International Journal of Science and Engineering Applications
Volume 13-Issue 10, 102 – 109, 2024, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1310.1022

Multi-Label Image Recognition Based on Attention and Multi-Scale Dynamic Graph Convolutional Network

Luo Litao
College of Electronic Information and Electrical Engineering
Yangtze University
Jingzhou City, Hubei Province
China

Lin Qihuang
College of Electronic Information and Electrical Engineering
Yangtze University
Jingzhou City, Hubei Province
China

Abstract: In order to introduce the semantic correlation between labels into the multi-label image classification model, ADD-GCN (Attention-Driven Dynamic Graph Convolutional Network) generates a dynamic graph for each image. The Dynamic Graph Convolutional Network (D-GCN) is used to model the relationship between the content-aware category representations generated by the Semantic Attention Module (SAM) to avoid frequency bias.However, ADD-GCN cannot automatically learn and selectively focus on important information in the input. When transmitting data, it is prone to instability and difficult to summarize all label semantic relationships. Aiming at the problem of ADD-GCN, a new multi-label image recognition model based on multi-scale dynamic graph convolutional network is proposed based on ADD-GCN.The model updates the multi-scale feature extraction module and integrates the Convolutional Block Attention Module (CBAM) to automatically focus on important information. In order to improve the data transmission of SAM and D-GCN, Gaussian Error Linear Units (GELU) is integrated to complete the mapping of neurons, and Adaptive Moment Estimation (Adam) and Binary Cross-Entropy With Logits Loss (BCEWithLogitsLoss) are introduced to make the data propagate stably. Compared with the original ADD-GCN model, the average accuracy of the algorithm in the MS-COCO data set and the PASCAL VOC data set reached 87.1 % and 95.6 %, respectively, which has better detection and recognition accuracy. It effectively improves the detection and recognition effect of the model for multi-label images.
Keywords: Multi-label image recognition ; semantic relevance ; graph convolution network ; attention mechanism ;Multi-scale dynamic graph convolutional network

1. INTRODUCTION

According to the amount of information contained in the image, the image can usually be divided into two categories : multi-label image and single-label image. The image in the real scene is usually a multi-label image. Different from single-label classification, multi-label image recognition needs to assign multiple labels to a single image. At present, many scholars use Bayesian network [1,2] and recurrent neural network (RNN) [3] to realize multi-label image recognition, but these methods have problems such as high computational cost, manual definition of adjacency matrix, and easy over-fitting. In recent years, graph convolutional network (GCN) [4] has achieved great success in modeling the relationship between vertices of a graph, which lays a solid foundation for multi-label recognition. Therefore, GCN has gradually entered the attention of scholars and become an option for multi-label graphics recognition.

In multi-label graph recognition, GCN is often used to focus on the label semantic relationship between classes or classes. By using the prior frequency of label co-occurrence in the target dataset, Chen et al. [5] constructed a complete graph to model the label correlation between each two categories, and created a Multi-Label Image Recognition with Graph Convolutional Networks (ML-GCN), which achieved an average accuracy of 83 % in the COCO dataset and achieved remarkable results. However, ML-GCN constructing a global graph for the entire data set may cause frequency deviation problems in most common data sets, resulting in poor results. In order to solve this problem, Ye et al. [6] created the ADD-GCN model for multi-label image recognition. Different from the previous graph-based methods, ADD-GCN models the semantic relationship of each input image by estimating the specific dynamic graph of the image. The dynamic graph created by ADD-GCN can capture the content-aware category relationship of each image, and can capture the category relationship of specific images in an adaptive way, which further enhances its representativeness and discriminative ability.

In order to further improve the performance of ADD-GCN, Reference [7] introduced Res2Net [8] into ADD-GCN for multi-label recognition of power equipment for the first time, and achieved 88.1 % on the self-built data set. However, the improved ADD-GCN cannot automatically learn and selectively focus on important information in the input, and it is difficult to summarize all label semantic relationships. With the introduction and iteration of various algorithms, it is possible to solve the above problems and further improve the performance of ADD-GCN. Reference [9] proposed a CBAM attention mechanism and combined CBAM with ResNet, which was 2 % higher than mAP using ResNet alone in CAM visualization results. Reference [10] added CBAM to ML-M-GAT model for multi-label tasks, and achieved good performance of 1.7 % and 2.1 % higher than mAP of the original ML-M-GAT model on VOC dataset and COCO dataset. In Reference [11], a new high-performance crowd counting method was proposed by combining CBAM with Res2Net. That is, CBAM can not only be simply combined with Res2Net, but also can be used for multi-label tasks. Reference [12] proposed an adaptive optimization algorithm Adam.Reference [13] applied this optimization algorithm to the GCN network. In the GCN, the gradient of Adam can be transmitted stably. In [14], a new activation function GELU is constructed by combining some properties of dropout [15],

