Adversarial Erasing method based on graph neural network

Shuhan Liu*
University of Technology, Tianjin, China

*Corresponding author e-mail: liushuhann@icloud.com

Abstract. Semantic segmentation is a traditional task that requires a large number of pixel-level ground truth label data sets, which is time-consuming and expensive. Recent developments in weakly-supervised settings have shown that reasonable performance can be obtained using only image-level labels. Classification is often used as an agent task to train deep neural networks and extract attention maps from them. The classification task only needs less supervision information to obtain the most discriminative part of the object. For this purpose, we propose a new end-to-end counter-wipe network. Compared with the baseline network, we propose a method to apply the graph neural network to obtain the first CAM. It is proposed to train the joint loss function to avoid the network weight sharing and cause the network to fall into a saddle point. Our experiments on the Pascal VOC2012 dataset show that 64.9% segmentation performance is obtained, which is an improvement of 2.1% compared to our baseline.

1. Introduction
At present, image semantic segmentation is a kind of visual recognition task that has received much attention in the field of computer vision. The computer learns image features to simulate the cognitive process of human beings on the image, and then classifies each pixel in the image semantically. Semantic segmentation has a very wide range of applications in the fields of intelligent driving, satellite remote sensing mapping, security monitoring, face recognition, and medical image analysis. However, the fully-supervised semantic segmentation method needs to use a large number of pixel-level annotations to train the network, and its creation process requires a lot of time and manual labor. Therefore, in recent years, some researchers have devoted themselves to weakly supervised semantic segmentation, using more easily available image-level annotations to complete semantic segmentation tasks [1], such as image-level tags [2,3], bounding box [4], graffiti [5] and point annotation [3] have all made good progress. We also focus on the use of image-level labels, which are the weakest of all supervision. The training graph neural network obtains the CAM activation map of a set of pictures, uses the saliency erasure method to remove the saliency area, and then sends it to the CNN network to achieve a more complete semantic segmentation result.

The previous method suggested to alleviate the problem of poor CAM area expansion by introducing adversarial erasure. Adversarial erasure sets a threshold on the CAM to obtain the activation map of the most discriminative part of the object. Generate a mask that can be used to remove the most discriminative object area from the image. Send the obtained image to another brand-new classification network to find the secondary discriminant area belonging to the same object. The current erasure method is divided into multiple steps to perform erasure, or implemented as a multi-stage training
method, or combined with multiple erasure networks for training in an integrated manner. This leads to a complicated training process and an excessively long training time. In order to ensure the antagonism of the two networks, we do not share the parameters of the two networks. Compared with the previous method, our setting avoids the network sharing a set of parameters, which causes the discrimination area to fall into a certain part.

The main contributions of this paper are as follows: (1) We propose an end-to-end anti-erasing network framework and obtain good segmentation performance, (2) We propose to integrate graph neural networks into the network framework, making the network more effective for a group of similar images have a better segmentation effect. (3) We have proved its effectiveness on the Pascal VOC 2012 data set, which is better than the baseline.

![Figure 1. Difference from baseline method.](image-url)

2. Related work

Attention map. Early work was achieved by visualizing the partial derivatives of predicted class scores. After a lot of work has been done to improve the CAM, CAMs can be classified by adjusting the global average pooling and fully connected layer, and display related areas. In semantic segmentation, the classification task is the pre-task to obtain the attention map, and the attention map method visualizes the classification results. CAM only shows the most discriminative part of the object, but the semantic segmentation task needs to capture the entire range of the target. In order to alleviate this problem, [6] first introduced the method of adversarial erasure. In this method, the most discriminative area found by CAM is first erased from the image.

Graph neural network. Graph neural network has advantages in processing data with similar graph structure and has attracted wide attention. Li [7] et al. applied graph neural network in the direction of weakly supervised semantic segmentation, and described the task as a grouping task of images with similar content. A group of images can merge semantic knowledge with each other. Inspired by this method, we also introduce this graph neural network into the anti-erasing network, so that it can be trained against the ordinary CNN network.

Adversarial Erasing. Recently, adversarial training ideas have received widespread attention. In the field of weak supervision, adversarial training has also been widely used. This means that deleting part of the image and training an auxiliary model on this new image does not introduce adversarial target formulas or independent models with different parameters.

