An Improved Fully Convolutional Network for Semantic Segmentation

Gaihua Wang 1, 2, a, Xizhou Wan 2, b, *, Xu Zheng 2, c and Zhao Guo 2, d

1 Hubei Collaborative Innovation Centre for High-efficiency Utilization of Solar Energy, Hubei University of Technology, Wuhan 430068, China.
2 School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan 430068, China.

a 20130006@hbut.edu.cn, b, * Corresponding author e-mail: xizhou.wan@foxmail.com, c zhengxuliong@163.com, d 823679338@qq.com

Abstract. A novel loss function based on category distribution for semantic segmentation is proposed in this paper. The new loss function is composed of two parts: (1) Pixel-wise cross-entropy. (2) Category distribution loss proposed. Most of existing semantic segmentation networks adopt a single pixel-wise cross-entropy loss function, which guides the network to independently predict the class each pixel belongs to. The downside is that ignoring global information of the image -- the distribution of all kinds of objects in the image, resulting in an unsatisfactory segmentation result. Category distribution loss proposed in this paper obtains category distribution information of the whole image by calculating the percentage of pixels to each class of objects. Acquired category distribution information can be represented as a feature vector, the distance of two vectors between prediction and ground truth forms category distribution loss. The new loss function measures the difference between pixels as well as the difference between global information. We apply the new loss to several classical networks, its pixel accuracy and mIoU accuracy on two benchmark datasets CamVid and Pascal VOC2012 are improved compared to using only pixel-wise cross-entropy. In addition, we also introduced the attention mechanism, which proved to be able to improve segmentation accuracy of the networks as well.

Keywords: Convolutional neural network, Semantic segmentation, Loss function.

1. Introduction

Deep Convolutional Neural Networks (DCNNs) [1] have achieved outstanding performance in many computer visions tasks such as image classification [2] and object detection [3], which is attributed to the excellent representation power of convolutional neural networks as feature extractor. This nice property of convolutional neural networks also applies to another important task in computer vision: Semantic segmentation, which aims at the classification and recognition of pixels so as to obtain the segmentation image with semantic information. Specifically, it will enable the computer to recognize the image content, thus achieve understanding and cognition of the scene.
The research on semantic segmentation can be roughly divided into two directions. One is to design different topology or new module for the network structure. Secondly, new losses and evaluation criterias are proposed to optimize network training.

Some representative work has been done in network structure innovation. For make full use of CNN in semantic segmentation task, long et al. [5] adopt convolution layers replace dense layers in classical network VGG [4], called Fully Convolutional Network (FCN). Ushering in a new era of end-to-end models, FCN is a substantial improvement over the previous methods based on regional feature extraction, which further proves that DCNNs have well ability of object recognizing. Subsequently, Badrinarayanan et al. [6] present an encoder-decoder architecture (SegNet). Specifically, the decoder uses pooling indices computed in the max-pooling step of the corresponding encoder to perform nonlinear upsampling. Ronneberger et al. [7] proposed a symmetrical "U-shaped" network structure, and concatenate the corresponding feature map of the encoder in the decoder stage to make full use of low-stage features for information fusion. Fisher Yu et al. [8] proposed dilated convolution to systematically aggregate multi-scale context information without losing resolution. It is widely used to reduce the number of downsampling to avoid large loss of resolution and to expand the receptive field. [9] believed that the features of each layer of the convolutional layer were helpful to the final result, so it designed a refinenet module, which could combine the high-stage features with the low-stage features. The feature map output by the final model contained all the information of the convolution layer, so it obtained high accuracy. To improve the network computing speed and make it more suitable for real-time tasks, Nekrasov et al [10] replaced the 3×3 convolution in Refinenet with 1×1 convolution. DeepLab family [11,12,13,14] apply atrous convolution to extract features at multiple scales in the form of pyramid, and CRF is used as a post-processing method to further improve the network segmentation result.

