J. Meas. Sci. Instrum., 2024, 15(2): 185-194. http://jmsi.publish.founderss.cn/
DOI: 10.62756/jmsi.1674-8042.2024019 jmsi@nuc.edu.cn

Improved image semantic segmentation algorithm based on EMA

DU Jiadong[1,2], LI Ting[1,2], GE Hongwei[1,2*]

1. School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China;
2. Jiangsu Provincial Engineering Laboratory for Pattern Recognition and Computational Intelligence, Jiangnan University, Wuxi 214122, China

***Corresponding author:** GE Hongwei (ghw8601@163.com)

Received: January 17, 2023 **Revised:** March 16, 2023 **Accepted:** April 15, 2023

Abstract: Aiming at the lack of semantic correlation between the parameters of expectation maximization attention (EMA) algorithm and images and the lack of attention to inter-channel information, a dual attention network EMA+ algorithm was proposed. Two modules were designed: spatial attention module and channel attention module. The EMA algorithm was used as the main structure by the spatial attention module. In the responsibility estimation step, the feature map itself was used as the initial parameter in the expectation maximization (EM) algorithm, and the semantic association between the parameter and the feature map was increased. Efficient channel attention (ECA) was used in the channel attention module by using one-dimensional convolution to learn the interactive information between channels. It avoided breaking the direct correspondence between channels and their weights due to dimensionality reduction operations. EMA+ significantly improved semantic segmentation tasks' performance by fusing spatial attention modules and channel attention modules. The experimental results showed that EMA+ has achieved better intersection-over-union than EMANet and other methods on PASCAL VOC 2012 and some more complex datasets, and had better generalization ability.

Key words: deep learning; image semantic segmentation; expectation-maximization attention (EMA); dual attention network (DANet); efficient channel attention (ECA)

0 Introduction

Semantic segmentation is a fundamental and challenging problem in computer vision, where the goal is to assign a semantic category to each pixel of an image. Semantic segmentation is the basis of visual analysis such as indoor navigation, geographic information system, human-computer interaction, automatic driving, virtual and augmented reality system, scene understanding, medical image processing, and object classification[1]. Various factors such as target diversification, irregular shapes, and object occlusion in complex environments have brought great challenges to semantic segmentation[2]. For example, in some cases, "grass" and "ground" have similar colors, and "person" may have different scales, characters, and clothes at different locations in the image. Meanwhile, the label space output by semantic segmentation is very compact, and the number of categories for a specific dataset is limited. Therefore, the task can be viewed as projecting data points in a high-dimensional noisy space into a compact sub-space. The essence is to de-noise these changes and capture the most important semantic concepts.

Many methods were proposed based on full convolutional neural (FCN)[3] to solve these problems. Due to the fixed geometry, they were inherently limited by local receptive fields and short-range contextual information. To capture long-range dependencies, some methods adopted multi-scale context fusion[4], such as dilated convolution[5], spatial pyramid[6], large kernel convolution[7], etc. Furthermore, to keep more detailed information, an encoder-decoder structure was proposed to fuse mid-level and high-level semantic features[8,9]. To aggregate information from all spatial locations, an attention mechanism was used, which enabled the features of a single pixel to fuse information from all other locations[10-12]. The basic idea of the attention mechanism was to ignore irrelevant information in the operation and focus on key information, learn contextual information through the attention mechanism, reduce dependence on external information, and capture the internal correlations of data or features[13]. Attention was widely used in various tasks such as machine translation, visual question answering, and

video classification. The contextual encoding of each position was computed by embedding a weighted sum over all positions in the sentence[10]. The self-attention mechanism was first adopted by non-local as a module for computer vision tasks, such as video classification, object detection, and instance segmentation[11]. Aggregate contextual information was learned for each location through a predicted attention map for PSANet[12]. A dual attention block was proposed to distribute and collect informative global features from the entire spatiotemporal space of an image for A^2Net[14]. Spatial and channel attention were applied to gather information around feature maps for DANet[15].

Li et al.[16] proposed a novel attention-based method based on the expectation-maximization (EM) algorithm, namely expectation-maximization attention (EMA). A randomly generated initial basis $\boldsymbol{\mu}$ was used as a parameter in the EMA module, which was updated by using sliding average in each batch training, such that the EMA module played a similar role to a convolutional layer. And the EMA module was not specific enough to extract contextual information and ignored detailed features. In addition, what EMA captured was the spatial interaction information, and it lacked attention to the interaction information between channels.

A dual attention network EMA+ algorithm was designed. The improved EMA algorithm was used on the spatial attention module, the EM algorithm was applied to each feature map by using the feature map itself as the initial $\boldsymbol{\mu}$, and the spatial interaction information of the feature map was extracted by EM algorithm. On the channel attention module, a method was adopted that did not use dimensionality reduction operations[17] to capture the correlations between channels as much as possible. The obtained extensive experimental results on three challenging semantic segmentation datasets PASCAL VOC 2012, Cityscapes, and ADE20K demonstrated that the proposed method outperformed other competitive state-of-the-art methods.

1 EMA

EMA consists of three operations, including responsibility estimation (A_E), likelihood maximization (A_M), and data re-estimation (A_R). Given an input $\boldsymbol{X}\in\mathbf{R}^{N\times C}$ and an initial basis $\boldsymbol{\mu}\in\mathbf{R}^{K\times C}$, A_E estimates the latent variable (or "responsibility") $\boldsymbol{Z}\in\mathbf{R}^{N\times K}$, so it functions as the E-step in the EM algorithm. A_M uses the estimate to update the basis $\boldsymbol{\mu}$, which is equivalent to M-step. The A_E and A_M steps are executed alternately for a pre-specified number of iterations. Then, through the converged $\boldsymbol{\mu}$ and $\boldsymbol{Z}$, A_R reconstructs the original $\boldsymbol{X}$ into $\tilde{\boldsymbol{X}}$ and outputs it.