International Journal of Science and Engineering Applications
Volume 13-Issue 10, 102 – 109, 2024, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1310.1022

zoneout [16] and RELUs [17]. In experiments, it is proved that the activation function GELU has better performance than ReLU and ELUs [18] in most scenarios. On the basis of using channel attention and SE-ResNeXt, BCEWithLogits-

Loss was used in the literature [19], and good research results were obtained, that is, BCEWithLogitsLoss would have better performance on channel attention and ResNet model. According to the above research, it can be seen that Res2Net, CBAM, GELU, Adam and BCEWithLogitsLoss are expected to be combined with ADD-GCN to form better algorithms, which is also the main research motivation of this paper.

The purpose of this paper is to improve the performance of ADD-GCN in multi-label image recognition. The proposed method takes ADD-GCN as the basic algorithm framework, and adjusts and improves its network structure accordingly, so as to improve the accuracy of the algorithm in multi-label image recognition. Experiments are carried out on VOC dataset and COCO dataset to verify the practicability and feasibility of the algorithm. Finally, ablation experiments were performed on the COCO dataset to verify the role of each module in ADD-GCN.

2. ADD-GCN model

ADD-GCN introduces a novel dynamic graph constructed based on content-aware category representation for multi-label image recognition. It first extracts the multi-scale features of the image through a convolutional neural network, and then decomposes the convolutional feature map into multiple content-aware category representations through SAM. Each category representation describes the content related to a specific label from the input feature map $X \in R^{H\times W\times D}$. SAM first calculates the activation map of a specific category $M = [m_1, m_2, \dots, m_c]$, and then uses them to convert the converted feature map into a category representation of the perceptible content $V = [v_1, v_2, \dots, v_c] \in R^{C\times D}$. Specifically, the representation of each class is expressed as a weighted sum of pairs, so that the resulting features related to its specific category can be selectively aggregated :

$$V_C = m_c^T X' = \sum_{i=1}^{H}\sum_{j=1}^{W} m_{i,j}^c x'_{i,j} \#(1)$$

Where $m_{i,j}^c$ and $x'_{i,j} \in R^{D'}$ are the weights of the first activation map and the feature vectors at the feature map (i, j), respectively.

These representations are input into D-GCN. Dynamic graph convolution propagates features through two joint graphs (static graph and dynamic graph). D-GCN takes the class representation of perceptible content as the input node features, and provides them to static GCN and dynamic GCN in order to generate discriminant vectors for multi-label classification. The single-layer static GCN is simply defined as $H = LReLU(A_s V W_s)$, where $H = [h_1, h_2, \dots, h_c] \in R^{C\times D_1}$, the activation function LReLU (·) is Leaky Relu. In the training process, the correlation matrix A_s and the state update weight W are randomly initialized by the gradient descent method. Since A_s is shared by all images, it is expected that A_s can capture the global rough classification dependency. Then the dynamic GCN is introduced to transform H, and the correlation matrix A_d is estimated adaptively according to the input feature H. The correlation matrix of static GCN is fixed, and all input samples are shared after training. The A_d of ADD-GCN is dynamically constructed according to the input features. Since each sample has a different A_d, the model improves its representativeness and reduces the over-fitting risk caused by the static graph. Formally, the output $Z \in R^{C\times D_2}$ of a dynamic GCN can be defined as :

$$Z = f(A_d H W_d), \text{where } A_d = \delta(W_A H') \quad (2)$$

Here, $f(\cdot)$ is the Leaky Relu activation function, $\delta(\cdot)$ is the Sigmoid activation function, $W_d \in R^{D1\times D_2}$ is the state update weight, $W_A \in R^{C\times 2D_1}$ is the weight of the convolution layer that constructs the dynamic correlation matrix A_d. $H' \in R^{C\times 2D_1}$ is obtained by concatenating H and its global representation $h_g \in R^{D_1}$. $h_g \in R^{D_1}$ is obtained by concatenating the global average pool and a layer of convolutional layer (the main purpose of this step is to fuse the vectors corresponding to each label to obtain the global feature vector, and then concatenate it to the feature vector of each original label, so that the network can better judge the correlation between labels based on visual features). Formally, H' is defined as :

$$H' = [(h_1; h_g), (h_2; h_g), \dots, (h_c, h_g)] \quad (3)$$

The dynamic graph is specific to each image and can capture content-related category dependencies. In general, D-GCN enhances the category representation of the perceived content from to through data-specific graphs and image-specific graphs.

