Semantic Segmentation using Light Attention Mechanism

Yuki Hiramatsu and Kazuhiro Hotta
Meijo University, Japan

Keywords: Semantic Segmentation, Attention Mechanism, Encoder-decoder Structure.

Abstract: Semantic segmentation using convolutional neural networks (CNN) can be applied to various fields such as automatic driving. Semantic segmentation is pixel-wise class classification, and various methods using CNN have been proposed. We introduce a light attention mechanism to the encoder-decoder network. The network that introduced a light attention mechanism pays attention to features extracted during training, emphasizes the features judged to be effective for training and suppresses the features judged to be irrelevant for each pixel. As a result, training can be performed by focusing on only necessary features. We evaluated the proposed method using the CamVid dataset and obtained higher accuracy than conventional segmentation methods.

1 INTRODUCTION

Convolutional neural network (CNN) (Krizhevsky, 2012) has achieved very high accuracy on image recognition. Semantic segmentation using CNN can be applied to various fields such as automatic driving (Badrinarayanan, 2017) and medical images (Ronneberger, 2015). Semantic segmentation refers to pixel-wise class classification. Typical methods for semantic segmentation using CNN include Fully Convolutional Neural Networks (Long, 2015) and encoder-decoder networks (Ronneberger, 2015). These methods are the basic structure of the semantic segmentation method. Many of latest methods have extracted features using very deep CNN. In general, it is said that the deeper the hierarchy of CNN, the better the feature extraction function and the higher the accuracy. However, the deepening of the CNN increases the amount of computation and the number of parameters.

In order to deal with this problem, we propose a light attention mechanism that can be introduced into the basic encoder-decoder network. In the network where an attention mechanism that we proposed is introduced, it pay attention to the extracted features and emphasizes the features judged to be effective for training and suppresses the features judged to be irrelevant for training. As a result, it can perform training while focusing only on the necessary features. Therefore, it can be considered that the increase in computational complexity and the number of parameters can be mitigated. In the experiment, we evaluate the proposed method using the CamVid dataset (Brostow, 2009) labelled with 11 classes of images taken by the in-vehicle camera. As a result, the proposed method could obtain higher accuracy than conventional segmentation methods.

2 RELATED WORKS

This section describes related works. Section 2.1 describes the encoder-decoder structure. Section 2.2 describes the attention mechanism (Wang, 2017, Hu, 2018).

2.1 Encoder-decoder

U-Net (Ronneberger, 2015) is proposed as a segmentation method using CNN. This method adopts the encoder-decoder structure. The encoder extracts features using convolution and down sampling, and the decoder restores the resolution of feature maps step by step while extracting features. In addition, a skip connection is introduced at each resolution between encoder and decoder, and feature maps obtained by the encoder is connected to the corresponding feature maps at decoder with the same resolution. This restores the information lost during feature extraction.

Hiramatsu, Y. and Hotta, K.
Semantic Segmentation using Light Attention Mechanism.
DOI: 10.5220/0009347206220625
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 4: VISAPP*, pages 622-625
ISBN: 978-989-758-402-2; ISSN: 2184-4321

