End-to-End Deep Residual Network for Semantic Segmentation

Yingzi Zhou\textsuperscript{1,a}, Kun Huang\textsuperscript{2,b}, Xiaoying Guo\textsuperscript{3,c}
\textsuperscript{1,2,3}Panzhihua University, Panzhihua, China
\textsuperscript{a}zhouyz@pzhu.edu.cn
\textsuperscript{b}zyzjoyce77@163.com, \textsuperscript{c}39827453@qq.com, \textsuperscript{c}25971025@qq.com

Abstract: Recent work has made significant progress in improving the pixelwise labeling with convolutional neural networks by using residual networks. In this paper, we explore the impact of residual network in semantic segmentation by residual encoder-decoder model. The residual block can improves the dimension of feature maps, and obtains more efficient feature maps, and then ensures the effective of deep network structure. Moreover, we select the activation function with zero-centered characteristics to speed up the model convergence. The SELU activation function is used to remove the BN function in the network due to normalized deployment, thereby reducing the amount of network calculations. The proposed residual encoder-decoder model can still output high-precision prediction results without any post-processing, and the convergence speed is significantly faster. Our approach has achieved 61.13% mIoU, 86.92% PA, 73.42% MPA on Camvid dataset.

1. INTRODUCTION

Semantic segmentation focuses on predicting the semantic label of a single pixel, that using multi-dimensional features automatically extracted by the network to describe a single pixel, and assigning the category probability of each pixel in the image through a certain feature organization, and then predicting pixel by pixel. Long [1] uses the fully convolutional network (FCN) to conduct pixel-level semantic segmentation research. Since then, deep learning-based semantic segmentation has been rapidly developed. The FCN is implementing pixel-level input and output, and increases the flexibility of the input image size, but the upsampling method of FCN is not fine enough, and the output predicted label map is rough. Subsequently, Kendall [2] proposed an encoder-decoder network structure in terms of image semantic segmentation, which greatly improved semantic segmentation results, but its detail segmentation accuracy and efficiency still need to be improved in practical applications. Therefore, the improved model [3-7] came into being, and through data enhancement, multi-scale fusion \cite{3, 8}, post-processing (CRF \cite{9}, etc.), additional features (elevation information, vegetation Index) \cite{10} and other optimization methods to improve the model’s ability to distinguish objects.

Current semantic segmentation methods are usually optimized and improved based on convolutional neural networks such as VGG \cite{11}, GoogleNet\cite{12}, ResNet\cite{13} and other models. For example, the RefineNet proposed by Lin et al. \cite{5}, the encoder uses the ResNet-101 residual model, and the decoder fuses the high resolution features of the encoder output and the low resolution features of the RefineNet decoder model. The RefineNet has achieved higher accuracy in image segmentation, but the network parameters and calculations are larger.
Deep neural networks require a lot of time to process a large amount of data. The convergence speed of the model is particularly important. Different activation functions have different training costs. The activation function with zero-centered characteristics can speed up the convergence speed of the model. Therefore, in addition to the optimization in the network structure, there are also many optimizations in the activation function to make the network model perform better. For example, the newly optimized activation functions ELU, Maxout, swish, etc. Thus the proposed network can benefit from the mutual reinforcement between residual networks and SELU [14]. Our contributions can be summarized as follows:

- Using residual networks to increase the dimension of feature descriptions helps to obtain better feature expressions, ensuring the effectiveness of deep network structure training and improving model accuracy.
- In order to make the network training model converge faster, we select the activation function with zero-centered characteristics to speed up the model convergence. We use the SELU activation function to remove the BN function due to the normalized deployment in the network, thereby reducing the amount of network calculation. Experiments show that the model proposed in this paper converges faster.
- The encoder-decoder segmentation network model of this article can still output prediction results with higher accuracy without any post-processing.

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 introduces the proposed method. Section 4 provides experiments and analysis, followed by the conclusion in Section 5.