3. Improvement of ADD-GCN model

On the basis of ADD-GCN, the proposed method adjusts and improves its network structure accordingly. The structure diagram is shown in Figure 1. The research and optimization are mainly on the multi-scale feature extraction module and the data processing transmission of SAM and D-GCN

3.1 Multi-scale Feature Extraction

Multi-scale feature extraction plays an important role in network model optimization. It can enhance the invariance and robustness of the model to objects and improve the adaptability of the model to images of different scales. Improving the multi-scale extraction ability of ADD-GCN can improve the model 's ability to understand and express images, thereby improving the performance of the model.

3.1.1 Res2net

ADD-GCN uses ResNet as the backbone to extract multi-scale information. In order to further obtain fine-grained feature information of multi-label images, we introduce Res2Net to extract multi-scale features of images. In the past, multi-scale feature extraction network structures usually used features of different resolutions to improve multi-scale capabilities. However, Res2Net constructs a residual connection similar to the hierarchical structure in a single residual block, represents multi-scale features at a finer-grained level, and increases the receptive field of each network layer. Reference [14] directly expounds the excellent characteristics of the module, which can replace ResNet in a complex network structure to achieve better results. In the Res2Net module, the input features are divided into **s** groups, which are denoted as X_i, $i \in 1,2,\dots,s$; the number of channels in each group of feature maps is 1 / **s** of the number of channels in the input feature map. Except for X_1,

International Journal of Science and Engineering Applications
Volume 13-Issue 10, 102 – 109, 2024, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1310.1022

each group of feature maps will go through a 3 × 3 convolution, and the convolution operation is recorded as $K_i()$. In addition to X_1 and X_2, the feature map X_i of the ith group is first added to the output of the previous group, and the $K_i()$ operation is performed on the added result.The above operation is shown in formula (4) :

$$y_i = \begin{cases} x_i & i = 1; \\ K_i(x_i) & i = 2; \\ K_i(x_i + y_{i-1}) & 2 < i \le s. \end{cases} \quad (4)$$

The output of this s group is spliced in the channel dimension, and then 1 × 1 convolution is performed for dimensionality reduction. The structure is shown in the blue part of the upper left corner of Fig.1. Obviously, the input of the convolution operation $K_i()$ in group i contains multiple sets of input features : $\{x_j, j \le i\}$. Due to this split hybrid connection structure, the output of the Res2Net module contains a combination of different receptive field sizes, which is conducive to extracting global and local information. In addition, in order to reduce the number of parameters, the convolution of the first segmentation part X_1 is omitted, which can also be regarded as a form of feature reuse.

Res2Net utilizes multi-scales at a finer-grained level, which is orthogonal to existing methods using hierarchical operations. Therefore, the Res2Net module can not only seamlessly replace the original ResNet, but also integrate with multiple structures to form a stronger backbone network.