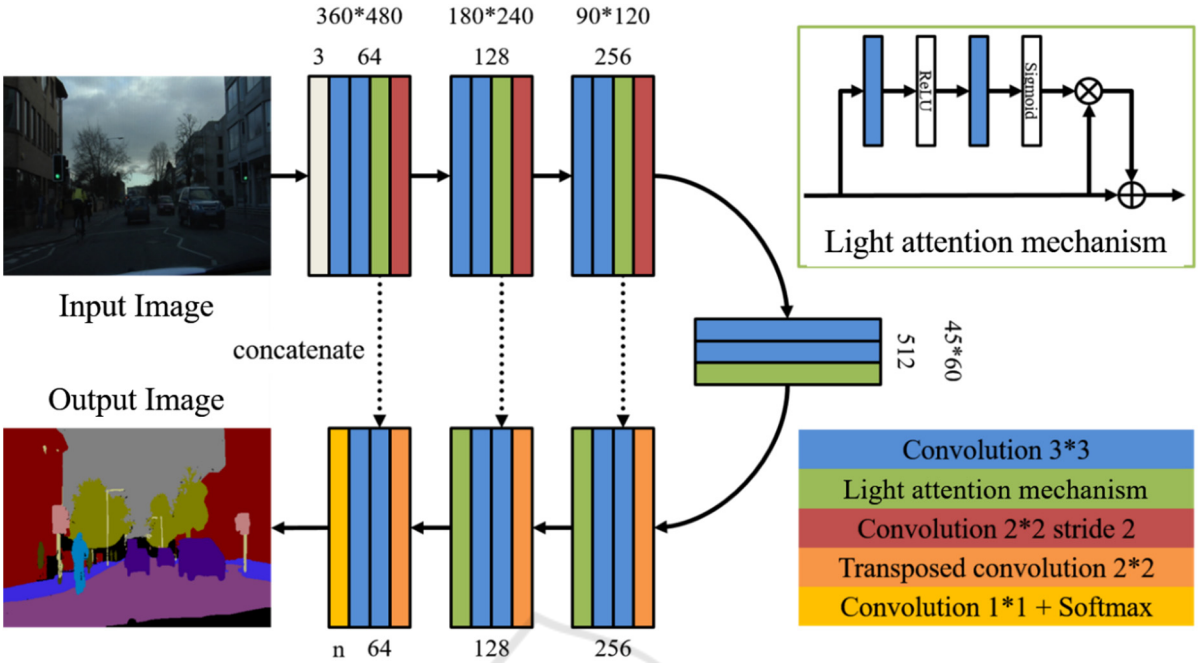

Figure 1: Overview of the proposed method.

2.2 Attention Mechanism

In the field of image recognition, Residual Attention Network (Wang, 2017) which introduced attention mechanism with the encoder-decoder structure into ResNet has been proposed. This method proposed attention residual learning which is similar to the residual block. This solved the problem of accuracy reduction due to the disappearance of the gradient due to deepening and the suppression of important features.

In the latest research, Squeeze-and-Excitation Networks (SENet) (Hu, 2018) has been proposed. This method proposed an attention mechanism (SE block) that adaptively weights feature maps for each channel. We confirmed that the training time was longer in the SENet. Therefore, we referred to the structure of an attention mechanism of the Residual Attention Network (Wang, 2017).

3 PROPOSED METHOD

This section describes the proposed method. Section 3.1 describes details of the network and Section 3.2 describes the proposed attention mechanism.

3.1 Network Details

Figure 1 shows the overview of the proposed method. Each encoder block consists of two 3×3 convolution layer (stride=1, pad=1), a 2×2 convolution layers (stride=2, pad=0) and an attention mechanism. On the other hand, each decoder block consists of a 2×2 transposed convolution layer (stride=2, pad=0), two 3×3 convolution layers (stride = 1, pad = 1) and an attention mechanism. The last layer is a 1×1 convolution layer (stride=1, pad=0) which compresses the number of dimensions of feature maps to the number of classes, and a Softmax function. “n” indicates the number of classes. In the convolution layer except for the final layer, Batch Renormalization [Ioffe, 2017] and ReLU function are used after convolution.

3.2 Light Attention Mechanism

An attention mechanism consists of two convolutional layers with different activation functions. In the first convolution layer, the Batch Renormalization and ReLU functions are used after the input is convolved. In the next convolution layer, Batch Renormalization and the sigmoid function are used after the convolution process. Finally, the output of an attention mechanism is calculated by multiplying the input and the output of the attention

Table 1: Accuracy comparison results for CamVid test dataset (Brostow, 2009).