2. RELATED WORK

Deep learning techniques for the semantic segmentation task have gained attention in the research community during the recent years. In great measure, this is due to the emergence of new challenges and segmentation datasets, from PASCAL_VOC2012, Berkeley Motion Segmentation Datasets, to more accurate and dense labeled ones as SYNTHIA Dataset, Camvid Dataset.

Semantic segmentation: FCN have become indispensable part of semantic segmentation. It does not contain a fully connected layer of fully convolutional network, and can adapt to any size input image. Recent efforts have focused on encoder-decoder based models, that extract the remote information, and the output of the encoder is transmitted to the decoder which produces high-resolution segmentation prediction. SegNet [15], U-Net and RefineNet are examples of such models that use different mechanisms to pass information from encoder to decoder. Another method to obtain long-term context information is spatial pyramid pooling. ParseNet [16] adds global context features for spatial features, DeepLabv2 [4] uses atrus spatial pyramid pool (ASPP), and PSPNet [6] introduces spatial pyramid pool on multiple scales to solve the segmentation problem.

We employ encoder-decoder SegNet as our segmentation model. It is very efficient in terms of memory and time, and has a smaller amount of training parameters, and it can use stochastic gradient descent for end-to-end training. SegNet’s encoder uses max-pooling step of the corresponding encoder to perform non-linear upsampling, decoder upsamples its lower resolution input feature maps. The upsampled maps are sparse and then convolved with a trainable filter to generate dense feature maps.

One difference between our work and SegNet is that our model is able to increase the dimension of feature descriptions by residual networks. Another is that our model converging faster by using SELU.

Residual network: The core idea of the residual network is to express the difference between the input mapping and the optimal solution with a function, deepen the network by optimizing the output, and solve the gradient dispersion. Its core structure is shown as Figure 1.
The structure contains 3 weight layers, \( x \) is input data, \( \sigma \) is the nonlinear activation function ReLU, and the weights of the convolutional layers are \( W_1, W_2, W_3 \) from left to right. After 3 layers of convolution operation, the output result \( (F) x \) is:

\[
F(x) = W_3 \sigma(W_2 \sigma(W_1 x)) \tag{1}
\]

Through the identity and the third nonlinear function, the desired underlying mapping as \( H(x) \):

\[
H(x) = F(x, \{W_i\}) + x \tag{2}
\]

When \( H(x) \) and \( x \) can meet the identity condition, the deep model is equivalent to the shallow model, which can avoid the problem of network degradation, so the identity condition can be transformed into a learning residual function \( F(x) = 0 \).

The biggest feature of residual network is to use the form of residual to increase the depth of network, and then to improve the dimension of feature description. It is helpful to learn objects more vividly, which ensure the effectiveness of deep network structure training and greatly improves the model accuracy. Current application researches on deep neural networks are usually optimized and improved based on residual network such as \cite{17}.

**SELU:** In deep learning, adding batch processing at each layer can make the network converge faster, and improve the effect. Klambauer \cite{14} proposed Spiking Neural Network (SNN) with SELU can achieve automatic normalization. The sample distribution is automatically normalized to 0 mean and unit variance by SELU. We choose to use SELU instead of ReLU as the activation function of our model that can remove the BN function in the network due to normalized deployment, thereby reducing the amount of network calculations.

### 3. Model

We propose a model based on encoder-decoder architecture to improve the semantic segmentation results. On the one hand, we use the residual block to increase the dimension of feature descriptions that improve the accuracy. On the other hand, the SELU used in the block reduces the amount of network calculations and speeds up the model convergence. For full pixel segmentation, the output should have the same resolution as the input by encoder-decoder.