Another kind of research focus on improving loss function, and also makes remarkable achievements. In the image classification task, in order to enhance the intra-class features, some researchers proposed contrastive loss [15, 16], triplet loss [17] and center loss [18]. These loss functions usually measure the difference between the predicted results and the label on a batch image. Contrastive loss and triplet loss require input image pairs or triples in each training iteration, which leads to a sharp increase in training samples and thus greatly increases the computational complexity. Center loss solves the above problem by introducing k-nearest neighbor algorithm (K-NN) [19] into softmax cross-entropy. During each training iteration, center loss calculates the distance between depth features and each class center of features through a small batch of images, and updates these centers after each iteration. Center loss can effectively reduce the intra-class variance while maintaining the separability of features of different classes. However, the computational cost of such methods is still high, not to mention that for image segmentation, each pixel is regarded as a training sample. In addition, the class distribution of most semantic segmentation data is unbalanced, so that the network training is biased to the main class [20]. In order to solve the problem of class imbalance, Lin et al. [21] revised the standard cross entropy loss, reduced the loss weight assigned by well-classified samples, and proposed focal loss.

Focus on improving loss function, we propose a new loss function based on the above research. New loss is constructed by measuring the distance of category distribution information between label and network prediction. Specifically, the proportion of pixels per class in the ground truth is computed as the category distribution information of the image, which is an one-dimensional distribution vector, and the network prediction results can obtain a distribution vector in the same way. Then distance between the two vectors constitutes a new loss. New loss function will be combined with the original pixel-wise cross-entropy to form a total loss function to guide the network training. The new loss is no longer to independently predict the category of a single pixel, but also to consider the category of objects in the whole image and the proportion of different categories in the image. The loss function was used to train on FCN [5], U-Net[7] and PSPNet[22], and the experimental results were improved in pixel accuracy and mIoU accuracy on two benchmark datasets Pascal VOC2012[23], CamVid[24]. In addition, we also introduce the residual attention module in [28] to further improve the network segmentation results.
2. Related Work

2.1. Penalty Cross-entropy
Class imbalance is a common problem in datasets of object detection and segmentation. For this problem, focal loss reduces the weight of the easily classified samples and makes the model focus more on the samples that are difficult to be classified. Moreover, the relationship between pixels is ignored when only a single cross-entropy loss is adopted. Some research [25,26] no longer only consider the loss of a single pixel, but construct region-level loss functions. By means of adaptive method, pixel regions belonging to the same category are found, and more weight is given to pixels have the same label.

2.2. Attention Module
Residual attention module [28] assists the network to enhance effective information and suppress invalid information, making the network pay more attention to objects rather than backgrounds. The output of residual attention unit can be expressed as

\[ H_{i,c}(x) = (1 + M_{i,c}(x)) \ast F_{i,c}(x) \]  

(1)

where \( F_{i,c}(x) \) is the feature map output by conv layers, \( M_{i,c}(x) \) is attention weight. In [28], three types of attention are proposed and the results show that mixed attention (channel attention and spatial attention) has the best effect, its expression is

\[ M_{i,c}(x) = \frac{1}{1 + \exp(-x_{i,c})} \]  

(2)

3. Method

In this section, we first detailedly introduce the specific process of constructing category loss function proposed. Then, we elaborate how to add the residual attention unit into several classical networks.

3.1. Category Loss Function

Cross-entropy loss is the most commonly used loss function in CNN, its formula is given as follow

\[ L_{\text{cross-entropy}}^i = -y_{i,\text{true}}^i \log y_{i,\text{pred}}^i \]  

(3)

Where \( y_{i,\text{pred}}^i \) is prediction of the network on pixel \( i \), that is the class pixel \( i \) belongs to. \( y_{i,\text{true}}^i \) is the category to which pixel \( i \) actually belongs. This loss means that the category of each pixel will be
predicted independently, which not only separate the relationship between pixels in the same class, but also ignores the distribution of objects of each category in the whole image. For the latter problem, we propose a new loss function

$$L_{\text{class}} = \frac{1}{n} \sum_{i=1}^{n} (C_{\text{true}} - C_{\text{pred}})^2$$