1.1 Responsibility estimation

The responsibility estimation (A_E) is used as the E-step in the EM algorithm. This step is to calculate the expected value of z_{nk} in $\boldsymbol{Z}$, corresponding to the responsibility of the k-th basis $\boldsymbol{\mu}_k$ to $\boldsymbol{x}_n$, where $1\leqslant k\leqslant K$ and $1\leqslant n\leqslant N$. The formula for the posterior probability of $\boldsymbol{x}_n$ in $\boldsymbol{X}$ under $\boldsymbol{\mu}_k$ is

$$p(\boldsymbol{x}_n,\boldsymbol{\mu}_k)=\mathcal{K}(\boldsymbol{x}_n,\boldsymbol{\mu}_k), \tag{1}$$

where $\mathcal{K}$ is a general kernel function. And Eq. (1) can be rewritten into a more general form, that is

$$z_{nk}=\frac{\mathcal{K}(\boldsymbol{x}_n,\boldsymbol{\mu}_k)}{\sum_{j=1}^{N}\mathcal{K}(\boldsymbol{x}_n,\boldsymbol{\mu}_j)}. \tag{2}$$

There are many options for $\mathcal{K}$, such as inner product $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b}$, index of inner product $\exp(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b})$, Euclidean distance $\|\boldsymbol{a}-\boldsymbol{b}\|_2^2$, RBF kernel $\exp(-\|\boldsymbol{a}-\boldsymbol{b}\|_2^2/\sigma^2)$ and so on. The choice of these functions makes only negligible difference in the final result, so the exponent of the inner product $\exp(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b})$ is used. In the experiment, Eq. (2) is implemented as matrix multiplication plus a Softmax layer. In summary, the calculation formula of A_E in the t-th iteration is

$$\boldsymbol{Z}^{(t)}=\mathrm{softmax}(\lambda\boldsymbol{X}(\boldsymbol{\mu}^{(t-1)})^{\mathrm{T}}), \tag{3}$$

where λ is a hyperparameter controlling the distribution of $\boldsymbol{Z}$.

1.2 Likelihood maximization

Likelihood maximization (A_M) was used as the M-step of the EM algorithm. Using the estimated $\boldsymbol{Z}$ value, A_M updates $\boldsymbol{\mu}$ by maximizing the full data likelihood. To keep the basis in the same embedding space as $\boldsymbol{X}$, the basis $\boldsymbol{\mu}$ is updated using a weighted sum of $\boldsymbol{X}$. So in the tth iteration of A_M, $\boldsymbol{\mu}_k$ is updated by

$$\boldsymbol{\mu}_k^{(t)}=\frac{z_{nk}^{(t)}\boldsymbol{x}_n}{\sum_{m=1}^{N}z_{mk}^{(t)}}. \tag{4}$$

1.3 Data re-estimation

After EMA alternately runs the A_E and A_M algorithms T times, the final $\boldsymbol{\mu}^{(T)}$ and $\boldsymbol{Z}^{(T)}$ are used to re-estimate $\boldsymbol{X}$. Using Eq. (5) to construct a new $\boldsymbol{X}$, namely $\tilde{\boldsymbol{X}}$, the formula is

$$\tilde{\boldsymbol{X}}=\boldsymbol{Z}^{(T)}\boldsymbol{\mu}^{(T)}. \tag{5}$$

2 Proposed method

2.1 EMA+

In this paper, a dual attention network EMA+ algorithm was designed, and the overall architecture is shown in Fig. 1. EMA+ contained a spatial attention module and a channel attention module.

The EMA mechanism was used on the spatial attention module. To make the initial basis contain the semantic features of the feature map as much as possible, in the responsibility estimation (A_E) step, the parameter $\boldsymbol{\mu}$ of the model was set to equal to the input feature map $\boldsymbol{X}$, and it was used to estimate the hidden variable $\boldsymbol{Z}$. The likelihood maximization (A_M) step updated the parameter $\boldsymbol{\mu}$ based on $\boldsymbol{X}$ and $\boldsymbol{Z}$. After A_E and A_M, T steps were performed alternately, the approximately converged $\boldsymbol{\mu}$ and $\boldsymbol{Z}$ re-estimated $\boldsymbol{X}$ (A_R). Compared with the $\boldsymbol{\mu}$ value initialized randomly by EMA, choosing $\boldsymbol{X}$ as the $\boldsymbol{\mu}$ value did not require a moving average update on each small batch, and the $\boldsymbol{\mu}$ value could be updated completely according to the EM algorithm.

