans and computer machines on scene understanding. The scenes parsing applications are object detection, text detection on video frames etc. Scene parsing is one of the key problems in computer vision and solved using Recurrent neural network. Labeling small objects is the main difficulties in the scene parsing. The proposed method is Fully Recurrent Network using Bi-Directional Recurrent neural network enables the network to model long-range semantic dependencies among image unit. Local representation and Local classification can be enhanced by Directed Acyclic Graph. Training RNN using gradient based method. Back propagation through time algorithm can be used to calculate the gradient method. Recognition accuracy can be improved with the class weighting function. The result can be predicted accuracy using Stifflow data set.

Index Terms: Image classification, image segmentation, scene parsing, object detection, deep learning, computer vision.

1 INTRODUCTION

The main beliefs of image processing are to improve pictorial information and image data processing. Understanding image is the primary importance for wide range of applications. Scene Parsing is one of the methods to reduce semantic gap between humans and computer machines on scene understanding. After a perfect scene labeling, every region and every object is tagged and delineated.

The objective of the scene parsing task is to parse every pixel in a scene with the category of the object it belongs to. To ensure a high visual coherence and high class accuracy, it is essential for a model to capture long range pixel dependencies in scenes. A method that uses a multilevel neural network used to extract dense features from raw pixels. That encodes regions of multiple sizes centered on each pixel. Associating one of the semantic classes to each pixel in a scene image. It is usually defined as a multi-class classification problem based on their surrounding image patches. Thus, how to equip local features with a broader view of contextual awareness is a pivotal issue in image labeling. RNNs can use their internal memory to process the inputs.

Consequently, a higher-level representation of scenes (their global context) is typically constructed based on the similarity of the low-level features of pixels and on their spatial dependencies using a graphical model. The graphical models construct the global dependencies based on the similarities of neighboring segments. The most popular graph-based approaches are Markov Random Fields (MRF) [2, 3] and Conditional Random Fields (CRF) [4, 5]. However, such methods require segmentation, superpixels. In order to improve modeling of long range dependencies, [1] first revealed the use of large input patches to consider larger contexts. However, that would reduce the resolution of the final label scenes, and a huge redundancy of overlapping regions would make the understanding incompetent. Hence, they introduced Recurrent Convolutional Networks (RCNNs) for scene labeling. RCNNs train different input patch sizes (the instances) recurrently to learn increasingly large contexts for each pixel, while ensuring that the larger contexts are coherent with the smaller ones.

In this proposed method Fully Recurrent Network (FRN) architecture take into account the local (pixel-by-pixel) and global (label-by-label) dependencies in a single process for scene labeling. It can uses any additional processing or conditions like multi-scale or different patch sizes to solve the object labeling task with machine’s effort. Figure 1 shows Fully recurrent network are used for sequence learning and layers. These networks include recurrently connected cells to learn the dependencies between two time frames, then transfer
the probabilistic inference to the next frame. The FRN memory block stores and retrieves this information over short or long periods of time.

![Hierarchical RNN layer](image)

**Figure 1** Hierarchical RNN layer

1.1 Problem definition

The pixel-wise accuracy is the main problem. The inaccurate measure can be obtained. Detecting small objects is often more important than accurately labeling every boundary pixel of the image. It can be labeling the large window and it fail to detect the small objects[6]. To solve this problem fully recurrent network can be used to labeling pixel wise.

The paper is organized as follows. Chapter 2 briefly presents related works. Chapter 5 describes the proposed strategy. Chapter 6 presents the results of our experiments in standard dataset: the SIFT Flow Dataset (33 classes) and compare the performance with other systems. Finally, Chapter 7 provides a conclusion.

2. RELATED WORK

Scene parsing is one of the most testing problems in computer vision. It has attracted more and more attention in recent years. The proposed model like to highlight and discuss three lines of works that are most relevant to ours. The first line of work is to explore the contextual modeling and local representation. One attempt is to encode context into local representation. For example, Farabet et al. [6] stacks surrounding contextual windows from different scales; Pinheiro et al. [1] increases size of input windows.

More recently, deep learning has become a very active area of research in scene understanding and vision in general. In [7], color and texture features from over segmented regions are merged by Recursive Neural Networks. This work has been extended by Socher et al. [8] who combined it with convolutional neural networks. Among deep learning approaches, Convolutional Neural Networks (CNNs) [9] are one of the most successful methods for end-to-end supervised learning.