(4)

Where $n$ is batchsize, $C_{\text{true}}$, $C_{\text{pred}} \in \mathbb{R}^{c\times W\times H}$ are the category distribution information of ground truth and prediction respectively, which are two distribution vectors. $c$ is the number of class. The specific process of constructing the category distribution loss function is shown in figure 1, the calculation procedure of $C_{\text{true}}$ and $C_{\text{pred}}$ is as follows

$$C_{\text{true}}^i = \frac{1}{W \times H} \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} Y_i(x,y), i = 0,1,\ldots,c-1$$

(5)

$$C_{\text{pred}}^i = \frac{1}{W \times H} \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} X_i(x,y), i = 0,1,\ldots,c-1$$

(6)

$X,Y \in \mathbb{R}^{W\times H\times c}$ are predicted results and labels respectively, $Y_i(x,y)$ is the pixel at $(x,y)$ in label, $X_i(x,y)$ is the pixel at $(x,y)$ in prediction. $Y$ is ground truth label after one-hot encoded, there are $c$ binary images, each one represents a class. If the image contains a class of objects, the pixel value belonging to this class in the binary image corresponding to this class is 1, and the pixel value not belonging to this class is 0. For this, we can easily get the distribution of various objects in the image by taking the global average of every binary image, then get a one-dimensional distribution vector $C_{\text{true}}$ of length $c$. Each value in the vector represents the proportion of each category of pixels in the whole image and reflects the distribution information of each category of pixels in the whole image. Similarly, $X_i$ is the predicted result of the network. That is, one image is input and $c$ images are output by the network, each image represents the segmentation result of each class. The distribution vector $C_{\text{pred}}$ can be obtained by global average pooling. Then the distance between two vectors constitutes a new loss called category loss function $L_{\text{class}}$. Finally, the new loss function will be combined with the original pixel-wise cross-entropy loss to form the total loss function:

$$L = (1 - \lambda)L_{\text{cross-entropy}} + \lambda L_{\text{class}}$$

(7)

Where $\lambda \in [0,1]$, equation 3-5 will become standard cross-entropy loss function when $\lambda = 0$. We set $\lambda$ to 0.5 in the experiments and get a better result.

The new loss is no longer to independently predict the category of a single pixel, but also to consider the category of objects in the whole image and the proportion of different categories in the image.
3.2. Residual Attention

Information fusion is widely used and low-stage feature fusion in decoder stage has been proved to be effective. Currently, there are mainly two feature fusion methods, add and concate. Most networks use these two methods to directly integrate low stage features and high stage features. The problem with this is that in the fusion process, without measuring the usefulness of each feature vector, the feature map is fused together, mixing a lot useless features with useful features. And attention mechanism assists the network to enhance effective information and suppress invalid information, making the network pay more attention to objects rather than backgrounds. Therefore, the introduction of attention module is feasible.

In this work, residual attention module (RA) proposed by Wang et al. [28] is introduced in FCN [5], U-Net [7] and PSPNet[22]. Figure 2 shows a schematic of adding residual attention module to encoder-decoder networks. In the feature fusion stage, low-stage features are first processed by residual attention module (RA) and then fused with high-stage features. Take the example of U-Net[7] to show more specifically how do we add attention blocks, its structure is shown in figure 3. In our networks, low-stage feature maps concatenate with high-stage feature maps after residual attention, the same as PSPNet[22]. Note that feature fusion in FCN[5] is conducted by “add”, we trained FCN8s with the same option for information fusion.
Figure 3. Residual attention module in UNet

4. Experiments
A series of ablation experiments were conducted on two benchmark datasets CamVid[24], Pascal VOC2012[23] to verify the proposed loss and residual attention module. Pixel accuracy and mean intersection-over-union (mIoU) were used as evaluation criteria.

