Pattern Recognition 84 (2018) 226–236

Contents lists available at ScienceDirect

Pattern Recognition

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

MoE-SPNet: A mixture-of-experts scene parsing network

Huan Fu[a,*], Mingming Gong[b], Chaohui Wang[c], Dacheng Tao[a]

[a] UBTECH Sydney AI Centre, SIT, FEIT, The University of Sydney, J12 Cleveland St, Darlington, NSW 2008, Australia
[b] Department of Biomedical Informatics University of Pittsburgh Cublicle 520c, 5607 Baum Bouevard, Pittsburgh, PA 15206, America
[c] Laboratoire d'Informatique Gaspard Monge - CNRS UMR 8049, Université Paris-Est, 77454 Marne-la-Vallée Cedex 2, France

ARTICLE INFO

Article history:
Received 7 January 2018
Revised 19 June 2018
Accepted 16 July 2018
Available online 17 July 2018

Keywords:
Scene parsing
Mixture-of-experts
Attention
Convolutional neural network

ABSTRACT

Scene parsing is an indispensable component in understanding the semantics within a scene. Traditional methods rely on handcrafted local features and probabilistic graphical models to incorporate local and global cues. Recently, methods based on fully convolutional neural networks have achieved new records on scene parsing. An important strategy common to these methods is the aggregation of hierarchical features yielded by a deep convolutional neural network. However, typical algorithms usually aggregate hierarchical convolutional features via concatenation or linear combination, which cannot sufficiently exploit the diversities of contextual information in multi-scale features and the spatial inhomogeneity of a scene. In this paper, we propose a mixture-of-experts scene parsing network (*MoE-SPNet*) that incorporates a convolutional mixture-of-experts layer to assess the importance of features from different levels and at different spatial locations. In addition, we propose a variant of mixture-of-experts called the adaptive hierarchical feature aggregation (*AHFA*) mechanism which can be incorporated into existing scene parsing networks that use skip-connections to fuse features layer-wisely. In the proposed networks, different levels of features at each spatial location are adaptively re-weighted according to the local structure and surrounding contextual information before aggregation. We demonstrate the effectiveness of the proposed methods on two scene parsing datasets including PASCAL VOC 2012 and SceneParse150 based on two kinds of baseline models FCN-8s and DeepLab-ASPP.

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

Scene parsing or semantic image segmentation, which predicts a category-level label (such as “sky”, “dog” or “person”) for each pixel in a scene, is an important component in scene understanding. A perfect parsing can contribute to a variety of applications including unmanned vehicles, environmental reconstruction, and visual SLAM. Many other fundamental computer vision problems can benefit from the parsing of an image, such as medical image analysis, tracking, and object detection [1–3]. However, scene parsing is a very challenging high-level visual perception problem as it aims to simultaneously perform detection, reconstruction, segmentation, and multi-label categorizing [4,5].

Since feature representation is critical to pixel-level labeling problems, classical methods focus on designing handcrafted features for scene parsing [6]. Since the handcrafted features alone can only capture local information, probabilistic graphic models such as conditional random fields (CRFs) are often built on these features to incorporate smoothness or contextual relationships between object classes [7]. Recently, deep learning approaches such as deep convolutional neural networks (DCNNs) have earned immense success in scene parsing. In particular, fully convolutional networks (FCNs)-based approaches have demonstrated promising performance on several public benchmarks [5,8–12].

A common strategy adopted in all the CNN-based methods is to aggregate multi-scale/level features from multiple CNN layers [5] or from a specific layer [8], which is a key component to obtain high-quality dense predictions because the multi-level features capture different levels of abstractions of a scene. The standard way to combine hierarchical features/predictions is to either concatenate multi-level features [13–20] or equivalently aggregate the prediction maps by average pooling [5]. However, the linear feature aggregation methods are not able to evaluate the relative importance of the semantic and spatial information in each level of features. The information at different scales is complementary because the higher-level convolutional features contain larger-scale contextual information which is beneficial for classification, while the lower-level features have higher spatial resolution which produces finer segmentation masks [21]. The information at different scales is also complementary since they are from different recep-

[*] Corresponding author.
E-mail addresses: hufu6371@uni.sydney.edu.au (H. Fu), mig73@pitt.edu (M. Gong), chaohui.wang@u-pem.fr (C. Wang), dacheng.tao@sydney.edu.au (D. Tao).

https://doi.org/10.1016/j.patcog.2018.07.020
0031-3203/© 2018 Elsevier Ltd. All rights reserved.

tive fields. There is thus a trade-off between the semantic and the spatial information. In addition, the average pooling ignores the spatial inhomogeneity of a scene, which is improper since different objects may prefer features from different scales/levels. For example, textured objects such as “grass” and “trees” can be easily distinguished from lower-level features while textureless objects like “bed” and “table” require higher-level features to capture the global shape information.