This method has been widely used in image classification [10], object recognition [12], face verification [11], and scene labeling. Farabet et al. [6] introduced multi-scale CNNs to learn scale-invariant features, but had problems with global contextual coherence and spatial consistency. These problems were addressed by combining CNNs with several postprocessing algorithms, i.e., superpixels, CRF, and segmentation trees. Later, Kekenc et al. [13] improved CNNs by combining two CNN models which learn context information and visual features in separate networks. Both mentioned approaches improved accuracy through carefully designed pre-processing steps to help the learning, i.e., class frequency balancing by selecting the same amount of random patches per class, and specific color space for the input data.

[14] Discussed a method to collective segments in a moderate fashion using a trained scoring function. The originality of the approach is that the feature vector of the combination of two segments is computed from the feature vectors of the individual segments through a trainable function. Like us, they use deep learning convolution neural network methods to train their feature extractors. But unlike us, their feature detection works on hand-engineering features. One of the main problem in scene labelling is how to label the small objects in the wide scene. [15] proposed to use the graph of labels extracted from a coarse scale as input to the labeller that looks at finer scales.

Like us, a number of authors have used group of segmentations and graph cut methods to generate candidate segments by aggregating elementary segments. The scene parsing can be done using the inference algorithm[16][17] based on the graph cut method and also segmentation trees. Other strategy using group of segmentations appeared in [18], [19]. None of the previous strategy for scene labeling used a purity criterion on the class distributions. Combined to the optimal cover strategy, this purity criterion is general, efficient and could be applied to solve different problems.

In order to improve modeling of long range dependencies, Pinheiro et al. [1] first revealed the use of large input patches to consider larger contexts. However, it decrease the resolution of the final parsing image, and a large redundancy of overlap pixels would make the learning inefficient. Hence, they introduced Recurrent Convolutional Networks (RCNNs) for scene labeling. RCNNs train different input patch sizes (the instances) recurrently to learn increasingly large contexts for each pixel, whilst ensuring that the larger contexts are coherent with the smaller ones.

3. FEATURE EXTRACTION FOR SCENE PARSING

This first representation typically suffered from two problems when using a classical convolutional neural network, where the image is divided following a grid pattern: (1) the window considered rarely contains an images that is properly centered and scaled, and therefore offers a poor observation basis to predict the class of the underlying object, (2) integrating a large context involves maximizes the grid size, and therefore the dimensionality P of the input; given a finite amount of training data, it is then necessary to enforce some invariance in the function f itself. It is achieved by pooling layers. In the second representation, the image is seen as an edge-weighted graph,[17] on which one or several over segmentations can be constructed.
4. RECURRENT NEURAL NETWORK

In temporal dependency modeling for chain structured data, Recurrent neural networks (RNNs) have achieved great success such as natural language and speeches. It is a class of artificial neural network that has recurrent connections, which provide the network with memory. It connections form a directed cycle. RNN can be perform scene labeling using the methods Long Short Term Memory network, Directed acyclic graph.

4.1. Long Short Term Memory network

Long Short Term Memory (LSTM) recurrent neural network architecture take into account the local and global dependencies in a single process for scene parsing. We can skip any additional processing or conditions like multi-scale or different patch sizes to solve the scene labeling task with the least human or machine’s effort. LSTM recurrent networks [20] were originally developed for sequence learning. Connected cells are included recurrently in these networks to learn the dependencies between two time frames, then transfer the probabilistic inference to the next frame. The LSTM memory block stores and retrieves this information over short or long periods of time.

LSTM networks have LSTM networks have been successfully applied to many tasks such as handwriting [21] and speech recognition [22]. They were extended to multi-dimensional learning which improved handwriting systems. Image processing tasks were also considered with the multi-dimensional model [23]. This work illustrate their usage for image labeling, binarization or texture analysis, but only on very simple data (simple digit images, graphical or texture data).

It focus particularly on natural scene image labeling. RNN-based approaches are known to be difficult to train with high noise and large data. Particularly in large data, long-term dependencies are vanished while the information is accumulated by the recurrence. Large scene images is supported which contain huge variations of instances for each label and which can have an enormous size. We show how LSTM networks can be generalized well to any vision-based task and efficiently learnt without any task specific features. The short and long-range contextual information are also learned by LSTM networks by end-to-end entirely supervised training.

4.2. Directed acyclic graph

Recurrent neural networks (RNNs) [24] are introduced to address this issue by modeling the contextual dependencies of local features. Specifically, we adopt undirected cyclic graphs (UCGs) to represent the connectivity of pixels in images. Due to the loopy property of UCGs, RNNs are not directly applicable to UCG structured images. so, the UCG can moved to several directed acyclic graphs (DAGs). In other words, an UCG can be mentioned as several DAG-structured images. Then, we introduce the DAG-RNNs, a simplification of RNNs [24], to process DAG structured images. Each hidden layer is generated independently through applying DAG-RNNs to the corresponding DAG-structured image, and they are aggregated to produce the context-aware feature maps. In this case, the local representations are able to implant the conceptual gist of the image, so their discriminative power is enhanced remarkably.