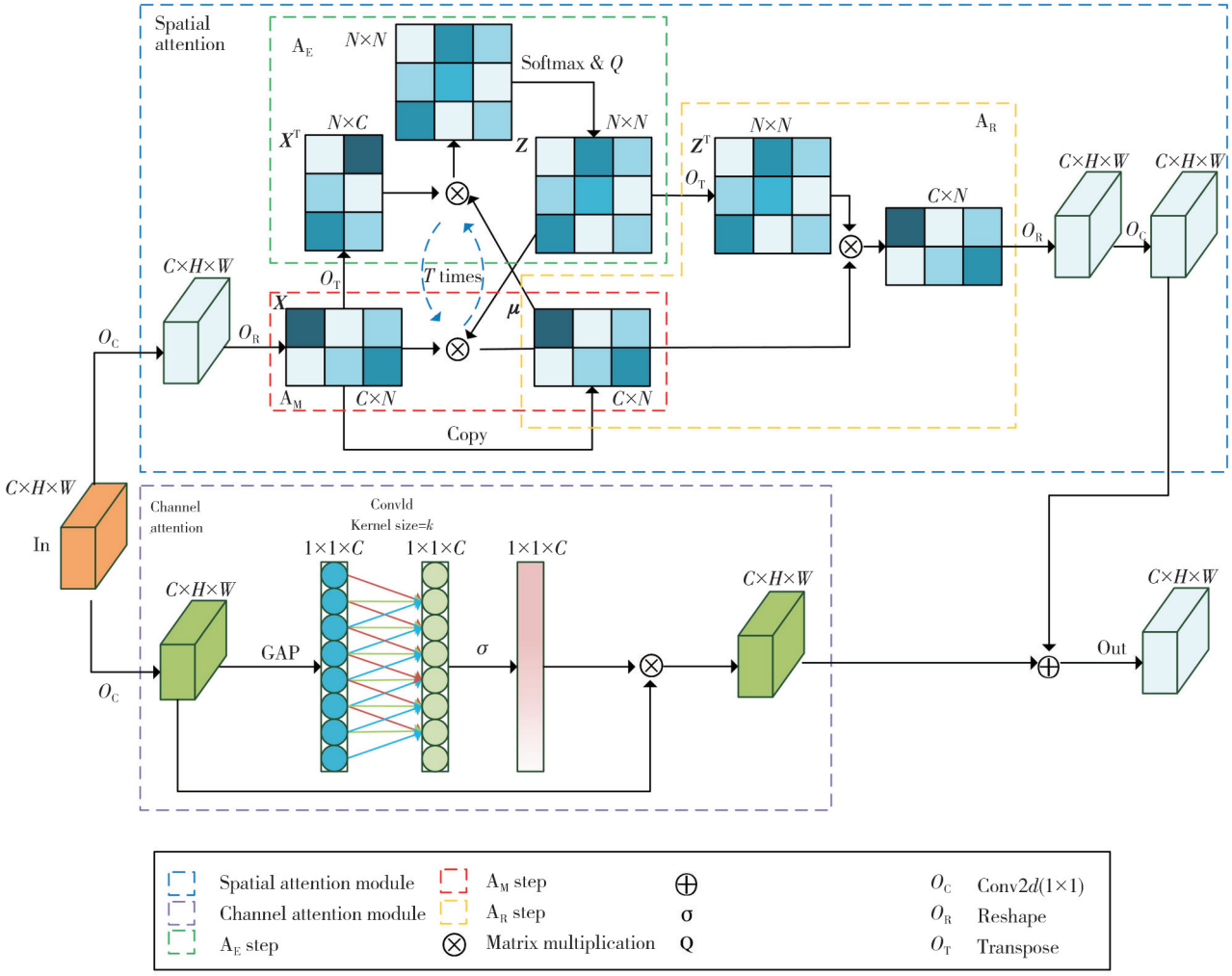

Fig. 1 Overall architecture of EMA+

A 1×1 convolution without ReLU activation before the EM module was used to convert the input value range from $(0, +\infty)$ to $(-\infty, +\infty)$, which could make the initial $\boldsymbol{\mu}$ also be located at $(-\infty, +\infty)$. At the end of spatial attention module, a 1×1 convolution was used to transform the re-estimated $\tilde{\boldsymbol{X}}$ and add the output of the channel attention module, which could be regarded as the residual space of $\boldsymbol{X}$ processed by the channel attention module, which increased the expressiveness of the network.

In the channel attention module, the efficient channel attention (ECA) [17] was drew on to extract the interactive information between channels through one-dimensional convolution to avoid the direct correspondence between the channel and its weight destroyed by the dimensionality reduction operation[18]. Since the position of the attention module in the whole network was fixed, that was, there was a fixed number of input channels, the optimal convolution kernel size k could be determined through experiments. Finally, the output

features of the two attention modules were fused, and the final prediction segmentation map was restored through upsampling.

2.2 Spatial attention

The input $\boldsymbol{X}$ of spatial attention was the intermediate activation of CNN, the size was $C\times H\times W$, through a 1×1 convolution without ReLU activation, the input value range was converted from $(0,+\infty)$ to $(-\infty,+\infty)$. $\boldsymbol{X}$ was reshaped to $C\times N$, where $N=H\times W$. For a given input $\boldsymbol{X}\in\mathbf{R}^{C\times N}$, the initial basis $\boldsymbol{\mu}=\boldsymbol{X}\in\mathbf{R}^{C\times N}$, $\mathrm{A_E}$ estimated the latent variable $Z\in\mathbf{R}^{N\times N}$. At this time it corresponds to the responsibility of the n_2th basis $\boldsymbol{\mu}_{n_2}$ to $\boldsymbol{x}_{n_1}$, where $1\leqslant n_1,\ n_2\leqslant N$. Then Eq. (2) can be expressed as

$$z_{n_1n_2}=\frac{\mathcal{K}(\boldsymbol{x}_{n_1},\boldsymbol{\mu}_{n_2})}{\sum_{j=1}^{N}\mathcal{K}(\boldsymbol{x}_{n_1},\boldsymbol{\mu}_j)}. \tag{6}$$

The kernel function $\mathcal{K}$ selects the index of the inner product, then the value $\lambda=1$ of Eq. (3) can be expressed as

$$\boldsymbol{Z}^{(t)}=\mathrm{softmax}(\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{\mu}^{(t-1)})). \tag{7}$$

When $t=1$, $\boldsymbol{\mu}^{(0)}=\boldsymbol{X}$ satisfies

$$\boldsymbol{Z}^{(1)}=\mathrm{softmax}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}). \tag{8}$$