3. Method

In this section, we introduce the graph neural network and anti-erasing technology used in our proposed network framework. Input a group of similar pictures into the graph neural network to get the CAM and classification results, and fit the CAM back to the original image to erase the high discriminative area. The erased image is sent to the new CNN network to calculate the CAM again, and the secondary discriminant area of the same entity is obtained. The two CAMs are superimposed to obtain the final segmentation mask.
3.1. Graph neural network

According to Li et al. they applied the graph neural network in the direction of weakly supervised semantic segmentation. Inspired by this, we use this graph neural network as the network to get the first CAM. GNNs usually model graph-like elements and approximate reasoning as a learnable neural network, and perform iterative reasoning to clearly discover the relationship between nodes. In semantic segmentation, a set of pictures with the same class is used to form a complete picture, and similar semantic information potentially existing in the data set is used. So that similar images can absorb semantic knowledge from each other. Obtain more accurate CAM images. The graph network performs step-by-step reasoning, and optimizes details through node creation, message passing, and message aggregation iterations. The entire network trains a cross-entropy loss function.

Specifically, we input a group of similar images into the graph neural network, and obtain the feature map \( h \) through the last layer of convolution of the classification network. Define a complete graph \( G(V, E) \) and use this group of feature graphs as the graph nodes of the graph neural network. The edges can be expressed as a set of \( E = \{e_{i,j} = (v_i, v_j)\} \). The edges of a fully connected graph can be represented by an \( n \times n \) matrix. The matrix is learnable and is also the learning parameter of the graph neural network. During inference, the different coefficients of the matrix absorb similar semantic information from each other. GNN iteratively improves the representation of the features at the nodes by aggregating its neighborhood features, so that each edge represents the similarity measurement matrix between the two nodes where it is located. The formula is expressed as:

\[
{e'_{ij}} = h_i^T W h_j^{tr} \in \mathbb{R}^{WH \times HW}
\]

Among them, \( h_i \in \mathbb{R}^{WH \times C} \) represents the class score vector obtained by the i-th image through the classification network. Here, in order to facilitate the calculation, the dimension of \( R \) is flattened from three dimensions to two dimensions. It is the learnable matrix of the network. \( t \) is the number of iterations. Calculate the softmax value for each row of the obtained edge matrix, and after iteration \( t \) times, take the average value of all the feature maps of each node and output it to calculate the CAM. The message function passed is as follows:
\[ m'_{i,j} = \text{Softmax}(e^{r+1})h^{r+1} \] (2)

For the training of the graph neural network, we will perform global average pooling on the obtained feature map to obtain the feature vector \( l \), and calculate the cross-entropy loss function with ground truth:

\[ L_{\text{loss}} = L_{\text{CE}}(l, l^{\text{GT}}) \] (3)

### 3.2. Adversarial Erasing

Usually the classification task is the previous step of generating the CAM map, and the classification requires only the weakest supervision to achieve very good results. In the semantic segmentation task, what we need to obtain is the entire target area in the image. In adversarial erasing, the CAM finds the most discriminative object area, and then erases it from the image. The erased image is then sent to another classification network to find the secondary discriminative object area belonging to the same entity. Finally, combine the CAMs obtained twice to obtain a segmentation mask. However, these two methods still require multiple training processes or inference steps, and the attention map needs to be integrated into the segmentation template. The baseline designed loss functions for the two CNNs and trained them. The difference between our work and the baseline is that we also designed loss functions for graph neural networks and CNN networks, but the entire network will have a joint loss function as follows:

\[ L_{\text{loss}} = \lambda L_{\text{CE}}(l, l^{\text{GT}}) + L_{\text{CE}}(l^{S}, l^{\text{GT}}) \] (4)

Among \( l \) is the weight coefficient, which is used to balance the training time of the two networks. The feature vector output by the CNN network in the second step is \( l^{S} \), which is calculated by cross-entropy loss with ground truth. Combined with the loss of the neural network in the previous figure, the loss of the entire network is the joint loss function.

### 4. Experiments

We evaluated the performance of the proposed method on the Pascal VOC 2012 segmentation dataset, which is the most widely used benchmark in weakly supervised semantic segmentation. The data set consists of 20 object classes and 1 background class, including 1464 training images, 1449 verification images, and 1456 test images, respectively. Based on previous work in the literature on weakly supervised semantic segmentation, we expanded the data set to obtain a total of 11971 training images. We report the joint mean intersection (mIoU) of the validation set and the test set to measure model performance.