In this paper, we propose a mixture-of-experts [22] scene parsing network (*MoE-SPNet*) which learns to aggregate multi-level convolutional features according to the image structures. Specifically, we treat each network branch that contains a specific level/scale of features/predictions as an expert and aggregate them using the weights generated by a trainable convolutional gating network. The gating network also has a convolutional architecture and outputs a weight map for the entire image. The proposed MoE-SPNet is motivated by the following three observations: (1) The lower-level convolutional features contain more precise boundary information but tend to yield more incorrect predictions, while the higher-level features contain more contextual and semantic information but less spatial information. (2) Different levels/scales of features reflect the visual properties of different-sized objects because they are extracted by receptive fields with different sizes. Notably, small objects are more likely to be misclassified to their background if using higher-level features because larger receptive fields introduces much noise to these small objects. (3) The relative importance of different levels of features varies with spatial location; it relies on the local image structure and surrounding contextual information. Obviously, a linear combination of these features by average pooling cannot capture the homogeneity of a scene and assess the importance of different feature levels. On the contrary, the proposed MoE-SPNet overcomes the limits of linear combination by aggregating different level of features in a nonlinear and adaptive way.

Since MoE-SPNet is only able to adaptively aggregate multi-scale features generated from a single CNN layer, we further propose a variant of MoE called adaptive hierarchical feature aggregation scheme (AHFA) which can be incorporated into the existing parsing networks that aggregate hierarchical features using skip-connections. For example, the original FCN architecture combines features from the last convolutional layer with previous layers by successive upsampling and aggregation. Employing AHFA will enable the parsing networks such as FCN to learn weights at each stage and aggregate the features adaptively as done in MoE-SPNet. In this paper, we focus on exploiting AHFA for the original FCN, leading to a new network architecture denoted as *FCN-AHFA*.

We demonstrate the effectiveness of our MoE-SPNet and FCN-AHFA on two challenging benchmarks for scene parsing, PASCAL VOC 2012 [23] and SceneParse150 [24], and achieve the state-of-the-art or comparable results. Also, the experimental results show that our MoE-SPNet and FCN-AFHA consistently improve the performance of all the evaluated baseline networks, and thus demonstrate the value of the proposed methods. In addition, the produced weight maps can help us understand the reason that some image structures prefer higher-level convolutional features while others prefer lower-level features.

2. Related work

Segmentation is a fundamental problem in scene understanding. While some works focus on low-level segmentation which segments a scene into some regions that share certain characteristics or computed property, such as color, intensity, or texture [25–28], high-level segmentation (scene parsing or semantic segmentation), which assigns a category-level label to each pixel of a scene, receives much attention recently.

In the past decade, the successful scene parsing methods rely on handcrafted local features like colour histogram and textons [6,29–33], and shallow classifiers such as Boosting [6,34], Random Forests [35,36], Support Vector Machines [37]. Due to the limited discriminative power of local features, a lot of efforts have been put into developing probabilistic graphical models such as CRFs to enforce spatial consistency and incorporate rich contextual information [7,38–40]. Recently, deep learning methods typified by DCNNs have achieved state-of-the-art performance on various computer vision tasks, such as image classification and multi-class object detection.

Also, the DCNN architectures such as VGG [41] and ResNet [42] originally developed for image classification have been successfully transferred to scene parsing. Specifically, Long et al. [5] proposed the fully convolutional network (FCN) which applied DCNNs to the whole image and directly produced dense predictions from convolutional features, making it possible to get rid of bottom-up segmentation steps [43] and train the parsing network in an end-to-end fashion.

The impressive performance of FCNs is largely due to the aggregation of multi-level or multi-scale features/predictions. There are mainly two types of aggregation methods: share-nets and skip-nets [44]. The skip-nets, which merge multi-level features/predictions from a single network, are computationally more efficient than the share-nets. Furthermore, they have been refined to enable end-to-end training by normalizing the features from different levels. For example, Hariharan et al. [4] concatenated the multi-level features together after certain normalization methods like L2 normalization. However, the concatenation of hierarchical features results in high-dimensional features and is thus time-consuming. The FCN-8s [5] model aggregated features from the last three convolutional blocks by averagely pooling over layers. Similarly, Chen et al. [45] combined the features which were extracted by applying multi-layer perceptrons on the original image and the pooling layers. However, linear combination of multi-scale features does not sufficiently exploit the geometric properties, contextual information, and the spatial-semantic tradeoff. Recently, Ghiasi and Fowlkes [21] found that directly summing up multi-scale features cannot achieve desirable results, as the learned parameters tended to down-weight the contribution of lower-level features (higher resolution) to suppress the effects of noisy predictions. They proposed the laplacian pyramid refinement approach which computed a boundary mask from higher-level semantic predictions to filter out the noisy predictions in lower-level features. However, we aim to learn the mask weights from multi-level features instead of calculating a boundary mask by manually designed mathematical operations.

Share-nets combine features from shared networks built on multiple rescaled images. For example, Farabet et al. [43] transformed the raw image through a laplacian pyramid, and each level of which was fed into a CNN. The produced sets of feature maps of all scales were concatenated to form the final representation. Similarly, Lin et al. [46] resized the original image to three scales and concatenated the multi-scale features. Aside from concatenation, average pooling [47] and max pooling [48] were adopted over scales to merge multi-scale features. However, average or max pooling either treats the multi-scale features equally or losses too much information. Targeting this problem, Chen et al. [44] proposed the scale attention method which uses the attention model [49] over scales to focus on the features from the most relevant scales. Instead of aggregating multi-scale features at one time, Pinheiro and Collobert [50] proposed a multi-stage approach which fed multi-scale images successively to a recurrent convolutional neural network. Although the share-nets obtain much better performance, they are computationally more expensive than the single scale networks. Most recently, Chen [8] developed an atros spa-

Download English Version:

https://daneshyari.com/en/article/6938661

Download Persian Version:

https://daneshyari.com/article/6938661

Daneshyari.com