$\mathrm{A_M}$ uses the estimated $\boldsymbol{Z}$ value to update the basis $\boldsymbol{\mu}$. So in the tth iteration of $\mathrm{A_M}$, $\boldsymbol{\mu}_{n_2}$ is updated by

$$\boldsymbol{\mu}_{n_2}^{(t)}=\frac{z_{n_1n_2}^{(t)}\boldsymbol{x}_{n_1}}{\sum_{m=1}^{N}z_{mn_2}^{(t)}}. \tag{9}$$

After the $\mathrm{A_E}$ and $\mathrm{A_M}$ steps are executed alternately T times, the final $\boldsymbol{\mu}^{(T)}$ and $\boldsymbol{Z}^{(T)}$ are used to re-estimate $\boldsymbol{X}$, that is $\tilde{\boldsymbol{X}}$, the formula is

$$\tilde{\boldsymbol{X}}=\boldsymbol{\mu}^{(T)}(\boldsymbol{Z}^{(T)})^{\mathrm{T}}. \tag{10}$$

After getting the revalued $\tilde{\boldsymbol{X}}$, the formula for the spatial attention output $\boldsymbol{Y}_{\mathrm{S}}$ can be given as

$$\boldsymbol{Y}_{\mathrm{S}}=f^{1\times1}(\tilde{\boldsymbol{X}})=f^{1\times1}(\boldsymbol{\mu}^{(T)}(\boldsymbol{Z}^{(T)})^{\mathrm{T}}), \tag{11}$$

where $f^{1\times1}$ is a 1×1 convolution.

2.3 Channel attention

For an input feature map $\boldsymbol{X}$ of size $C\times H\times W$, a 1×1 convolution without ReLU activation was first used to convert the input $\boldsymbol{X}$ range from $(0,+\infty)$ to $(-\infty,+\infty)$, then through the aggregated features obtained by global average pooling (GAP), at this time $\boldsymbol{X}'\in\mathbf{R}^{1\times1\times C}$. In order to avoid the loss of information caused by the dimensionality reduction operation on the image, one-dimensional convolution was used to generate channel weights. The convolution kernel size k of one-dimensional convolution was determined by experiment, then the result activated by the Sigmoid function was multiplied by the input $\boldsymbol{X}$ to restore the size of $C\times H\times W$. Then, the output $\boldsymbol{Y}_{\mathrm{C}}$ calculation formula of channel attention is

$$\boldsymbol{Y}_{\mathrm{C}}=\boldsymbol{X}\otimes(\sigma(f^{1\times1}(g(\boldsymbol{X})))), \tag{12}$$

where $g(\boldsymbol{X})=\dfrac{\sum_{i=1,j=1}^{W,H}x_{ij}}{WH}$ is the channel-level global average pooling (GAP), and σ is the Sigmoid function.

So the formula for calculating the total output $\boldsymbol{Y}$ of the dual attention module is

$$\boldsymbol{Y}=\boldsymbol{Y}_{\mathrm{S}}+\boldsymbol{Y}_{\mathrm{C}}. \tag{13}$$

Substituting $\boldsymbol{Y}_{\mathrm{S}}$ and $\boldsymbol{Y}_{\mathrm{C}}$ into the Eq. (13), there is

$$\boldsymbol{Y}=f^{1\times1}(\boldsymbol{\mu}^{(T)}(\boldsymbol{Z}^{(T)})^{\mathrm{T}})+\boldsymbol{X}\otimes(\sigma(f^{1\times1}(g(\boldsymbol{X})))). \tag{14}$$

3 Experiments

To evaluate the proposed method, comprehensive experiments were conducted on the PASCAL VOC 2012 dataset[19], the ADE20K dataset[20], the cityscapes dataset[21], the PASCAL context dataset[22], and the COCO stuff 10K dataset[23]. Experimental results showed that the proposed method achieved state-of-the-art performance on all datasets. Next, the dataset and implementation details was first introduced, and then a series of parameter selection experiments and ablation studies were conducted on the PASCAL VOC 2012 dataset. Finally, the proposed method was compared with some state-of-the-art techniques on three datasets.

3.1 Datasets and evaluation index

3.1.1 Datasets

1) The PASCAL VOC 2012 data set had 17 125 image data for different tasks. The pixel size of each image was different. The horizontal image size was about 375×500 pixels, and the vertical image size was about 500×375 pixels. For the semantic segmentation task, there were a total of 2 913 images. These data contained objects of 4 major categories and 20 small categories (excluding background categories). The training set had 1 464 images and the verification set had 1 449 images.

2) The cityscapes dataset contained 5 000 images from 50 different cities. Each image was 2 048×1 024 pixels with high-quality pixel-level labels for 19 semantic categories. There were 2 975 images in the training set, 500 images in the validation set, and 1 525 images in the test set.

3) The ADE20K dataset was a new scene parsing benchmark and contained 20 210 images for training,

2 000 images for validation, and 3 352 images for testing. There were a total of 150 categories in this dataset, including 35 object classes (e.g., walls, sky, roads) and 115 discrete object classes (e.g., cars, people, tables). The unbalanced distribution of labels and images of different scales made the dataset more challenging.

4) The PASCAL context dataset was an extension of the PASCAL VOC 2010 detections. The dataset had a total of 459 labeled categories and contained 10 103 images, of which 4 998 were used for training and 5 105 were used for validation. Usually, the 59 categories with the highest frequency were selected as semantic labels. The rest of the classes were labeled as background classes.

5) The COCO stuff 10 K dataset was a large-scale scene understanding and semantic segmentation dataset, including 80 categories of objects and 91 categories of "stuff". Object categories included people, vehicles, animals, furniture, etc., while stuff categories included background elements in scenes such as sky, ground, water, grass, and buildings. The dataset included 10 000 images, about half of which were training set images from the COCO dataset and the other half were validation set images randomly selected from the COCO dataset.