4.1. CamVid
The CamVid is a dataset consisting of 11 class objects, which contains 367 training images, 101 validate images and 233 test images. These images are uniformly cropped to 320×320 size in our experiments. For training, we use the Adam optimizer and initial learning rate is 0.001, batchsize is 8, 500 epoches for all networks. Test results on three classic networks are shown in table 1, some of the visualizations are shown in figure 4.

| Method   | \(L_{\text{class}}\) | RA | mean IoU(%) | Pixel Acc(%) |
|----------|----------------------|----|-------------|--------------|
| FCN      |                      |    | 65.42       | 73.65        |
| FCN      | \(\sqrt{\phantom{0}}\) |    | 69.63       | 76.51        |
| FCN      | \(\sqrt{\phantom{0}}\) | \(\sqrt{\phantom{0}}\) | 67.98 | 75.18 |
| FCN      | \(\sqrt{\phantom{0}}\) | \(\sqrt{\phantom{0}}\) | 68.33 | 76.06 |
| U-Net    | \(\sqrt{\phantom{0}}\) |    | 79.22       | 84.24        |
| U-Net    | \(\sqrt{\phantom{0}}\) | \(\sqrt{\phantom{0}}\) | 82.57 | 86.35 |
| U-Net    | \(\sqrt{\phantom{0}}\) | \(\sqrt{\phantom{0}}\) | 80.24 | 84.93 |
| U-Net    | \(\sqrt{\phantom{0}}\) | \(\sqrt{\phantom{0}}\) | 81.21 | 85.13 |
| PSPNet   | \(\sqrt{\phantom{0}}\) |    | 74.83       | 82.15        |
| PSPNet   | \(\sqrt{\phantom{0}}\) | \(\sqrt{\phantom{0}}\) | 78.56 | 84.55 |
| PSPNet   | \(\sqrt{\phantom{0}}\) | \(\sqrt{\phantom{0}}\) | 75.73 | 82.51 |
| PSPNet   | \(\sqrt{\phantom{0}}\) | \(\sqrt{\phantom{0}}\) | 76.56 | 83.14 |

Table 1. Test set results of CamVid on different networks
As shown in table 1, compared with the original networks, after using the new loss function, the Pixel Acc and mean IoU of three networks are greatly improved, respectively, by about 2.5% and 4%. The residual attention module also shows its effect, with a small boost compared to the original network. However, we found that the effect of using both the new loss function and residual attention module is not as good as using the new loss function alone, but it is better than using the attention module alone.

Figure 4. Some of the visualizations (left is original image, middle is ground truth, right is prediction)

4.2. Pascal VOC2012
The Pascal VOC2012 consists of 20 object categories and a background category. There are a total of 2913 images for semantic segmentation, including 1464 training sets and 1449 validation sets. Like CamVid, we cropped the images uniformly to 320×320, and use Adam optimizer with initial rate 0.001, batch size 4. Test results on three classic networks are shown in table 2.

| Method   | $L_{\text{class}}$ | RA | mean IoU(%)  | Pixel Acc(%) |
|----------|--------------------|----|--------------|--------------|
| FCN      | √                  |    | 64.23        | 75.52        |
| FCN      |                    | √  | 66.38        | 80.00        |
| FCN      |                    | √  | 65.21        | 78.31        |
| FCN      |                    | √  | 67.85        | 82.54        |
| U-Net    | √                  |    | 65.42        | 75.44        |
| U-Net    |                    | √  | 66.86        | 81.20        |
| U-Net    |                    | √  | 65.95        | 80.03        |
| U-Net    |                    | √  | 68.54        | 83.66        |
| PSPNet   | √                  |    | 62.51        | 74.36        |
| PSPNet   |                    | √  | 67.55        | 82.23        |
| PSPNet   |                    | √  | 66.19        | 81.11        |
| PSPNet   |                    | √  | 70.95        | 85.64        |
Similarly, the loss function proposed has also made progress on the dataset VOC. It can be seen from Table 2 that the loss function and attention module proposed in this paper both play a role. At the same time, under their joint action, the Pixel Acc and mean IoU of the three networks are the highest, which is relatively higher than the original network. In the dataset CamVid, the network achieves the best results when only the new loss function is used, while the dataset VOC achieves the best results when the loss function and attention module are used at the same time, which may be caused by the difference between the datasets, VOC contains more types of targets than CamVid, but the scene is simpler and the target objects are more prominent, so the attention mechanism can play a better effect in VOC.