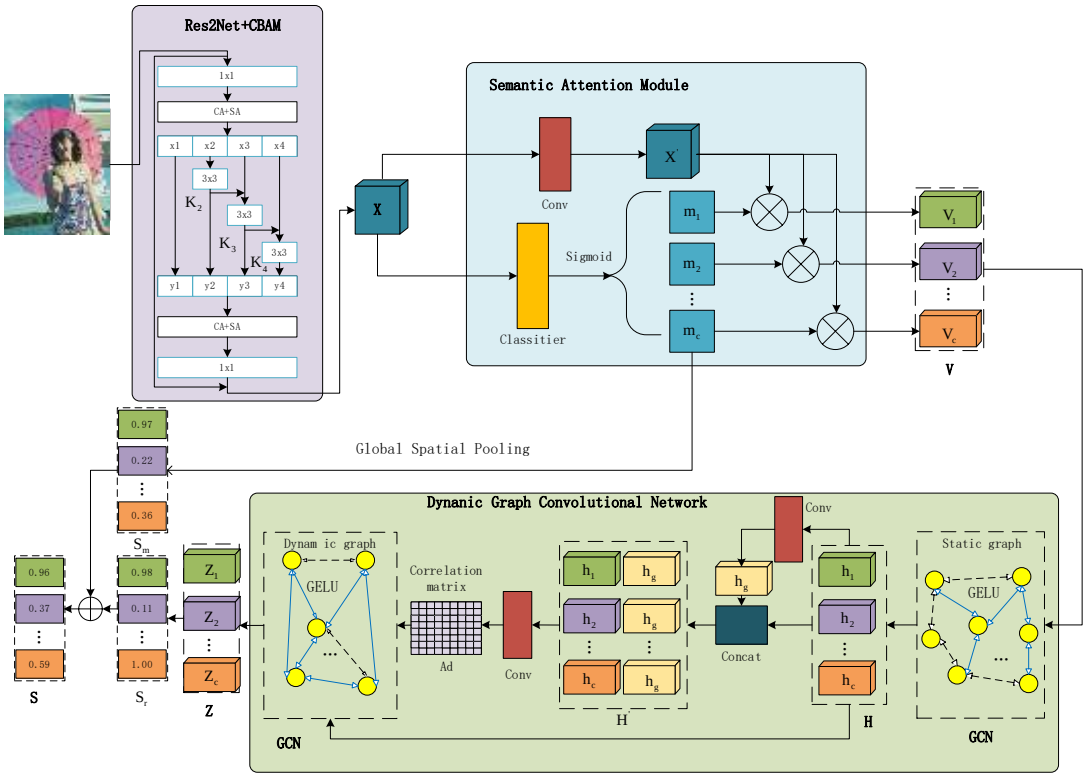


Fig.1 Improved network structure of ADD-GCN mode

3.1.2 CBAM

Inspired by References [16] and [17], the former uses CBAM for multi-label tasks, and the latter combines CBAM with Res2Net, both of which have achieved good results. Considering the difficulty of classification caused by the need to deal with a large number of labels in multi-label recognition, this paper introduces the CBAM module and adds it to Res2Net. Unlike literature [17], which uses CBAM for Res2Net output, this paper uses CBAM in the channel dimension of Res2Net to minimize the amount of calculation while improving the representation ability of Res2Net. In this paper, Res2Net is combined with CBAM module to strengthen the feature learning between different levels and different channels of Res2Net, which greatly saves parameters and computing power. In multi-label recognition tasks, this combination can make the network pay more attention to important feature channels and spatial locations, thereby improving the representation ability of the network. At the same time, due to the characteristics of CBAM, this combination can also reduce the sensitivity of the model to noise and unnecessary information, thereby improving the clarity of the feature map and making the model more lightweight and efficient.

3.2 Data processing and transmission

ADD-GCN is represented by SAM decomposition-aware categories, and D-GCN is introduced to adaptively transform their coherent correlation for multi-label recognition. The multi-label problem usually involves complex interactions and dependencies, so it is particularly important to improve the

International Journal of Science and Engineering Applications
Volume 13-Issue 10, 102 – 109, 2024, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1310.1022

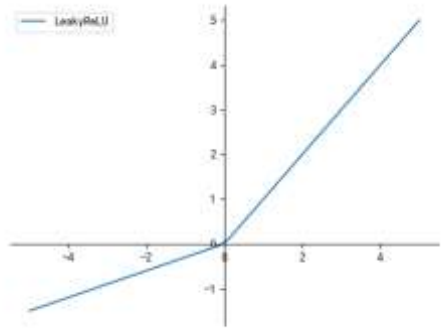
processing and transmission capabilities of data in SAM and D-GCN.

3.2.1 GELU activation function

SAM and D-GCN use the Leaky ReLU activation function when passing class representation, which is in the form of :

$$LReLU(x) = \begin{cases} x & x > 0 \\ ax & x \le 0 \end{cases} \tag{5}$$

The figure is as follows.



It is not difficult to see that when the function is differential, both parts are linear. When the neural network is only composed of linear operations, even if there are multiple hidden layers, its expression ability is equivalent to that of single-layer neural network. At the same time, the value is a hyperparameter and needs to be manually set.

Reference [19] proposed a high-performance activation function GELU, which is an approximation of the cumulative distribution function based on Gaussian distribution. This approximation property helps to better model the distribution of data. The form of GELU is :

$$xP(X \le x) = x \int_{-\infty}^{x} \frac{e^{-\frac{(X-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dX \tag{6}$$

The calculation results are about :

$$GELU(x) = 0.5x\left(1 + \tanh\left(\sqrt{\frac{2}{\pi}} \cdot (x + 0.044715x^3)\right)\right) \tag{7}$$

The GELU activation function and the Sigmoid function have approximate properties, so they can also be expressed as :

$$x\sigma(1.702x) \tag{8}$$

Where, $\sigma(x_n)$ is the Sigmoid function.