3.1.2 Evaluation index

In order to evaluate the performance of the algorithm in this paper, the commonly used semantic segmentation evaluation index mIoU (mean intersection over union) [24] was used, which was the average value of the accumulated IoU values of each class of image pixels. The detailed calculation formula is

$$mIoU=\frac{1}{k}\sum_{i=1}^{k}\frac{p_{ii}}{t_i+\sum_{j=1}^{k}(p_{ji}-p_{jj})} \tag{15}$$

where k is the number of categories of pixels; p_{ii} is the number of pixels whose actual category is i and the predicted category is also i; t_i is the total number of pixels of category i; p_{ji} is the number of pixels whose actual category is i and predicted category is j.

3.2 Implementation details

CPU was Intel(R) Xeon(R) Gold 6240 CPU @ 2.60 GHz. RAM was 32 GB. GPU was NVIDIA Tesla V100 32 GB. Operating system was Ubuntu 18.04.3 LTS, using python3.7.6 and deep learning framework Pytorch1.8. ResNet[25] (Pretrained on imageNet[26]) was used as backbone. The number of channels was reduced from 2 048 to 512 using 3×3 convolution kernels, and then the attention modules were stacked on top of it. The entire network was called as EMA+Net. Following previous work[6,7,9], a multivariate learning rate strategy was adopted, where the initial learning rate was multiplied by $(1-iter/total\_iter)^{0.9}$ after each iteration. The initial learning rate for all datasets was set to 0.009. The momentum and weight decay coefficients were set to 0.9 and 0.000 1, respectively. For data augmentation, common scaling factors (0.5 to 2.0) are applied to crop and flip images to augment the training data. The input size for all datasets is set to 513×513 pixels.

For training and evaluation on all datasets, the backbone output stride was set to 8. Restricted by the experimental equipment, a Tesla V100 GPU had a memory capacity of 32 GB, and the batch size of all our experiments was 8. Parameter selection experiments and ablation studies were conducted on the PASCAL VOC 2012 dataset. In order to compare the proposed model with other advanced network models, these models were trained on ResNet-101[17] the same number of iterations on each dataset.

3.3 Results on PASCAL VOC 2012 dataset

Experiments were conducted on the PASCAL VOC 2012 dataset, and all parameter selection experiments and ablation studies on ResNet-50[25] were conducted to speed up the training process. Because the number of iterations T is a parameter of the spatial attention module and the convolution kernel size k is a parameter of the channel attention module, there is no negative correlation between their impact on the performance of the model. Therefore, the parameter k was first preset as a common value to conduct a selection experiment on the parameter T, and then a selection experiment was conducted on the parameter k after determining the optimal parameter T value.

3.3.1 Selection experiment of iteration number T

Set the most commonly used convolution kernel size $k=3$, and trained in the range of $1\leqslant T\leqslant 8$. The results are shown in Fig. 2. When $T\leqslant 4$, as the number of iterations increased, the performance of the model improved, and reached the optimal value when $T=4$. When $T>4$, the performance began to fluctuate and decline, and the performance was generally better when T was even than when T was odd.

3.3.2 Selection experiment of convolution kernel size k

After confirming that the optimal effect could be achieved when the number of EM iterations T was 4, set the number of iterations $T=4$, and trained in the range of $3\leqslant k\leqslant 9$ (k was an odd number). The results are shown in Table 1. When $k=3$, the performance was the best, and as k increased, the performance began to decline.

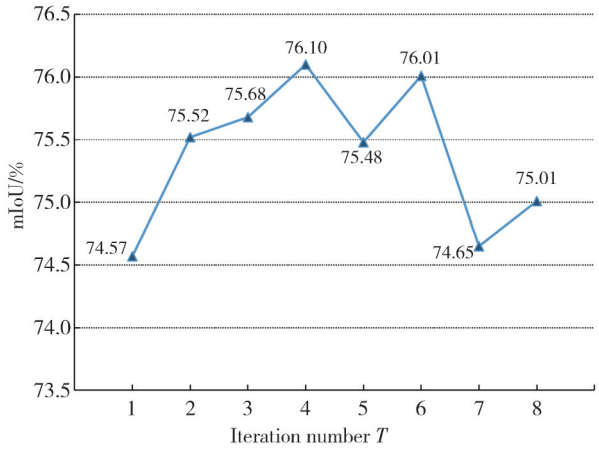

Fig. 2 Influence of mIoU on parameter iteration number *T*

Table 1 Influence of mIoU on parameter convolution kernel size *k*

k	mIoU/%
3	**76.10**
5	75.99
7	74.95
9	74.74

3.3.3 Ablation study of attention module

Set the number of EM iterations $T=4$, the one-dimensional convolution kernel size $k=3$, and conduct ablation studies on the attention module. The experimental results are shown in Table 2. Compared with the basic ResNet network without the attention module, the spatial attention module improved mIoU metrics by about 2%, and the channel attention module improved mIoU metrics by about 2%. The combination of spatial attention and channel attention improved the mIoU index by about 3.5%. It showed that the attention network had significantly improved the performance of semantic segmentation.

Table 2 Ablation study on attention module

Spatial attention module	Channel attention module	mIoU/%
×	×	72.60
×	√	74.82
√	×	75.23
√	√	**76.10**

3.3.4 Comparisons with other advanced network models

To further evaluate the performance of the proposed network model, it was trained with other state-of-the-art models on ResNet-101[25], performing the same 60 000 iterations on each dataset. The results are shown in Table 3, and the proposed method outperformed other competitive advanced network models listed under the same experimental conditions.

