Road Scene Understanding for Autonomous Driving Based on Deep Residual Learning: Post-print

Authors: Song Rui, Zhiping Shi, Qu Ying, Shao Zhenzhou, Guan Yong

Date: 2018-06-19T00:00:00+00:00

Abstract

With the rapid development of road scene understanding technology, significant progress has been achieved in the field of autonomous driving. In related tasks, the real-time performance and accuracy of road segmentation, classification, and vehicle detection constitute a critical safety concern. To address this, we propose a method based on deep residual learning with an encoder-decoder network architecture. On the one hand, the encoder network architecture employs residual networks at different levels to extract abstract features in high-dimensional space, which are shared among the three subsequent tasks; on the other hand, the decoder network architecture adopts a parallel computation mechanism for sub-tasks, wherein road segmentation, vehicle detection, and road classification tasks are executed simultaneously. Additionally, a fully convolutional neural network is utilized to upsample the extracted image features to solve the road segmentation problem. Ultimately, experimental results demonstrate that the processing frame rate can achieve over 15 fps while maintaining high accuracy.

Full Text

Preamble

Road scene understanding for autonomous driving via deep residual learning

Song Rui[1] , Shi Zhiping[1] †, Qu Ying[2], Shao Zhenzhou[1] , Guan Yong[1] (1. a. Beijing Advanced Innovation Center for Imaging Technology; b. Beijing Key Laboratory of Light Industrial Robot & Safety Verification, College of Information Engineering, Capital Normal University, Beijing 100048, China; 2. Department of Electrical Engineering & Computer Science, The University of Tennessee, Tennessee 37996, USA)

Abstract: The autonomous driving field has made significant progress with the rapid development of road scene understanding techniques. Safety is a critical concern that depends on real-time and accurate performance in related tasks including road segmentation, road classification, and vehicle detection. To address this, we propose an approach based on deep residual learning with an encoder-decoder network structure. On one hand, the encoder network structure uses different layers of residual networks to extract abstract features in high dimensions, which are shared across the three subsequent tasks. On the other hand, the decoder network structure adopts a parallel computing mechanism for sub-tasks, where road segmentation, vehicle detection, and road classification tasks are executed simultaneously. Additionally, fully convolutional networks are used to upsample the extracted features to specifically solve the road segmentation problem. Experimental results demonstrate that the processing rate can effectively reach over 15 fps while guaranteeing high accuracy.

Key words: road scene understanding; deep residual learning; encoder-decoder structure; fully convolutional networks

0 Introduction

With the rapid development of artificial intelligence technology, autonomous driving has attracted increasing attention as it transforms people' s daily travel modes. Based on safety considerations for human life, autonomous driving technology requires high stability, accuracy, and the ability to process various complex road scenes in real-time. Currently, deep learning technology [1] is the mainstream approach in this field and has been widely applied to road segmentation, classification, and vehicle detection tasks to enhance autonomous vehicles' understanding of driving scenarios (Figure 1). Therefore, achieving faster and more accurate road scene understanding is of great research significance in autonomous driving.

Currently, typical solutions for the three aforementioned road scene understanding tasks are as follows:

a) **Road segmentation task.** As a semantic segmentation task applied to autonomous driving scenarios, Long et al. [2] proposed using deep neural network structures to solve road segmentation problems. They were the first to achieve end-to-end semantic segmentation using fully convolutional neural networks. Subsequently, Paszke et al. [3] proposed an encoder-decoder network model that uses neural networks for image feature extraction to improve algorithm generalization and network operation speed.

b) **Vehicle detection task.** Ren et al. [7] proposed a region proposal-based approach that improves upon the large-scale computational problems of traditional sliding window methods. It first uses a region proposal network to generate multiple candidate boxes for detected objects, then trains different neural network models to improve confidence, and finally selects the

result with maximum confidence as the detection outcome. Additionally, Redmon et al. [8] proposed a revolutionary region proposal-based object detection framework that divides the entire image into S×S grids and predicts candidate boxes for all grids simultaneously, achieving end-to-end real-time object detection.

c) **Road classification task.** Since Krizhevsky et al. [9] proposed the AlexNet structure and achieved breakthrough progress in applying neural networks to classification tasks, deep neural networks have developed rapidly. In the ILSVRC challenge, many network structures with complex hierarchies emerged, such as VGG and GoogleNet [10]. In 2015, He et al. [11] proposed the deep residual network based on original network structures, introducing the residual concept for the first time and using block structures to manage network layers. This greatly improved the overfitting problem caused by excessively deep networks and considered the impact of low-dimensional image features discarded during convolutional downsampling on classification tasks, significantly improving object classification accuracy.