5. Conclusion
In this work, we introduced a novel loss function based on category distribution, which combines with the weighted original pixel-wise cross-entropy to form the total loss. The category information contained in Ground Truth is further utilized to guide the network to train towards the object distribution situation closer to the real image, so as to improve the accuracy of network segmentation. After a series of experiments, the validity of the new loss proposed in this work was verified, and the segmentation result can be further improved by adding the attention mechanism.

References
[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” in Proc. IEEE, 1998.
[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in NIPS, 2013.
[3] Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.
[4] Ciresan, D.C., Gambardella, L.M., Giusti, A., Schmidhuber, J.: Deep neural net-works segment neuronal membranes in electron microscopy images. In: NIPS. pp.2852–2860 (2012)
[5] Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation.arXiv:1411.4038, 2014.
[6] Vijay Badrinarayanan, Ankur Handa, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. CoRR, abs/1505.07293, 2015.
[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In MICCAI, 2015.
[8] Yu F., Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions[J]. 2015.
[9] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian D. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In CVPR, 2017.
[10] V. Nekrasov, C. Shen, and I. Reid, “Light-weight refinenet for realtime semantic segmentation,” in BMVC, 2018.
[11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. CoRR, abs/1412.7062, 2014.
[12] Chen, L.C., Papandreou, G., Schro , F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587 (2017)
[13] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. TPAMI (2017)
[14] Chen L C , Zhu Y , Papandreou G , et al. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation[J]. 2018.
[15] Raia Hadsell, Sumit Chopra, and Yann Lecun. Dimensionality reduction by learning an invariant mapping. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 1735–1742, 2006.
[16] Yuheng Chen, Yuheng Chen, Xiaogang Wang, and Xiaou Tang. Deep learning face representation by joint identification-verification. In International Conference on Neural Information Processing Systems, pages 1988–1996, 2014.

[17] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In IEEE Conference on Computer Vision and Pattern Recognition, pages 815–823, 2015.

[18] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A Discriminative Feature Learning Approach for Deep Face Recognition. Springer International Publishing, 2016.

[19] K Fukunage and P. M Narendra. A branch and bound algorithm for computing k-nearest neighbors. IEEE Transactions on Computers, C-24(7):750–753, 1975.

[20] Samuel Rota Bulo, Gerhard Neuhold, and Peter Kontschieder. Loss max-pooling for semantic image segmentation. 2017.

[21] Tsung Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. pages 2999–3007, 2017.

[22] Zhao H, Shi J, Qi X, et al. Pyramid Scene Parsing Network[J]. 2016.

[23] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The pascal visual object classes VOC challenge. IJCV, 2010.

[24] Brostow, Gabriel J., et al. "Segmentation and Recognition Using Structure from Motion Point Clouds." European Conference on Computer Vision Springer, Berlin, Heidelberg, 2008.

[25] Ke, Tsung Wei, et al. "Adaptive Affinity Field for Semantic Segmentation." 2018.

[26] Guo J, Ren P, Gu A, et al. Locally Adaptive Learning Loss for Semantic Image Segmentation[J]. 2018.

[27] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In Neural Information Processing Systems, 2014.

[28] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In IEEE Conference on Computer Vision and Pattern Recognition, 2017.