The GELU function image is shown in Figure 4 :

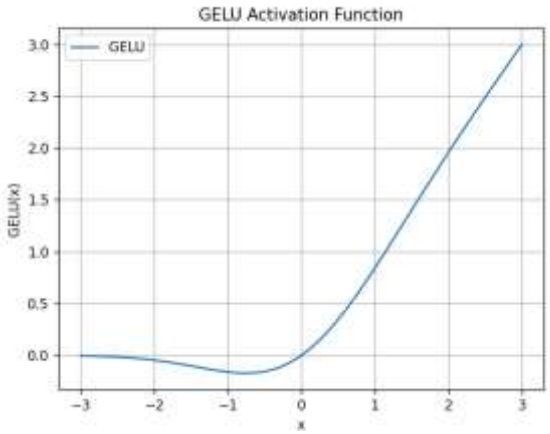


Fig.4 Graph of the GELU Activation function

Comparing the function images of the two activation functions, it can be seen that GELU is an activation function with nonlinear properties. This property can transform continuous input into discrete output, endow the network with nonlinear modeling ability, so that the deep neural network can learn and represent the complex nonlinear function mapping between input and output, so as to solve the problem that the linear model cannot solve. At the same time, GELU is a smooth activation function, which contributes to the stable propagation of gradients, promotes the neural network to better fit these complex relationships, and makes the training process more stable. In the multi-label image recognition task, these properties of GELU ensure that the model can adapt to multiple labels, not only can reduce the death of neural units, but also improve the numerical stability, especially when the output of the model is near the extreme value. The smoothness and nonlinearity of GELU make it greatly improve the numerical stability.

3.2.2 Optimizer and loss function

ADD-GCN uses a combination of SGD optimizer and MultiLabelSoft MarginLoss loss function. This combination is prone to problems of sensitivity to learning rate and numerical instability. In order to solve these problems and further improve the processing and transmission capabilities of data in SAM and D-GCN, this paper introduces a combination of Adam optimizer and BCEWithLogitsLoss loss function.

The SGD of ADD-GCN selects a mini-batch each time, instead of all samples, and uses gradient descent to update the model parameters. First calculate the gradient of the current batch, and then update the parameters :

$$\theta_{t+1} = \theta_t - \alpha * g_t \tag{10}$$

It first calculates the gradient of the current batch, and then updates the parameters, where θ_{t+1} is the parameter at the next moment, θ_t is the parameter at the current moment, α is the learning rate, and g_t is the gradient at the current moment. From the formula, it is not difficult to see that SGD is sensitive to the learning rate, and the update direction depends entirely on the current sample. There is a SGD optimizer that solves the problem of random small batch samples, but there are still problems such as adaptive learning rate and easy to get stuck in small gradient points. At the same time, due to the small amount of data used in each parameter update, the oscillation amplitude of the gradient update is large, which is easily affected by outliers, and there will be increased fluctuations near the optimal solution.

The Adam optimizer is not sensitive to the set initial learning rate, and can be optimized to a better parameter in a wide range of intervals to avoid the above problems. The Adam update rules are as follows :

1) calculate the gradient of the current batch

2) Update the first moment estimation (mean) :

$$m_t = \beta_1 * m_{t-1} + (1 - \beta_1) * g_t \quad (10)$$

Among them, m_t is the first-order moment estimation at the current moment, and β_1 is the exponential decay rate of the first-order moment estimation, which is usually close to 1, such as 0.9.

3) Uncentered variance is updated :

$$v_t = \beta_2 * v_{t-1} + (1 - \beta_2) * g_t^2 \quad (11)$$

Among them, v_t is the second-order moment estimation of the current moment, β_2 is the exponential decay rate of the second-order moment estimation, usually also close to 1, such as 0.999.

4) Bias correction : Bias correction is needed because it may be very small at the beginning, especially in the early stage of training. This is to reduce the deviation of the optimization process.

$$m_{t_{hat}} = \frac{m_t}{1 - \beta_1^t} \#(12)$$

$$v_{t_{hat}} = \frac{v_t}{1 - \beta_2^t} \#(13)$$

Where, t represents the number of iterations at the current moment.

5) Updating parameters : Using the corrected first-order moment estimation and second-order moment estimation to update the model parameters :

$$\theta_{t+1} = \theta_t - \alpha * \frac{m_{t_{hat}}}{sqrt(v_{t_{hat}}) + \varepsilon} \quad (14)$$

θ_{t+1}, which is the parameter at the next moment, is the learning rate, which is a small constant, usually 1e-7, used to avoid zero denominator.