In addition, for the crucial feature extraction component in neural networks, the VGG network structure with simple and uniform architecture is often used for high-dimensional image feature extraction. However, the VGG network structure has some limitations, as its large number of parameters reduces operation speed and cannot meet real-time application requirements in autonomous driving.

To address these problems, this paper adopts a typical encoder-decoder network for road scene understanding tasks in autonomous driving. First, the encoder structure uses deep residual networks (ResNet) to extract image features. Deep residual networks introduce shortcut connection structures that better fuse low-dimensional and high-dimensional image features, greatly improving accuracy while increasing depth. The decoder structure utilizes the extracted features to simultaneously complete road segmentation, vehicle detection, and road classification tasks. Finally, experiments and training on the KITTI dataset [12] compare different network layers and structures, ultimately improving the processing frame rate to over 15 fps, which greatly enhances image processing speed, improves vehicle perception of road environments, and ensures the stability, accuracy, and timeliness of autonomous driving technology.

1 Encoder-Decoder Network Structure

Encoder-decoder network structures can fully utilize both deep-level and shallow-level salient features of images by combining features from different layers to improve task accuracy. In this paper, the encoder portion involves feeding images into neural networks with complex convolutional structures to extract deep-level abstract features [13], which can be shared across multiple sub-tasks. The decoder portion then connects to corresponding task-specific

processing. The network structure of this paper and the specific layer outputs and parameter settings of important encoder and decoder layers are shown in Figure 2 [Figure 2: see original paper], which effectively completes road segmentation tasks as well as vehicle detection and road classification tasks.

Among many typical segmentation methods, the VGG network [4] is commonly used for feature extraction tasks, where networks like SegNet [5] and MultiNet [6] adopt this structure for high-dimensional abstract feature extraction to complete road segmentation tasks, achieving good operation speed and accuracy.

2 Feature Extraction Based on Deep Residual Learning

Deep convolutional neural networks have the powerful advantage of obtaining complex, deeper-dimensional image features from large-scale training data. This paper adopts the deep residual network structure that won the championship in the ILSVRC & COCO 2015 challenges. Unlike previous neural network structures, this network introduces residual learning modules with shortcut connections. Based on original convolutions, it introduces a linear connection between the input and output of layers, which not only effectively avoids overfitting problems caused by excessive depth but also better utilizes low-dimensional image features, effectively improving accuracy, as shown in Figure 3 [Figure 3: see original paper].

Additionally, the network uses standard 3×3 convolution kernels and ReLU activation functions for activation, including typical convolution and max pooling operations. The model contains 50, 101, and up to 152 layers. Compared with the VGG network structure, it reduces network model parameters and adds residual branches, using block structures to manage network layers. Moreover, this network has strong transfer capabilities and can effectively complete multiple tasks including road classification, vehicle detection, and road segmentation. It can also adopt different network layer structures according to different tasks and training data volumes. Therefore, this paper uses pre-trained residual networks for image feature extraction tasks. Table 1 provides detailed parameter configurations of the convolutional layers for the different network structures used in our experiments. Due to the large number of parameters, all experiments in this paper adopt a fine-tuning approach based on pre-trained models to optimize network parameters and improve model adaptation to specific datasets.

3.1 Road Segmentation Based on Fully Convolutional Networks

The fully convolutional network structure represents a key advancement in semantic segmentation, first achieving end-to-end semantic segmentation of images. It proposes an operation inverse to convolution. Under the premise that convolutional downsampling in feature extraction discards various low-dimensional image features, the features trained by the residual network are passed through a 1×1 convolutional layer to readjust dimensions for the seg-

mentation task. The introduction of upsampling operations avoids the repetitive storage and computation problems caused by using pixel blocks. The method adopts an approach opposite to convolution, using deconvolution to complete upsampling operations. Skip connections are introduced to organically combine low-dimensional features with high-dimensional features. Additionally, by combining with conditional random fields and introducing dilated convolution structures [14], the receptive field range can be increased without reducing dimensions, obtaining more accurate segmentation results.