#### 3.1. Encoder

The input \( x \) of the encoder is an RGB image, and the output \( f = \{f_1, f_2, ..., f_k\} \) is a set of features at different resolutions. The architecture of the encoder is illustrated as the part (on the left) in Figure 2. Each encoder in the encoder network convolutes with the pooling layer to produce a set of feature maps. The residual block include 3 convolutional layers, the convolutional kernel are \( 1 \times 1, 3 \times 3, 1 \times 1 \) from left to right that optimizes the feature maps between two max-pooling layers. The activation function is SELU that instead of ReLU and
Figure 2. End-to-end residual network for semantic segmentation. A encoder with residual block by SELU sub-samples its input to produce feature maps and stores the pooling indices for decoder. A decoder up-samples its input using the pooling indices form its encoder to produce feature maps. The feature maps at the output of the final decoder being the input of a trainable soft-max classifier that can classifies each pixel independently.

3.2. Decoder

Figure 2 depicts the decoder architecture as the part (on the right) in Figure 2. Each encoder layer has a corresponding decoder layer, decoder network alternately uses upsampling and residual blocks, and finally uses a multi-class soft-max classifier for each pixel independently classification. The decoder in the decoder network uses the max-pooling indices stored in the corresponding encoder feature maps to up-sample its input features. This step produces sparse feature maps. Then residual block optimizes the feature maps between two up-pooling layers and uses the SELU automatic normalization. The high dimensional feature maps at the output of the final decoder being the input of a trainable soft-max classifier that can classifies each pixel independently. The output of the soft-max is a K channel which K is the number of classes. Segmentation prediction is the category with the largest probability for each pixel.

4. Experiments

We use the Camvid dataset to benchmark the performance of the model. All of them are real street view images taken by car camera. There are 701 images, consisting of 468 training and 233 testing RGB images at 360×480 resolution. The dataset is the first one for road scene and the shooting time is day and evening, with different light intensity and different image complexity. It contains 11 target categories, and the unmarked categories are classified as unlabeled categories. The target category can be divided into two categories according to the proportion of pixels. The first category is medium and large target objects, such as sky, building, lane, sidewalk, tree and car. The second category is small target objects, such as poles, fences, signal lights, pedestrians.

4.1. Training

The encoder and decoder weights were all initialized using the model pre-trained on ImageNet. We use random gradient descent with a step learning rate and momentum of 0.9 and mini-batch size of 4. Figure 3 shows qualitative results comparing the training accuracy and loss of SegNet and ours. It shows
that the model proposed in this paper can learn features faster and converge faster, and the speed of improving the accuracy rate in the training process is faster than SegNet.

4.2. Analysis

We use three commonly used performance measures. Pixel accuracy (PA) which is the ratio of the correct pixels set to the total pixels set. Mean pixel accuracy (MPA) which calculates the proportion of pixels predicted by policy in the total pixels of each category to get $PA_i$, and then calculates the average value of all categories to get the average pixel accuracy MPA. Mean intersection over union (MIoU) which is the ratio of intersection and union of test pixel set and label pixel set.

The result in Table 1 shows the proposed model in this paper obtaining competitive results when compared with other methods. This shows the ability of the residual block to extract meaningful features from the input images and feature maps to accurate class segment labels. Compared with SegNet, the global pixel accuracy of ours is 0.97% higher, while the mean pixel accuracy is 2.47% higher, and the mean intersection over union is 1.92% higher. This shows that the model with residual block can extract more efficient feature information then improve the result.

| method     | performance measures |
|------------|----------------------|
|            | PA   | MPA  | MIoU |
| SegNet     | 85.95 | 70.95 | 59.21 |
| ours       | **86.92** | **73.42** | **61.13** |

The result in Table 2 shows the pixel accuracy of each class. The third and fourth rows in the table are the results of SegNet method, and the fifth and sixth rows are the results of proposed method. Our model has obvious advantages in buildings, sky, roads, trees, signal lights, fences, bicyclists, and has a significant performance improvement. It is able to segment small and large classes both (see Fig.4). The first line is five test samples, which were taken in the daytime and evening respectively. The complexity of background is different, which is used to verify the robustness of the method. The second line is ground truth, and the third line is the test results of SegNet, the last line is our test results.