It can be seen from the expression that Adam is an updated step size calculation, which can be adaptively adjusted from the two angles of gradient mean and gradient square, rather than directly determined by the current gradient. It combines the advantages of AdaGrad [20] and RMSProp [21]. The first-order moment estimation and second-order moment estimation of the gradient are considered comprehensively, and the update step size is calculated. At the same time, the update of the parameters is not affected by the scaling transformation of the gradient, which is suitable for large-scale data and parameter scenarios. In the multi-label recognition task, it can accelerate the training of the neural network and improve the performance of the model through adaptive learning rate, momentum and bias correction techniques, so as to better fit the training data and obtain better results in various deep learning tasks.

Literature [18] combined channel attention with SE-ResNeXt and BCEWithLogitsLoss, and obtained good research results. Because this paper uses CBAM attention mechanism and Res2Net, inspired by this, this paper also introduces BCEWithLogitsLoss. BCEWithLogitsLoss is usually used with the Sigmoid activation function to measure the cross-entropy loss between the output of the deep learning model and the actual label, which helps to guide the model training to achieve accurate classification. The GELU used in this paper is an approximation of the Sigmoid function, which can inherit this advantage and make the model converge to the optimal parameter configuration faster. Compared with ADD-GCN, MultiLabelSoftMarginLoss of ADD-GCN does not have this characteristic. When it is far away from the extreme value, it is easy to be numerically unstable, which affects the data transmission of SAM and D-GCN.

The form of BCEWithLogitsLoss is as follows :

Assuming that there is a batch, each batch predicts a label, then Loss is :

$$Loss = -[y_n \cdot \log(\sigma(x_n)) + (1 - y_n) \cdot \log(1 - \sigma(x_n))] \#(15)$$

Here $\sigma(x_n)$ is the Sigmoid function, x can be mapped to the interval of (0,1).

The function image of BCEWithLogitsLoss is :

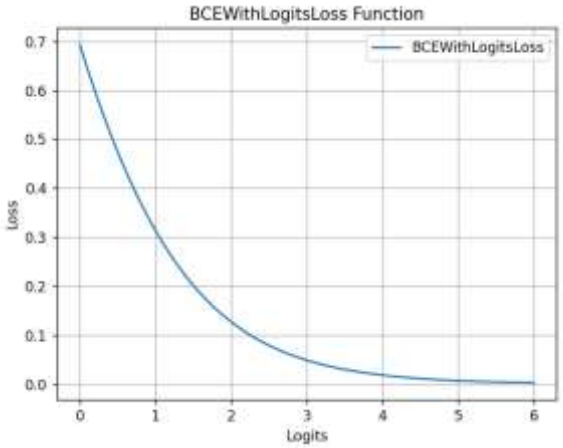


Fig.5 Graph of the BCEWithLogitsLoss Function

4. Experimental Results and Analysis

4.1 Data Set

In order to prove the effectiveness of the improved model of the algorithm, we conducted experiments on the data set MS-COCO [22] and the data set Pascal VOC 2007 [23] to evaluate the proposed method.

1) MS-COCO consists of a training set of 82081 images and a validation set of 40137 images. The dataset covers 80 common object categories, with approximately 2.9 object labels per image.

2) PASCAL VOC 2007 includes 20 different types of objects. The data set contains a total of 9963 images : 5011 training sets and 4952 test sets.

4.2 Experimental Evaluation Index

In this paper, the mAP, CR, CF1, OP, OR and OF1 of ' ALL ' and ' Top-3 ' will be used to evaluate the performance of the model.

4.3 experiment setting

In order to compare with the current multi-label recognition network, this paper uses the same parameters and experimental steps as them, and uses NVIDIA RTX A6000 graphics card to train 50 rounds in MS-COCO data set and Pascal VOC 2007 data set respectively. At the same time, the environment used in this experimental training is the Ubuntu20.04 operating system. The

International Journal of Science and Engineering Applications
Volume 13-Issue 10, 102 – 109, 2024, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1310.1022

Python language is used to build the operating environment based on the Pytorch framework, and the CUDA 11.3 acceleration toolbox is used.

4.4 Experimental comparison and result analysis

According to the experimental settings, in order to prove the effectiveness of the proposed algorithm, this paper compares it with other algorithms, especially the original ADD-GCN.

The comparison results on the MS-COCO dataset are shown in Table 1. The method in this paper has achieved the most advanced results among all evaluation indicators.

The comparison effect of Pascal VOC 2007 data set is shown in Table 2, and the evaluation index is the average value of each average accuracy and average accuracy. The method of this paper has achieved advanced results in all evaluation indicators.