3.2 Vehicle Detection Based on Proposals

The vehicle detection task mainly draws on proposal-based methods, primarily adopting the FastBox approach inspired by successful models such as YOLO [8], using region-of-interest pooling to fully utilize high-dimensional features obtained from deep residual network training. Similar to the segmentation task, the encoder features first need to be passed through a 1×1 convolutional layer to adjust network dimensions, followed by a bottleneck layer composed of multiple 1×1 convolutions that adjusts the output to 6 channels. The first two channels represent the semantic meaning of the detected object, with values indicating confidence within the bounding box; the last four channels represent the coordinates and dimensions of the bounding box. This yields a coarse estimation result. However, such predictions are inaccurate, so this paper introduces a rescaling layer that corrects the original prediction results by utilizing high-dimensional and implicit features from image regions other than those selected by non-maximum suppression. Through region-of-interest pooling, the final detection results are obtained by adjusting dimensions via 1×1 convolution.

3.3 Road Classification Based on Fully Connected Structure

For the road classification problem, this paper adopts a fully connected layer structure typical in neural networks. The features trained by the residual network are first passed through a 1×1 convolution to adjust image dimensions. Using a multi-classifier with a fully connected layer structure employing softmax activation functions, the final prediction classification results are obtained using one-hot encoding based on final proportional scores.

4 Experimental Results and Analysis

To evaluate the performance of the autonomous driving road scene understanding algorithm based on deep residual networks, this paper conducted two sets of experiments. The first set primarily verifies the algorithm' s generalizability and necessity for solving practical problems. In the second set, the proposed algorithm is compared with ENet [3], FCN [2], SPL [15] for road segmentation, and algorithms from the KITTI leaderboard for vehicle detection.

The experiments primarily use the KITTI dataset [12], which is rich in autonomous driving domain data. For road segmentation and classification meth-

ods, the KITTI Road dataset [16] is used for evaluation, which includes 289 training images and 290 test images. Figure 4 [Figure 4: see original paper] shows sample images from the KITTI Road dataset, mainly containing three types of road images: single lane line, multiple lane lines, and no lane lines. The first row shows single lane line data, the second row shows multiple lane line data, and the third row shows no lane line data. For the vehicle detection task, the KITTI Object dataset is used for training and evaluation, with detected objects divided into three difficulty levels: easy, moderate, and hard. The segmentation task is evaluated using maximum F1 value [16] and average precision, while the detection task uses average precision for these three categories as the evaluation standard, and the classification task uses average precision for evaluation. The machine configuration for this experiment is shown in Table 2 .

4.1.1 Loss Function

The loss functions mainly include those for road segmentation, classification, and vehicle detection tasks. Since road segmentation and classification tasks use the same type of loss function, we take the segmentation task as an example here. The segmentation task uses cross-entropy as the loss function, defined as in Equation (1).

Where: p is the predicted value; q is the ground truth value; c is the set of categories; I is the member of the mini-batch.

The loss function for vehicle detection consists of the sum of cross-entropy for confidence and L1 loss for bounding box coordinates, defined as follows:

Where: p is the predicted value; q is the ground truth value; I is the member of the mini-batch. The bounding box mainly contains four parameters: the center point coordinates and height of the bounding box.

4.1.2 Initialization

The encoder stage is initialized using deep residual network weights pre-trained on ImageNet. The vehicle detection decoder weights are initialized randomly, while the segmentation decoder weights are initialized using residual network weights, with the skip connections contained within initialized randomly.

4.1.3 Optimizer and Regularization

The neural network training in this paper uses the Adam [17] optimizer with a learning rate of 1e-5. The dropout percentage is set to 0.5, and weight decay for all layers uses 5e-4.