Table 3 Comparisons with other advanced network models on PASCAL VOC 2012

Method	Backbone	mIoU/%
EANet[27]	ResNet-101	75.8
PSPNet[6]	ResNet-101	76.4
Deeplabv3-JFT[5]	ResNet-101	76.7
PSANet[12]	ResNet-101	76.7
EncNet[28]	ResNet-101	76.9
DFN[29]	ResNet-101	77.2
EMANet[16]	ResNet-101	78.7
EMA+Net (Proposed)	ResNet-101	**79.9**

In Table 3, EMANet was one of the current optimal network models. In order to visually show the performance improvement of this method compared with EMANet, a visual comparison on the output prediction map is made, as shown in Fig.3.

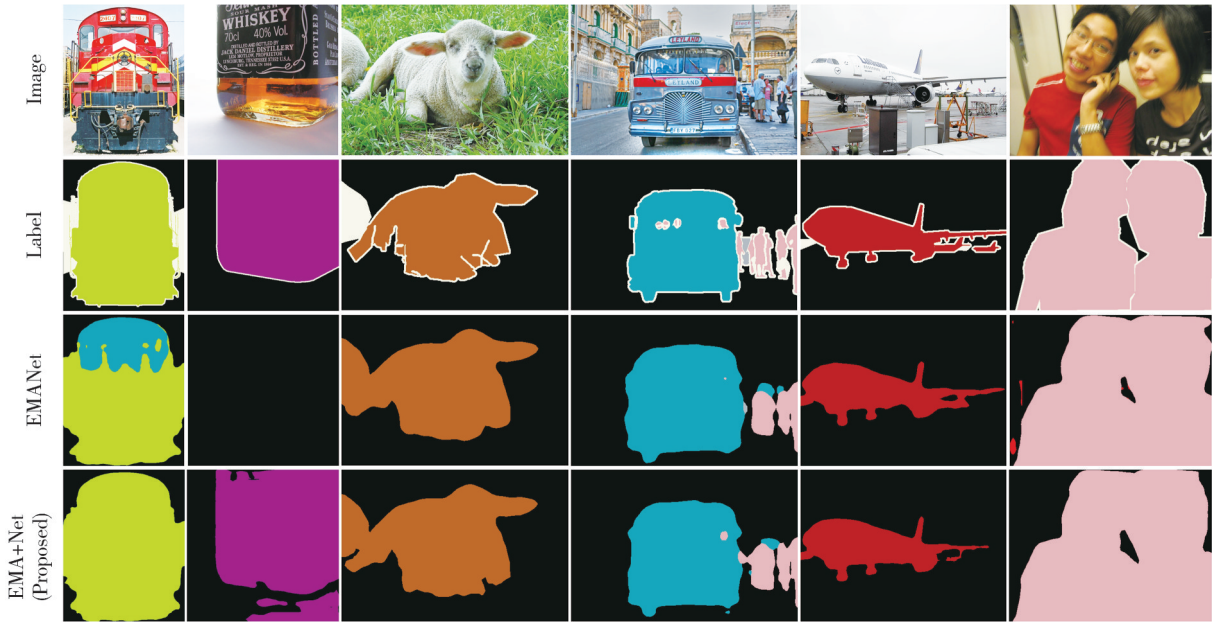

Fig. 3 Visual comparison of segmentation results between EMA+Net and EMANet on PASCAL VOC 2012 dataset

In the prediction of objects that accounted for a large proportion of the entire image area, the proposed method was more accurate than EMANet, which confirmed the theory that the double attention network was more accurate for its own feature extraction.

3.4 Results on cityscapes dataset

Experiments on the cityscapes dataset were conducted to validate the effectiveness of the proposed method. The quantitative results on the cityscapes dataset are shown in Table 4. The results showed that the proposed method also exhibited excellent performance on the cityscapes dataset. Several representative pictures are selected for comparison, as shown in Fig.4.

Table 4 Comparisons with other advanced network models on cityscapes

Method	Backbone	mIoU/%
DeepLabv3[5]	ResNet-101	70.3
PSANet[12]	ResNet-101	70.4
SPNet[30]	ResNet-101	71.0
CAA[31]	ResNet-101	71.6
EMANet[16]	ResNet-101	74.8
EMA+Net (Proposed)	ResNet-101	**76.5**

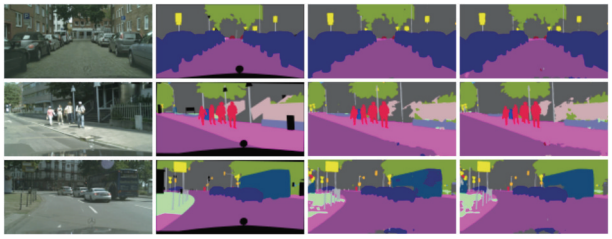

Fig. 4 Visual comparison of segmentation results between EMA+Net and EMANet on cityscapes dataset

3.5 Results on ADE20K dataset

To further evaluate the effectiveness of the proposed method, experiments on the ADE20K dataset were also conducted. The comparison with some current state-of-the-art methods is shown in Table 5. It was worth noting that the mIoU of the proposed method reached 56.3%, greatly outperforming previous methods, including some transformer methods.

Table 5 Comparisons with other advanced network models on ADE20K

Method	Backbone	mIoU/%
EMANet[16]	ResNet-101	50.4
BeiT-L[32]	ViT+UperNet	55.6
MaskFormer[33]	SwinL	55.6
CSWin-L[34]	UperNet	55.7
SeMask[35]	SeMask Swin-L MaskFormer	56.2
EMA+Net (Proposed)	ResNet-101	**56.3**

3.6 Results on PASCAL context dataset

Experiments on the PASCAL context dataset were also conducted, and the comparison results with some current advanced methods are shown in Table 6. The results showed that the proposed method had better performance, and the mIoU index reached 53.9%, which was 0.8% higher than EMANet. In order to display the results more intuitively, several representative pictures are selected for visual comparison, as shown in Fig.5. It could be seen that the proposed method also had good segmentation results on some samples that EMANet could not segment well.