Compared with traditional network architectures (RNN, ResNet, SSGRL), they cannot make full use of the spatial structure information in image data. The GCN architecture can capture the spatial relationship and local features between pixels, and learn the relationship between nodes in the graph structure, so as to better capture the relationship between various parts of the image and achieve better results.

Since the algorithm in this paper introduces the Adam optimizer, it can perform better in identifying large data sets. In the COCO dataset, compared with the traditional ADD-GCN model, mAP, CF1, and OF1 reached an increase of 2.9 %, 2.5 %, and 1.9 %, respectively.

4.5 ablation experiment

In this paper, ablation experiments are carried out on the MS-COCO data set to prove that the proposed improvement of each

module is effective. In this paper, the improved modules will be added in turn and compared with the original ADD-GCN method. Table 3 shows the combination of various modules and the experimental results.

It can be seen from Table 3 that the combination of Res2Net + CBAM has significantly improved on mAP, CF1, and OF1. At the same time, the addition of CBAM can help the model to better focus on important features, reduce complexity, and thus reduce Batchtime. The GELU activation function can effectively improve the training accuracy and further shorten the training time. Compared with the original loss function and optimizer, the BCEWithLogitsLoss + Adam combination has a great improvement in LOSS and training process. Under the condition of opening pre-training, the best effect can be achieved basically in the eighth epoch.

Table 1 The proposed method is compared with other multi-label recognition algorithms on the COCO dataset

Method	ALL							Top-3					
	mAP	CP	CR	CF1	OP	OR	OF1	CP	CR	CF1	OP	OR	OF1
ResNet-101	79.7	82.7	67.4	74.3	86.4	71.8	78.4	85.9	60.5	71	90.2	64.2	75
DecoupleNet	82.2	83.1	71.6	76.3	84.7	74.8	79.5	–	–	–	–	–	–
ML-GCN	83	85.1	72	78	85.8	75.4	80.3	89.2	64.1	74.6	90.5	66.5	76.7
SSGRL	83.8	**89.9**	68.5	76.8	**91.3**	70.8	79.7	**91.9**	62.5	72.7	**93.8**	64.1	76.2
ADD-GCN	84.2	85.1	73.6	78.9	86.2	76.6	81.1	89.2	65.2	**78.9**	90.9	67.1	77.2
Our	**87.1**	86.8	**76.6**	**81.4**	87.1	**79.2**	**83**	90.4	**67**	76.9	91.4	**68.9**	**78.6**

International Journal of Science and Engineering Applications
Volume 13-Issue 10, 102 – 109, 2024, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1310.1022

Table 2 Comparison of the proposed method with other multi-label recognition algorithms on VOC datasets

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	hors e	mbik e	perso n	plant	shee p	sofa	trai n	tv	mAP
CNN-RNN	96.7	83.1	94.2	92.8	61.2	82.1	89.1	94.2	64.2	83.6	70.0	92.4	91.7	84.2	93.7	59.8	93.2	75.3	99.7	78.6	84.0
ResNet-101	99.1	97.3	96.2	94.7	68.3	92.9	95.9	94.6	77.9	89.9	85.1	94.7	96.8	94.3	98.1	80.8	93.1	79.1	98.2	91.1	90.8
ML-GCN	99.5	98.5	**98.6**	98.1	80.8	94.6	97.2	98.2	82.3	95.7	86.4	98.2	98.4	96.7	99.0	84.7	96.7	84.3	98.9	93.7	94.0
SSGRL	99.7	98.4	98.0	97.6	85.7	96.2	98.2	**98.8**	82.0	98.1	89.7	**98.8**	98.7	97.0	99.0	86.9	98.1	85.8	99.0	93.7	95.0
ADD-GCN	**99.8**	98.6	98.1	98.3	85.8	97.2	98.3	98.1	83.1	**98.3**	88.2	98.6	98.6	**97.4**	99.0	**88.3**	**98.7**	87.1	99.2	94	95.24
Our	**99.8**	**98.7**	98.3	**99**	**86.7**	**98.1**	**98.5**	98.3	**85.8**	**98.3**	**88.9**	**98.8**	**99**	**97.4**	**99.2**	**88.3**	**98.7**	**90.7**	**99.5**	**96.1**	**95.6**

Table 3 The performance of the method in this paper on each module on the VOC dataset