#### 4.1. Experimental Setup

Use the Pascal VOC2012 validation set as the data set for the evaluation model. The FCN network used in this article is DeepLab-V2 fully convolutional network. The graph neural network parameters of CAM obtained for the first time are shown in Table 1.

| Graph neural network parameter selection. |
|------------------------------------------|
| Parameter                                |
| Number of message delivery cycles        | \( t=3 \) |
| Number of graph nodes                    | \( n=4 \) |
| Best combination of parameters           | \( n=4, t=3 \) |

The backbone network is VGG16, epoch is set to 10, momentum is set to 0.9, and weight decay is set to 5e-4. For the second time to obtain CAM, the CNN setting learning rate is 0.01 and the learning rate is reduced by 10 times per 1.5k rounds of iteration, the epoch is set to 20, the momentum is set to 0.9, and the weight decay is set to 5e-4.

The frameworks of our two networks are based on Pytorch settings, and the training is completed on 4 NVIDIA RTX 2080Ti GPUs with 11GB of video memory. The testing process is completed on 1 GPU.
4.2. Comparison with baseline

Now, we compare the proposed network effect with the baseline. Table 2 lists the comparison with the baseline for each category. We found that the network framework we proposed has improved the overall performance very well. In the small target category, the effect of our network framework is better than that of the baseline. In the animal category, our improvement is also more obvious. The activation areas of these classes are concentrated on the animal's head and feet, which will cause the network to learn only a part of the object and ignore the whole. The graph neural network we proposed solves this problem very well. It learns through a set of similar images and absorbs similar semantic information from each other.

Table 2. Segmentation accuracy of each class.

| Method | bkg | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow |
|--------|-----|------|------|------|------|--------|-----|-----|-----|-------|-----|
| EADER  | 88.2| 54.9 | 31.3 | 84.1 | 58.2 | 70.9   | 83.0| 76.2| 82.1| 24.4  | 80.6|
| Ours   | 90.2| 60.1 | 32.3 | 83.2 | 57.1 | 80.1   | 86.4| 77.1| 85.7| 21.8  | 83.1|

We show the segmentation results of some networks, as shown in Figure 4. The first four columns show some successful cases of segmentation, and the last column shows cases of failed segmentation. It can be seen that through the learning of the joint loss function, the attention of the CAM can be improved to spread to most of the objects.

Figure 4. Graph Neural Network Simplified Diagram.

4.3. Comparison to the State-of-the-Art

We compare our work with the baseline and some recent studies. These works are based on weakly supervised semantic segmentation tasks. The comparison results are shown in Table 3. It can be seen that we have some improvements compared with recent work and there is not much difference compared with recent stronger work, but our training process and training takes less time.

Table 3. Comparison to the State-of-the-Art.

| Method      | Backbone | val mIoU (%) | test mIoU (%) |
|-------------|----------|--------------|---------------|
| MCOF_CVPR-18| ResNet-101| 60.3         | 61.2          |
| CIAN_CVPR-19| ResNet-101| 64.1         | 64.7          |
| ICD_CVPR-20 | ResNet-101| 64.1         | 64.3          |
| RRM-AAAI-20 | ResNet-38 | 62.6         | 62.9          |
| Ours        | ResNet-101| 63.3         | 64.9          |
5. Conclusion
In this paper, we propose a new end-to-end adversarial erasure method to solve the segmentation and localization problem inherent in weakly supervised semantic segmentation tasks. This method adds a graph neural network and does not require iterative classifiers, post-processing, weight sharing or saliency masks, which is different from many previous adversarial erasure methods. We further show that the graph neural network improves the final segmentation performance, making the network attention spread to most of the objects. End-to-end adversarial erasure improves the performance on the Pascal VOC 2012 data set, especially for most animals, achieving better positioning results.

References
[1] Dong J, Cong Y, Sun G, et al. Semantic-transferable weakly-supervised endoscopic lesions segmentation [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 10712-10721.
[2] Li Z, Zhang J, Tan T, et al. Deep learning methods for lung cancer segmentation in whole-slide histopathology images-the acdc@ lunghp challenge 2019 [J]. IEEE Journal of Biomedical and Health Informatics, 2020.
[3] Bearman A , Russakovsky O , Ferrari V , et al. What's the Point: Semantic Segmentation with Point Supervision [C]// European Conference on Computer Vision. Springer, Cham, 2016.
[4] Khoreva A , Benenson R , Hosang J , et al. Simple Does It: Weakly Supervised Instance and Semantic Segmentation [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition . IEEE Computer Society, 2016.
[5] Di L , Dai J , Jia J , et al. ScribbleSup: Scribble-Supervised Convolutional Networks for Semantic Segmentation [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition . IEEE, 2016.
[6] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1568 – 1576, 2017.
[7] Li X, Zhou T, Li J, et al. Group-wise semantic mining for weakly supervised semantic segmentation [J]. arXiv preprint arXiv:2012.05007, 2020.