Method	Building	Tree	Sky	Car	Sign	Road	Pedestrian	Fence	Pole	Pavement	Bicyclist	Mean IoU
SegNet	68.7	52.0	87.0	58.5	13.4	86.2	25.3	17.9	16.0	60.5	24.8	46.4
FCN8	77.8	71.0	88.7	76.1	32.7	91.2	41.7	24.4	19.9	72.7	31.0	57.0
FC-DenseNet67	80.2	75.4	**93.0**	78.2	40.9	**94.7**	58.4	30.7	38.4	81.9	52.1	65.8
FC-DenseNet103	**83.0**	**77.3**	**93.0**	77.3	43.9	94.5	**59.6**	**37.1**	37.8	**82.2**	50.5	66.9
Ours(Without Attention)	80.9	73.5	92.2	75.0	43.6	92.2	54.8	36.5	38.8	77.9	50.0	65.0
Ours(With SEBlock)	79.9	71.8	91.9	74.0	41.7	92.2	54.1	28.6	36.5	77.3	49.5	63.4
Ours(With Attention)	81.7	74.7	91.9	**79.4**	**48.7**	94.0	57.0	35.2	**40.3**	81.8	**55.0**	**67.3**

mechanism calculated in the following as

$$H(x) = \big(1 + F(x)\big) * x \tag{1}$$

where x is the input, F (x) is the output of an attention mechanism. Features judged to be necessary for training is emphasized and features judged to be irrelevant is suppressed by this process. Therefore, training can proceed while focusing on only the necessary features and efficient learning can be performed. The proposed an attention mechanism is different from SE block and has a normal convolution layer, so attention can be paid in pixel units. This is the difference from SE block.

4 EXPERIMENTS

This section describes the experimental results of the proposed method. Section 4.1 describes the dataset used in experiments. Section 4.2 describes the details of training of the proposed method. Section 4.3 describes evaluation experiment results.

4.1 Dataset

We evaluate the proposed method using the CamVid dataset (Brostow, 2009). The size of images is 360×480 pixels and the number of classes is 11 classes. It consists of 367 training, 101 validation and 233 test images. As data augmentation, we used random left-right flipping during training.

4.2 Training

Since the classes are unbalanced in images, we use class balancing weight (Badrinarayanan, 2017) where the weight is assigned to each class in the cross-entropy loss function. The weight is defined as

$$w_c = \text{median}_{\text{frequency}} / \text{frequency}(c) \tag{2}$$

where median frequency is the median of all class frequencies and frequency (c) is the number of pixels of class c in training images. This gives a large weight to classes with a small ratio and a small weight to classes with a large ratio.

We set the batch size to 2 and used an Adam optimizer. The initial learning rate was 1e-3 and trained up to 1000 Epoch.

4.3 Results

We evaluated the accuracy of the proposed method with test images. Table 1 shows the comparison results for the test images and Fig. 2 shows the example of the segmentation results. Table 1 shows that the proposed method improved the Mean IoU accuracy compared to the conventional segmentation method. Compared with FC-DenseNet103 (Simon, 2017), IoU accuracy for each class can be confirmed to improve IoU accuracy for Sign class by about 4.8% and IoU accuracy for Bicyclist by about 4.5%. From these results, it can be confirmed that the proposed method can improve the classification accuracy for the class with small size. On the other hand, Figure 3 shows that bus was misclasfified as building class. This is due to the fact that objects such as buses appear very rarely in the dataset and learning is insufficient.