	Network					Attribute types					
	Res2Net	CBAM	GELU	BCEWithLogitsLoss	Adam	mAP	CF1	OF1	Batch_time	Average Loss	epochs
a						84.2	78.9	81.1	0.51	0.08	/
b	√					85.8	80.1	82	0.52	0.05	/
c	√	√				86.7	81.3	82.9	0.49	0.05	13
d	√	√	√			86.9	81.9	83	0.5	0.05	11
e	√	√	√	√	√	87.1	82.1	83	0.46	0.03	8

5. Tag

ADD-GCN has good performance and accuracy in dealing with multi-label recognition tasks, but it cannot automatically learn and selectively focus on important information in the input. When transmitting data, it is prone to instability and difficult to summarize all label semantic relationships. In order to solve the above problems, this paper proposes an improved algorithm based on ADD-GCN model. Compared with the ADD-GCN model, the model updates the multi-scale feature extraction module and integrates the convolutional attention mechanism module to automatically focus on important information. The Gaussian error function is integrated to complete the mapping of neurons, and the adaptive moment estimation and binary cross entropy loss are introduced to make the data propagate stably. Finally, the performance advantages of the proposed method are verified by experiments on the MS-COCO dataset and Pascal VOC 2007 dataset. According to the experimental results, the proposed method is superior to the original ADD-GCN and other multi-label recognition algorithms in the values of OF1, CF1 and mAP, which significantly improves the performance of multi-label recognition tasks.Future research directions can further explore how to improve and optimize multi-label recognition technology. For example, it can further improve the performance of the algorithm in this paper, add the latest modules or perform lightweight operations on the algorithm in this paper. In addition, deep learning and computer graphics methods can be combined to improve the connection between semantic labels, thereby improving the accuracy of multi-label recognition.

6. REFERENCES

[1] Goan E, Fookes C. Bayesian neural networks: An introduction and survey[J]. Case Studies in Applied Bayesian Data Science: CIRM Jean-Morlet Chair, Fall 2018, 2020: 45-87.

[2] Cao Jie.Study on learning and application of Bayesian network structure [J].PhD thesis. Hefei : University of Science and Technology of China, 2017.

[3] Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization[J]. arXiv preprint arXiv:1409.2329, 2014.

[4] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv:1609.02907, 2016.

[5] Chen Z M, Wei X S, Wang P, et al. Multi-label image recognition with graph convolutional networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5177-5186.

[6] Ye J, He J, Peng X, et al. Attention-driven dynamic graph convolutional network for multi-label image recognition[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16. Springer International Publishing, 2020: 649-665.

[7] Yan Y, Han Y, Qi D, et al. Multi-label image recognition for electric power equipment inspection based on multi-scale dynamic graph convolution network[J]. Energy Reports, 2023, 9: 1928-1937.

[8] Gao S H, Cheng M M, Zhao K, et al. Res2net: A new multi-scale backbone architecture[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 43(2): 652-662.

[9] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[10] Shi Xiuyun, Li Shunyong, Han Xiang.Multi-label image classification based on multi-head graph attention network and graph model 1 [J].2023.

[11] Chen Jiangchuan, Wu Yuntao, Kong Quan. Crowd counting algorithm based on CBAM-Res2Net [J].Journal of Wuhan University of Engineering, 2022,44 (6) : 664-669.

[12] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[13] Kipf T N , Welling M .Semi-Supervised Classification with Graph Convolutional Networks[J]. 2016.DOI:10.48550/arXiv.1609.02907.

[14] Hendrycks D , Gimpel K .Gaussian Error Linear Units (GELUs)[J]. 2016.DOI:10.48550/arXiv.1606.08415.

[15] Related S O R , Ofdevil R , Related S O R ,et al.Dropout: A Simple Way to Prevent Neural Networks from Overfitting [91][J].[2023-12-08].

[16] Krueger D , Maharaj T ,Kramár, János,et al.Zoneout: Regularizing RNNs by Randomly Preserving Hidden Activations[J]. 2016.DOI:10.13140/RG.2.1.1027.2889.

[17] Agarap A F M .Deep Learning using Rectified Linear Units (ReLU)[J]. 2018.DOI:10.48550/arXiv.1803.08375.

[18] Djork-Arné Clevert, Unterthiner T , Hochreiter S .Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)[J].arXiv e-prints, 2015.

[19] Li Jie. Research on classification of cerebral hemorrhage based on deep learning [D].Hangzhou University of Electronic Science and Technology, 2021.

[20] Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research, 12(Jul), 2121-2159.

[21] Ruder S .An overview of gradient descent optimization algorithms[J]. 2016.DOI:10.48550/arXiv.1609.04747.

[22] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014: 740-755.

[23] Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (voc) challenge[J]. International journal of computer vision, 2010, 88: 303-338.