Table 6 Comparisons with other advanced network models on PASCAL context dataset

Method	Backbone	mIoU/%
RefineNet[36]	ResNet-101	47.3
PSPNet[6]	ResNet-101	47.8
EncNet[28]	ResNet-101	51.7
DANet[15]	ResNet-101	52.6
EMANet[16]	ResNet-101	53.1
EMA+Net (Proposed)	ResNet-101	**53.9**

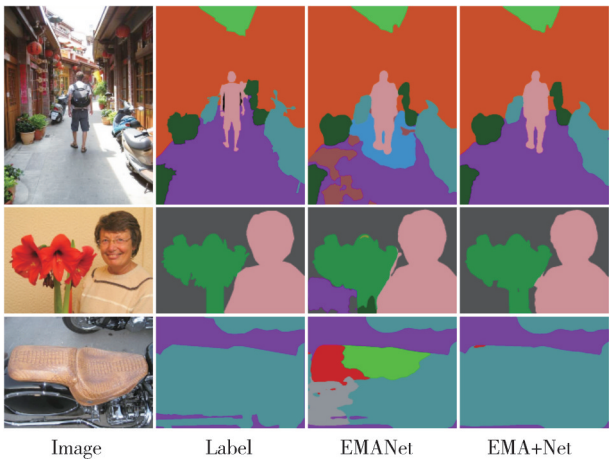

Fig. 5 Visual comparison of segmentation results between EMA+Net and EMANet on PASCAL context dataset

3.7 Results on COCO stuff 10K dataset

In order to continue to verify the generalization ability of the proposed method, experiments were conducted on the COCO stuff 10 K dataset which was a more complex scene analysis dataset, and the comparison results with some current advanced methods are shown in Table 7. The results showed that the performance of the proposed method was better than other methods, and the mIoU index was 0.3% higher than that of EMANet. Some examples of the COCO stuff 10 K validation set are shown in Fig. 6. It could be seen that the proposed method performed well in face of large area object segmentation.

Table 7 Comparisons with other advanced network models on COCO stuff 10K dataset

Method	Backbone	mIoU/%
RefineNet[36]	ResNet-101	33.6
SVCNet[37]	ResNet-101	39.6
DANet[15]	ResNet-101	39.7
EMANet[16]	ResNet-101	39.9
EMA+Net (Proposed)	ResNet-101	**40.2**

Fig. 6 Visual comparison of segmentation results between EMA+Net and EMANet on COCO stuff 10 K dataset

4 Conclusions

Due to the lack of semantic correlation between parameters and images and insufficient attention to inter-channel information in the EMA algorithm in image semantic segmentation tasks, a dual attention network EMA+ algorithm was proposed. This method designed two modules: spatial attention module and channel attention module. The improved EMA algorithm in the spatial attention module was used to increase the semantic association between parameters and feature maps. In the channel attention module, the lighter ECA was used to capture the interactive information between channels. EMA+ significantly improved the performance of semantic segmentation tasks. The EMA+ algorithm was evaluated on PASCAL VOC 2012 and some more complex datasets. A large number of experiments showed that the proposed method achieved high segmentation accuracy and had good generalization ability. However, there is still a lot of room for improvement in the segmentation performance of the algorithm, and the attention mechanism needs to capture more contextual information, so combining other classical machine learning algorithms with the attention mechanism is a further research direction.

Acknowledgement

This work was supported by National Natural Science Foundation of China (No.61806006); Priority Academic Program Development of Jiangsu Higher Education Institutions; 111 Project (No.B12018).

Declaration of conflicting interests

The authors have no conflict of interests related to this publication.

References

[1] LUO H L, ZHANG Y. A survey of image semantic segmentation based on deep network. Acta Electronica Sinica, 2019, 47(10): 2211-2220.

[2] XU H, ZHU Y H, ZHEN T, et al. Survey of semantic methods based on deep neural network. Journal of Frontiers of Computer Science and Technology, 2021, 15(1): 47-59.

[3] SHELHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651.

[4] LI X, WU J, LIN Z, et al. Recurrent squeeze-and-excitation context aggregation net for single image deraining//European Conference on Computer Vision, September 8-14, Munich, Germany. Berling: Springer, 2018: 254-269.

[5] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation. Computer Vision and Pattern Recognition, arXiv: 1706.05587.

[6] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network//IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, Honolulu, HI, New York: IEEE, 2017: 6230-6239.

[7] PENG C, ZHANG X, YU G, et al. Large kernel matters-improve semantic segmentation by global convolutional network//IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, New York: IEEE, 2017: 1743-1751.

[8] YU C, WANG J, PENG C, et al. Learning a discriminative feature network for semantic segmentation//IEEE Conference on Computer Vision and Pattern Recognition, June 18-23, Salt Lake City, UT, New York: IEEE, 2018: 1857-1866.

[9] CHEN L C, ZHU Y, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation//European Conference on Computer Vision, September 8-14, 2018, Munich, Germany. Berling: Springer, 2018: 833-851.

[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need//The 31st International Conference on Neural Information Processing Systems (NIPS'17), December 4-9, Red Hook, NY, USA, Cambridge: MIT Press, 2017: 6000-6010.

[11] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks//IEEE Conference on Computer Vision and Pattern Recognition, August 20-24, Salt Lake City, UT, USA, New York: IEEE, 2018: 7794-7803.

[12] ZHAO H, ZHANG Y, LIU S, et al. Psanet: Point-wise spatial attention network for scene parsing//European Conference on Computer Vision. Munich, September 8-14, 2018, Munich, Germany. Berling: Springer, 2018: 270-286.