| Method | classes |
|--------|---------|
|        | sky | building | pole | road | Sidewalk | tree | Sign-symbol | fence | car | pedestrian | bicyclist |
|        | PA  | 93.83    | 85.79 | 39.99 | 94.65 | **88.48** | 81.73 | 50.87 | 40.18 | **79.88** | 69.38 | 33.78 |
|        | IoU | 90.77    | 77.70 | 27.15 | 91.48 | 77.15 | 71.11 | 31.11 | 30.01 | **74.04** | 49.00 | 31.83 |
| Ours   | PA  | **94.21**| 89.20 | 47.13 | **97.82**| 87.33 | **83.64**| **55.13**| **44.01**| 75.90 | **71.81**| **62.22**|
|        | IoU | **91.20**| 79.14 | 26.42 | 91.91 | 78.23 | 73.63 | 37.21 | 33.11 | 68.54 | **45.70**| **47.22**|
Figure 4. Results on Camvid test samples.

5. CONCLUSIONS
We presented an end-to-end residual network for semantic segmentation which is efficient both in terms of extracting feature maps and speeding up model training convergence. The encoder network with residual block improves the dimension of feature description and optimizes the feature maps between two max-pooling layers. It is helpful to learn objects more vividly, which ensure the effectiveness of deep network structure training and greatly improves the model accuracy. In addition, SELU instead of ReLU can remove the BN function in the network due to normalized deployment and reduce the amount of network calculations. And then the residual block with SELU improves the convergence speed of model training. End-to-end learning and fast convergence of deep segmentation architectures is a big challenge and we will continue to work hard to this important problem.

ACKNOWLEDGMENT
The authors would like to acknowledge support from NSFC (No.11176018) and research funds for the Science and technology benefiting people of Chengdu (2015-HM01-00293-SF).
1. NSFC (No.11176018)
2. Research funds for the Science and technology benefiting people of Chengdu (2015-HM01-00293-SF).

REFERENCES
[1] Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2015: 3431-3440.
[2] Kendall A, Badrinarayanan V, Cipolla R. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding[J]. Computerence, 2015.
[3] Audebert N, Saux B L, Lefèvre, Sébastien. Semantic Segmentation of Earth Observation Data Using Multimodal and Multi-scale Deep Networks[J]. 2016.
[4] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4):834-848.
[5] Lin G, Milan A, Shen C, et al. RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.

[6] Zhao H, Shi J, Qi X, et al. Pyramid Scene Parsing Network[C]// Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 6230-6239.

[7] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation[J]. 2015.

[8] Yu F, Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions[C]// International Conference on Learning Representations (ICLR). 2016.

[9] Zhao J, Zhong Y, Shu H, et al. High-Resolution Image Classification Integrating Spectral-Spatial-Location Cues by Conditional Random Fields[J]. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 2016, 25(9):4033-4045.

[10] Zhou X, Takayama R, Wang S, et al. Deep learning of the sectional appearances of 3D CT images for anatomical structure segmentation based on an FCN voting method[J]. Medical Physics, 2017, 44(10).

[11] Paisitkriangkrai S, Sherrah J, Janney P, et al. Effective semantic pixel labelling with convolutional networks and Conditional Random Fields[C]// Computer Vision & Pattern Recognition Workshops. IEEE, 2015.

[12] Szegedy C, Liu W, Jia Y, et al. Going Deeper with Convolutions[J]. 2014:1-9.

[13] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[J]. 2016:770-778.

[14] Klambauer, Günter, Unterthiner T, Mayr A, et al. Self-Normalizing Neural Networks[J]. 2017.

[15] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017:1-1.

[16] Wei Liu, Andrew Rabinovich, and Alexander Berg. Parsenet: Looking wider to see better. arXiv preprint.arXiv:1506.04579, 2015.

[17] Yongliang Tang, Jiashui Huang, Faen Zhang, et al. Deep residual networks with a fully connected deconstruction layer for single image super-resolution. 2020, 405:186-199.