5 CONCLUSIONS

We proposed an attention mechanism that can be introduced in the encoder-decoder network. This achieved higher accuracy than conventional segmentation methods on the CamVid dataset. However, incorrect segmentation may occur for

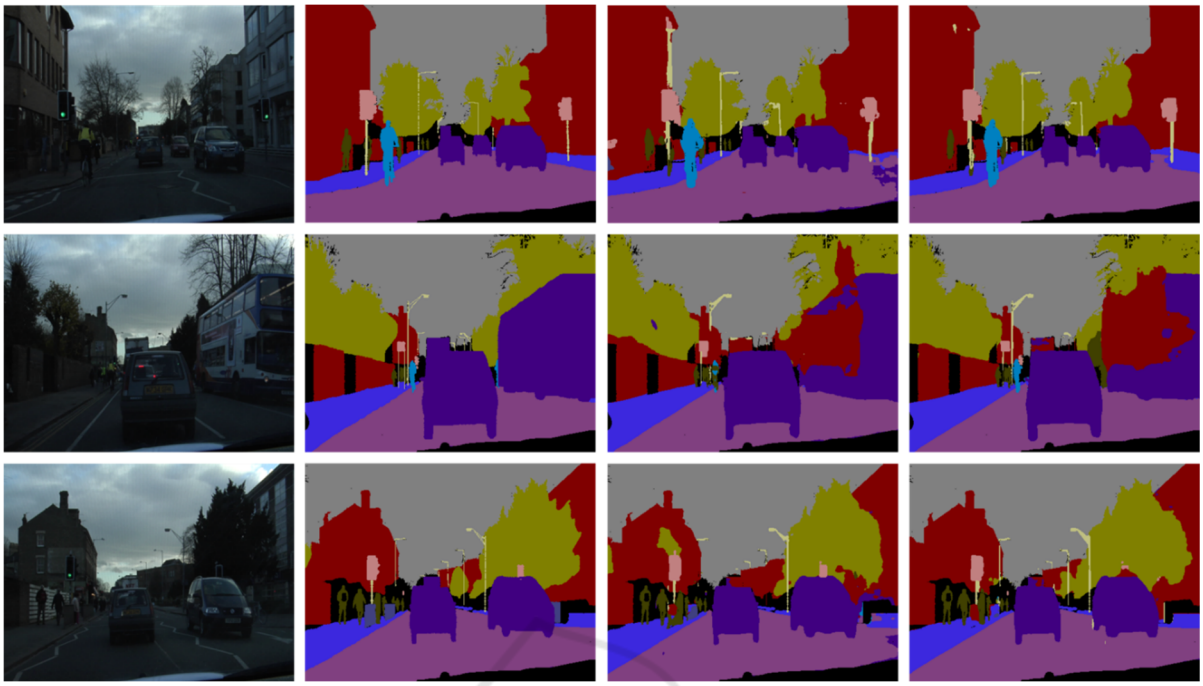

Figure 2: Example of segmentation results for CamVid test dataset (Brostow, 2009) (Left: Input Image, middle-left: Ground truth, middle-right: Conventional method, right: Our method).

objects that do not appear frequently. In the future, we are planning to validate an attention mechanism against other encoder-decoder networks and improve the feature extraction ability by optimising the structure of an attention mechanism.

REFERENCES

A. Krizhevsky, I. Sutskever, G. E. Hinton, “ImageNet classification with deep Convolutional neural networks”, In Advances in neural information processing systems, pp.1097-1105, 2012.

V. Badrinarayanan, A. Kendall, R. Cipolla, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.39, pp.2481-2495, 2017.

O. Ronneberger, P. Fischer, T. Brox. “U-Net: Convolutional networks for biomedical image segmentation”, Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241, 2015.

J. Long, E. Shelhamer, and T. Darrell. "Fully Convolutional Networks for Semantic Segmentation”, Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440, 2015.

F. Wang, et al. “Residual attention network for image classification”, Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 6450-6458, 2017.

J. Hu, L. Shen, and G. Sun. “Squeeze-and-excitation networks”, Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

J. Simon, et al. "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation", Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1175-1183, 2017.

G. Brostow, J. Fauqueur, and R. Cipolla, “Semantic object classes in video: A high-definition ground truth database,” Pattern Recognition Letters, Vol. 30, No.2, pp. 88-97, 2009.

S. loffe, “Batch Renormalization: Towards Reducing Minibatch Dependence in Batch-Normalized Models”, Advances in neural information processing systems, 2017.