[13] YANG Z Q, FAN Y S, YU H Y. An improved image semantic segmentation algorithm based on U-Net network. Journal of North University of China (Natural Science Edition), 2023, 44(4): 397-402.

[14] CHEN Y, KALANTIDIS Y, LI J, et al. A^2-nets: Double attention networks//The 32nd International Conference on Neural Information Processing Systems (NIPS' 18), December 3-8, Red Hook, NY, USA. Cambridge: MIT Press, 2018: 350-359.

[15] FU J, LIU J, TIAN H, et al. Dual attention network for scene segmentation//IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 15-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 3141-3149.

[16] LI X, ZHONG Z, WU J, et al. Expectation-maximization attention networks for semantic segmentation//IEEE/CVF International Conference on Computer Vision, October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE, 2019: 9166-9175.

[17] WANG Q, WU B, ZHU P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks // IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 13-19, Seattle, WA, USA. New York: IEEE, 2020: 11531-11539.

[18] HU J, SHEN L, SUN G. Squeeze-and-excitation networks// IEEE Conference on Computer Vision and Pattern Recognition, June 18-23, Salt Lake City, UT, USA. New York: IEEE, 2018: 7132-7141.

[19] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, et al. The pascal visual object classes (voc) challenge. International Journal of Computer Vision, 2010, 88(2): 303-338.

[20] ZHOU B, ZHAO H, PUIG X, et al. Scene parsing through ADE20K dataset//IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 5122-5130.

[21] CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding//IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 3213-3223.

[22] MOTTAGHI R, CHEN X, LIU X, et al. The role of context for object detection and semantic segmentation in the wild//IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014: 891-898.

[23] CAESAR H, UIJLINGS J, FERRARI V. Coco-stuff: Thing and stuff classes in context//IEEE Conference on Computer Vision and Pattern Recognition, August 20-24, Salt Lake City, UT, USA. New York: IEEE, 2018: 1209-1218.

[24] GARCIA-GARCIA A, ORTS-ESCOLANO S, OPREA S, et al. A review on deep learning techniques applied to semantic segmentation. Computer Vision and Pattern Recognition, arXiv:1704.06857.

[25] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition//IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[26] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3): 211-252.

[27] GUO M H, LIU Z N, MU T J, et al. Beyond self-attention: external attention using two linear layers for visual tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(5): 5436-5447.

[28] ZHANG H, DANA K, SHI J, et al. Context encoding for semantic segmentation//IEEE Conference on Computer Vision and Pattern Recognition, August 20-24, Salt Lake City, UT, USA. New York: IEEE, 2018: 7151-7160.

[29] YU C, WANG J, PENG C, et al. Learning a discriminative feature network for semantic segmentation// IEEE Conference on Computer Vision and Pattern Recognition, August 20-24, Salt Lake City, UT, USA. New York: IEEE, 2018: 1857-1866.

[30] HOU Q, ZHANG L, CHENG M M, et al. Strip pooling: Rethinking spatial pooling for scene parsing//IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 13-19, 2020, Seattle, WA, USA, New York: IEEE, 2020: 4002-4011.

[31] HUANG Y, KANG D, JIA W, et al. Channelized axial attention for semantic segmentation-considering channel relation within spatial attention for semantic segmentation. Computer Vision and Pattern Recognition, arXiv: 2101.07434.

[32] BAO H, DONG L, PIAO S, et al. Beit: Bert pre-training of image transformers. Computer Vision and Pattern Recognition, arXiv: 2106.08254.

[33] CHENG B, SCHWING A, KIRILLOV A. Per-pixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems, 2021, 34: 17864-17875.

[34] DONG X, BAO J, CHEN D, et al. Cswin transformer: A general vision transformer backbone with cross-shaped windows//IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-24, 2022, New Orleans, Louisiana, USA, New York: IEEE, 2022: 12114-12124.

[35] JAIN J, SINGH A, ORLOV N, et al. Semask: Semantically masked transformers for semantic segmentation. Computer Vision and Pattern Recognition, arXiv: 2112.12782.

[36] LIN G, MILAN A, SHEN C, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation//IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 5168-5177.

[37] DING H, JIANG X, SHUAI B, et al. Semantic correlation

promoted shape-variant context for segmentation//IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 15-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 8877-8886.

基于EMA改进的图像语义分割算法

杜佳栋[1,2], 李 婷[1,2], 葛洪伟[1,2*]

1. 江南大学 人工智能与计算机学院, 江苏 无锡 214122;
2. 江南大学 江苏省模式识别与计算智能工程实验室, 江苏 无锡 214122

摘 要: 针对期望最大化注意(EMA)算法参数与图像的语义关联不足以及缺少对通道间信息关注的问题, 本文提出一种双重注意力网络EMA+算法。该算法设计了2个模块: 空间注意力模块和通道注意力模块。空间注意力模块以EMA算法为主体架构, 在责任估计步骤采用特征图作为期望最大化(EM)算法的初始参数, 增加参数与特征图语义上的关联。通道注意力模块使用高效通道注意力(ECA), 通过使用一维卷积学习通道之间交互信息, 避免由于降维操作导致的破坏通道与其权重之间的直接对应关系。EMA+通过融合空间注意力模块和通道注意力模块, 显著提高了语义分割任务的性能。实验结果表明, EMA+在PASCAL VOC 2012和一些更复杂的数据集上均取得了较EMANet等方法更优的交并比指标, 有较好的泛化能力。

关键词: 深度学习; 图像语义分割; 期望最大化注意; 双重注意力网络; 高效通道注意力模块

引用格式: DU Jiadong, LI Ting, GE Hongwei. Improved image semantic segmentation algorithm based on EMA. Journal of Measurement Science and Instrumentation, 2024, 15(2): 185-194.