4.2 Performance Evaluation

To investigate the impact of using deep residual networks on performance in autonomous driving road scene understanding tasks, this paper conducted ex-

periments on road segmentation, vehicle detection, and road classification tasks using different network structures on the same dataset. For road segmentation tasks, maximum F1 value (MaxF1) is used as the comparison metric; for vehicle detection tasks, average precision for moderately difficult objects is compared; and classification average precision (AP) serves as the evaluation standard for classification problems, as shown in Table 3 .

Table 3 shows the results of using VGG network structures and deep residual networks for feature extraction tasks. For road segmentation tasks, the segmentation accuracy using deep residual networks for feature extraction improved by 6.5% compared to VGG network structures. For vehicle detection tasks, the average precision also improved by 2.15%. Additionally, for traditional classification tasks, the average precision also saw a slight improvement. Experimental results further demonstrate that deep residual networks are more beneficial than VGG networks for improving task accuracy in autonomous driving road scene understanding tasks.

For vehicle detection tasks, this paper uses different network structures and layers for image feature extraction to complete vehicle detection tasks. As shown in Table 5 and Figure 6 [Figure 6: see original paper], object detection difficulty is divided into three categories: easy, moderate, and hard. The three different types of objects are trained and evaluated using VGG networks and deep residual networks with different layers. The comparison results show that while maintaining operation speed, the deep residual network structure used in this paper significantly improves vehicle recognition accuracy. Using the moderate difficulty as the final evaluation metric, the 152-layer deep residual network improves recognition accuracy by 4.9%. The author speculates that because the KITTI Object dataset contains multiple categories and rich data for pretraining, the multi-layer, large-scale neural network models are fully trained, thus significantly improving recognition accuracy.

The experiments compare the performance of ENet [3], FCN [2], SPL [15], and the proposed algorithm for road segmentation tasks. The ENet network uses an encoder-decoder architecture that backpropagates classification to the original image for semantic segmentation. The FCN network is the first typical network structure to achieve end-to-end semantic segmentation. Additionally, SPL introduces an unsupervised approach for label generation to complete road segmentation tasks. Table 6 compares the accuracy of different methods on the road segmentation task.

As can be seen from Table 6 and Figure 7 [Figure 7: see original paper], the method using deep residual encoder-decoder network structure for road segmentation achieves significantly better segmentation accuracy than methods that do not use deep residual encoder-decoder network structures (ENet, FCN, SPL), with the proposed method achieving the best segmentation accuracy. Compared with the traditional semantic segmentation method FCN, the proposed method' s accuracy improved by 5.16%. This is because while processing segmentation tasks, the encoder-decoder structure and deep residual network for feature ex-

traction deeply fuse high-dimensional abstract features with low-dimensional boundary texture features, improving the network model' s generalization capability and thus improving segmentation accuracy. Additionally, this paper only conducts training and evaluation on the KITTI road dataset without leveraging other larger-scale datasets, which saves significant training time and dataset resources compared to the SPL method that uses the KITTI Object dataset for training models.

For vehicle detection tasks, the proposed method is compared with different excellent vehicle detection methods from the KITTI Object leaderboard. The compared algorithms have hardware running environments basically consistent with the proposed method, so the detection results are compared. Table 7 compares the running speeds of different detection algorithms under similar accuracy conditions.

As can be seen from Table 7 and the corresponding Figure 8 [Figure 8: see original paper], when comparing the proposed method with detection algorithms with consistent or even superior hardware environments, the proposed algorithm shows significant speed improvement while ensuring no large gap in detection accuracy. Under the premise of guaranteeing high accuracy, the running time reaches 65 ms. Since this paper only performs parameter fine-tuning on KITTI data based on pre-trained deep residual network models, this provides significant improvement in shortening running time and increasing speed.

Figure 9 [Figure 9: see original paper] visualizes road segmentation task results in an intuitive manner. The shadow regions in the first row mark the algorithm' s output road segmentation areas; the second row shows the actual effective road area in the original images; the third row shows the road annotation areas displayed by the actual labels in the KITTI Road dataset. Figure 10 [Figure 10: see original paper] shows road classification and vehicle detection results. The first row shows original images of single and multiple lane line roads from the KITTI Road dataset; the top-left corner of images in the second row displays the road category, while vehicles are marked with bounding boxes indicating detected vehicle positions. The results demonstrate that the proposed method can effectively complete road segmentation, vehicle detection, and road classification tasks.

