Pattern Recognition 80 (2018) 32–41

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Deep contextual recurrent residual networks for scene labeling

T. Hoang Ngan Le*, Chi Nhan Duong, Ligong Han, Khoa Luu, Kha Gia Quach, Marios Savvides

Department of Electrical and Computer Engineering, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA

ARTICLE INFO

Article history:
Received 20 May 2017
Revised 25 October 2017
Accepted 7 January 2018
Available online 31 January 2018

Keywords:
Recurrent network
Residual learning
Visual representation
Context modeling
Scene labeling

ABSTRACT

Designed as extremely deep architectures, deep residual networks which provide a rich visual representation and offer robust convergence behaviors have recently achieved exceptional performance in numerous computer vision problems. Being directly applied to a scene labeling problem, however, they were limited to capture long-range contextual dependence, which is a critical aspect. To address this issue, we propose a novel approach, **Contextual Recurrent Residual Networks (CRRN)** which is able to simultaneously handle rich visual representation learning and long-range context modeling within a fully end-to-end deep network. Furthermore, our proposed end-to-end CRRN is completely trained from scratch, without using any pre-trained models in contrast to most existing methods usually fine-tuned from the state-of-the-art pre-trained models, e.g. VGG-16, ResNet, etc. The experiments are conducted on four challenging scene labeling datasets, i.e. SiftFlow, CamVid, Stanford background and SUN datasets, and compared against various state-of-the-art scene labeling methods.

1. Introduction

Scene labeling has played an important role in many applications in computer vision and machine learning. This problem is known as semantic segmentation or scene parsing and refers to associating each pixel with one semantic class in a scene image. This task is very challenging as it implies solving jointly detection, segmentation and multi-label recognition problems. To address this issue, a large body of researches have recently proposed different approaches mainly focusing on contextual information via graphical model [2–7] or recurrent neural network [5,8,9] or enriching visual representations via convolutional neural network [10–15]. Yang et al. [5] focuses on rare object classes to achieve richer semantic understanding of visual scenes, compared to common background classes

However, scene labeling problem in the real world needs both information of the context dependencies and visual representation. For example, powerful visual representation is capable to discriminate a road from a beach or sea from the sky; but it may not effective enough to tell a patch of sand belongs to the side of a road or to a beach and it is almost impossible to distinguish a pixel belonging to sky from a pixel belonging to sea by only looking at a small patch around them. In such circumstance, pixels can not be labeled based only on short-range context i.e. a small region around them. Clearly, the context presented in the whole scene can show its advantage to describe them.

Indeed, the roles of contextual information i.e. short-range and long-range context and powerful descriptive visual representation are equally important in the scene labeling problem.

To effectively address the scene labeling problem, we propose a novel deep network named **Contextual Recurrent Residual Network (CRRN)** that inherits all the merits of sequence learning information and residual learning in order to simultaneously model *long-range contextual information* and learn *powerful visual representation* within a *single deep network*. Our proposed CRRN deep network consists of three parts corresponding to sequential input data, sequential output data and hidden state. Each unit in hidden state is designed as a combination of two components: a context-based component via sequence learning and a visual-based component via residual learning. That means, each hidden unit in our proposed CRRN simultaneously (1) learns long-range contextual dependencies via context-based component. The relationship between the current unit and the previous units is performed as sequential information under an undirected cyclic graph (UCG) and (2) provides powerful encoded visual representation via residual component which contains blocks of convolution and/or batch normalization layers equipped with an identity skip connection. Furthermore, unlike previous scene labeling approaches [8,9,16], our method is not only able to exploit the long-range context and vi-

* Corresponding author.
E-mail addresses: thihoanl@andrew.cmu.edu (T.H.N. Le), chinhand@andrew.cmu.edu (C.N. Duong), kluu@andrew.cmu.edu (K. Luu), kquach@andrew.cmu.edu (K.G. Quach), msavvid@ri.cmu.edu (M. Savvides).

https://doi.org/10.1016/j.patcog.2018.01.005
0031-3203/

sual representation but also formed under a fully-end-to-end trainable system that effectively leads to the optimal model. In contrast to other existing deep learning network which are based on pretrained models, our fully-end-to-end CRRN is completely trained from scratch.

2. Related work

Scene labeling is arguably one of the hardest challenges in computer vision. It requires the algorithms to have much more finesse than those that are only required to tackle image scale object recognition for instance. Nonetheless, a lot of studies have focused on this challenging problem in the past and have made considerable progress recently. Generally, scene labeling methods can be divided into three categories as follows.