5. OUR APPROACH

In this paper the proposed method can be fully recurrent network it can be used the Bi directional recurrent neural network for parsing the images. Figure 2 shows the structure of Bi directional RNN use a finite sequence to predict or label each element of the sequence based on both the past and the future context of the element. The outputs of BRNN can be processed by the sequence from left to right, the other one from right to left is done. The combined outputs are the predictions of the teacher-given target signals. Bi directional RNN can be combined with the Long Short Term Memory to produce the output. Comparing to another recurrent neural network like Time Delay neural network have limitation on input data occurrence. In Bi Direction RNN there is no such limitation and number of input can be used in the system. The basic idea of BRNNs is to connect two hidden layers of opposite directions to the same output. By this structure, the output layer get information from past and future states. The principle of BRNN is to split the neurons of a regular RNN into two directions, one for positive time direction(forward states), and another for negative time direction(backward states). BRNN can be trained using similar algorithms compared to RNN. Training a BRNN can be achieved by gradient-based method. Back Propagation through time (BPTT) [25] is usually used to calculate the gradients. The BRNN can be processed,

\[
\Pr(c_1, c_2, \cdots, c_T | x_T^{T}) \\
= \prod_{t=1}^{T} \Pr(c_t | c_{t+1}, c_{t+2}, \cdots, c_T, x_t^{T}) \\
= \prod_{t=1}^{T} \Pr(c_t | c_1, c_2, \cdots, c_{t-1}, x_t^{T}).
\]

The probability term within the product is the conditional probability of an output class given all the input to the right- and left-hand side plus the class sequence on one side of the currently evaluated input vector.
6. Experiment

The SiftFlow dataset has 1500 images generally captured from 8 typical outdoor scenes. Every image has 256 × 256 pixels, which belong to one of the 33 semantic classes. We adopt the training/testing split protocol (2400/200 images) provided by [15] to perform our experiments. Following the 85%-15% criterion, the class frequency threshold η = 0.04. Statistically, out of 33 classes, 26 of them are regarded as infrequent class. The graphical visualization of the weights for different classes.

![Figure 2.png](image)

**Figure 2.** shows the structure of Bi Directional RNN

![Figure 3.png](image)

**Figure 3.** Input image

![Figure 4.png](image)

**Figure 4.** Scene Parsing using B-RNN

The quantitative results are listed in Table 1, where the upper part presents the performance of methods under set-ting 1. Our baseline BRNN achieves very promising results, which proves the effectiveness of the convolution layer. We also notice that results of CNN-65 fall behind CNN-65-ENN on the average class accuracy.

![Table 1.png](image)

**Table 1:** Quantitative performance of our method on the siftFlow dataset. The numbers (in brackets) following the BRNN denote the neighbor-hood system of the UCG.

This result shows that the proposed class weight function significantly the recognition accuracy for rare classes. By adding DAG-RNN(24), our network reaches 81.1% (48.2%) on global (class) accuracy 4, which outper-forms the baseline (CNN-65-ENN) by 5% (11.5%). Mean-while, we observe promising accuracy gain (global: 0.6% / class: 5.6% ) by switching Directed acyclic graph-RNN to BRNN , in which we believe that long-range dependencies are better captured as information propagation paths in BRNN(24) are shorter than those in DAG.

7. Conclusion

In this paper, Fully Recurrent network used Bi Directional RNN to model the contextual dependencies of local features and small objects. BRNN are able to encode the global gist of images to local features, so their discriminative power can be
enhanced. Furthermore introduced class weighting function is experimentally proved to be effective towards the recognition enhancement for rare classes. Integrate with the convolution layer, our BRNN achieve state-of-the-art results on one challenging scene labeling benchmarks. We also demonstrate that useful contexts are captured by our BRNN, which is helpful for generating smooth and semantically sensible labeling maps in practice.

8. FUTURE WORK

In this section, we investigate the effects of our BRNN for each class. The class accuracy for the SiftFlow dataset is listed in Table 1. That the contextual information encoded through our Directed acyclic graph-RNN(24) is beneficial for almost all classes. The local representations are highly discriminative in this situation, our BRNN further tremendously improves their representative power for rare classes. Statistically, Observe a phenomenal 8.6% accuracy gain for rare classes. Under both settings, modeling the dependencies among local features enables the classification to be contextual aware. Therefore, the local ambiguities are mitigated to a large extent. However, to ob-serve commensurate accuracy improvements for extremely-small-size and rare ‘object’ classes we conjecture that the weak local information may have been overwhelmed by context.

References