Based on the above experiments, this paper also uses the Cityscapes dataset [18] for training and testing on deep residual networks to extract road features and complete road segmentation tasks. The Cityscapes dataset provides an image segmentation dataset for autonomous driving environments, used to evaluate visual algorithms' capability in urban scene semantic understanding. It provides 5,000 finely annotated images from 50 cities across different scenes and seasons, making it a very complete dataset for autonomous driving environments. The test results on this dataset are shown in Figure 11 [Figure 11: see original paper]. As can be observed from Figure 11, the deep residual network demonstrates practical effects on road segmentation tasks across different road scene datasets. Unlike the KITTI dataset, the Cityscapes dataset annotates 30 different objects.

In this experiment, for the specific road segmentation task, only road features are learned while other redundant features are treated as background. The first row shows original data images from the Cityscapes dataset; the second row shows segmentation result images using the residual network in this paper. It can be clearly observed that the proposed method also has good generalization capability and practical value when tested on different datasets, and can be well migrated to other well-annotated datasets for autonomous driving scene testing.

5 Conclusion

This paper proposes an encoder-decoder network structure based on deep residual learning to address road scene understanding problems in autonomous driving. The method uses deep residual networks as encoders for high-dimensional image feature extraction tasks and shares the extracted high-dimensional features among parallel road segmentation, vehicle detection, and road classification tasks to improve operation speed and task accuracy. Experiments on the KITTI dataset demonstrate that the algorithm can effectively improve road segmentation operation speed while guaranteeing segmentation precision and can improve the accuracy of vehicle detection and road classification tasks to a certain extent. The algorithm improves vehicles' perception of road environments, thereby ensuring the stability, accuracy, and timeliness of autonomous driving technology, and has broad application prospects in the autonomous driving field.

References

[1] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]// Proc of International Conference on Neural Information Processing Systems. New York: Curran Associates Inc, 2012: 1097-1105.

[2] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3431-3440.

[3] Paszke A, Chaurasia A, Kim S, et al. Enet: a deep neural network architecture for real-time semantic segmentation [J]. arXiv: 1606.02147, 2016.

[4] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. arXiv: 1409.1556, 2014.

[5] Badrinarayanan V, Kendall A, Cipolla R. Segnet: a deep convolutional encoder-decoder architecture for image segmentation [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2017, 39 (12): 2481-2495.

[6] Teichmann M, Weber M, Zoellner M, et al. Multinet: real-time joint semantic reasoning for autonomous driving [EB/OL]. (2016). https://arxiv.org/pdf/1612.07695.pdf.

[7] Ren Shaoqing, He Kaiming, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [C]// Advances in Neural

Information Processing Systems. 2015: 91-99.

[8] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2016: 779-788.

[9] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]// Proc of International Conference on Neural Information Processing Systems. New York: Curran Associates Inc, 2012: 1097-1105.

[10] Szegedy C, Liu Wei, Jia Yangqing, et al. Going deeper with convolutions [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE Computer Society, 2015: 1-9.

[11] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[12] Geiger A. Are we ready for autonomous driving? The KITTI vision benchmark suite [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE Computer Society, 2012: 3354-3361.

[13] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks [C]// Proc of European Conference on Computer Vision. [S. l.]: Springer, 2014: 818-833.

[14] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions [J]. arXiv: 1511.07122, 2015.

[15] Wang Weiyue, Wang Naiyan, Wu Xiaomin, et al. Self-paced cross-modality transfer learning for efficient road segmentation [C]// Proc of IEEE International Conference on Robotics and Automation. 2017.

[16] Fritsch J, Kuhnl T, Geiger A. A new performance measure and evaluation benchmark for road detection algorithms [C]// Proc of International IEEE Conference on Intelligent Transportation Systems. [S. l.]: IEEE Press, 2014: 1693-1700.

[17] Kingma D, Ba J. Adam: a method for stochastic optimization [EB/OL]. (2014). https://arxiv.org/pdf/1412.6980.pdf.

[18] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE Computer Society, 2016.

Note: Figure translations are in progress. See original paper for figures.