2.1. Graphical model approaches

In the past, using traditional vision techniques, scene labeling was approached from a undirected graphical model paradigm utilizing Markov Random Fields (MRF) and Conditional Random Fields (CRF). In [2], He, et al. proposed contextual feature incorporated into CRF which combines the outputs of several components i.e. image-label mapping, patterns within the label field, fine/coarse resolution patterns. The context information was then continue studied under an auto-context algorithm in [17]. The context information is learn via a discriminative probability maps which is then applied to vision tasks and 3D Brain image segmentation. In contrast to much previous work on structured prediction, Munoz et al. [4] directly trained a hierarchical inference procedure inspired by the message passing mechanics of some approximate inference procedures in graphical models. Different from the previous approaches, which usually employ Conditional Random Fields (CRFs) or hierarchical models to explore contextual information, Zhou et al. [7] propose a novel flexible segmentation graph (FSG) representation to capture multi-scale visual context for scene labeling problem by establishing a contextual fusion model to formulate multi-scale context. When CNN have shown great ability of learning features and attained remarkable performance, Bu et al. [6] design a Embedded Deep Networks (IEDNs) that aims to inherits the merits of both CNN and CRFs by considering CRFs model as one type layer of deep neural networks. Through the IEDNs, the network can learn hybrid features, the advantages of which are that they not only provide a powerful representation capturing hierarchical information, but also encapsulate spatial relationship information among adjacent objects

2.2. ConvNet-based approaches

In recent years, deep learning techniques have started to become ubiquitous in scene labeling. One of the first studies to apply convolutional neural networks (deep CNNs) to scene labeling was [10], which stacked encompassing windows from different scales to serve as context. This inspired other studies in which fully convolutional networks were used instead [11] utilizing higher model complexity. Both these techniques used filter based models to incorporate context. Recently, recurrent models have started to gain popularity. For example [12], where the image is passed through a CNN multiple times in sequence *i.e.* the output of the CNN is fed into the same CNN again. As an interesting study, Zheng et al. [13] modeled a CRF as a neural network that is applied iteratively to an input, thereby qualifying as a recurrent model. Inference is done through convergence of the neural network output to a fixed point. Recently, deep residual networks (ResNets) [18] have emerged as a family of extremely deep architectures showing compelling accuracy and desirable convergence behaviors. They consist of blocks of convolutional and/or batch normalization layers equipped with an identity skip connection. The identity connection helps to address the vanishing gradient problem and allows the ResNets to robustly train using standard stochastic gradient descent despite very high model complexity. This enables ResNets to extract very rich representations of images that perform exceedingly well in image recognition and object detection challenges [19]. The extremely deep architectures in ResNets show compelling accuracy and robust convergence behaviors and achieve state-of-the-art performance on many challenging computer vision tasks on ImageNet [20], PASCAL Visual Object Classes (VOC) Challenge [21] and Microsoft Common Objects in Context (MS COCO) [22] competitions. To deal with rich structures exist in SAR (Synthetic aperture radar) images, Duan et al. [14] makes use of convolutional-wavelet neural networks (CWNN) and Markov Random Field (MRF) by replacing the conventional pooling in CNN with a wavelet constrained pooling layer. CNN has been proved to be effective in many areas from natural images [10,11,14,18,19] to facial analysis [23–25] to drive safety [15,26–28]. Recently, Le et al. [15] incorporates grammatical structure telling the relationship between parts into CNNs to support driver behavioral situational awareness (DB-SAW). Le et al. [15] uses prior knowledge of deep probability map to define within-subgraph and between-subgraph which deep features capable of representing both information of feature and shape. Nonetheless, they are feed forward models that do not explicitly encode contextual information and typically cannot be applied to sequence modeling problems. On the other hand, they are able to *effectively learn visual representations but limited to model long-range context explicitly*.

2.3. Recurrent-based approaches

In recent years, vision data is being interpreted as sequences leading to the successful application of RNNs (and their variants, e.g. Long-Short Term Memories (LSTMs), Gated Recurrent Units (GRUs), etc.) to vision problems. For instance, Zuo et al. [29] and Graves [30] applied 1-D RNNs and multi-dimensional RNNs to model contextual dependencies in object recognition/image classification and offline handwriting recognition respectively. 2-D LSTMs instead were applied to scene parsing [16], Lately, scene labeling [8], object segmentation [31], have been reformulated as sequence learning, thereby allowing RNNs to be applied directly. Scene labeling, in particular, has seen the use of RNNs coupled with Directed Acyclic Graphs (DAGs) to model an image as a sequence [8,9,16]. There have also been a few studies that utilize RNNs to compute visual representations [32,33]. Be designed as similar fashion as [8,9], ION [34] inherits the merits from both CNN and RNN. However, the two components CNN and RNN are treated separately as two separated consecutive elements, namely, CNN is for feature extraction at different scales whereas RNN is contextual learning at different directions. Clearly, RNN-based approaches are *effective in context modeling but lack the ability to learn visual representation.*

Drawbacks of the current approaches are:

(1) The ConvNet-based approaches model makes use of convolutional filters which allow them to learn the short-range context of surrounding neighbors designed by these filers. Therefore they are limited to generalize to long-range contexts dependencies. (2) The RNN-based approaches usually utilize a feature extractor that is independent of the sequence modeling framework, in many cases being trained component wise and not end-to-end [1]. (3) Purely RNN based approaches fail to extract robust visual features during sequence learning itself. This is due to simple linear models being used as the recurrent internal models.