Research on Capsule Network Based on Attention Mechanism

Yan Jiao
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: jiaoyan@st.xatu.edu.cn

Hexin Xu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 809015737@qq.com;

Li Zhao
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 332099732@qq.com;

***Abstract*—The capsule network has good spatial recognition and has good accuracy in classification and recognition tasks. However, because of the dynamic routing algorithm in the capsule network, the training speed of the capsule network is slow. In order to make better use of the capsule network, reduce For its training cost, this paper proposes a capsule network based on the attention mechanism, and adds the CBAM attention module to the original capsule network to improve the network's ability to extract information in the feature map channel and information in the feature map space, and improve the network's learning ability, To reduce the number of network training, thereby reducing the cost of training. This paper conducts experiments based on the original neural network to verify the effectiveness and feasibility of adding the CBAM module to the capsule network. The final result is that the CBAM module can speed up the convergence speed of the capsule network by 50%.**

Keywords-Component; Capsule Net; CBAM; Attention

I. INTRODUCTION

The concept of capsule network started in 2011 and was proposed by Professor Hinton [1], and then in 2017, the first generation of capsule network (CapsNet) was realized. The capsule network includes three layers. The first two layers of the network use the traditional convolutional layer, and the capsule layer is added after the convolutional layer. And pioneered the use of dynamic routing algorithms in the network as the connection between the primary caps and digit caps to achieve information transfer. For the general neural network structure at this stage, the output of each layer of neurons is a scalar result, and the scalar data contains limited information. Correspondingly, if the output is a vector, the output generated can more accurately represent the posture, etc. Related Information. The capsule network implements this theory, and the capsule is equivariant. Different inputs of the capsule network will have different outputs. When the same object changes, such as movement, rotation, size, etc., the capsule will output a module with the same length but different internal data Vector. The length of the vector output by the capsule network is used to indicate the possibility of the existence of the object, and the internal elements of the vector output by the capsule network represent a certain feature of the object, such as the writing angle of the number and the thickness of the stroke. The capsule network uses a dynamic routing algorithm to transfer between capsules, which not only achieves the purpose of reasonably transmitting the information contained in the upper layer of the network to the neurons of the next

DOI: 10.21307/ijanmc-2021-011

layer, but also successfully makes the vector form of neurons in the neural network. To be realized.

The advantages of the capsule network are as follows:

(1) A key feature of the capsule network is to learn the posture information of objects in the network. Therefore, when detecting an object, there is no need for multiple training at different angles. The capsule network will autonomously learn the position relationship of the object and show it in the output vector. In a conventional convolutional neural network, there are usually multiple convolutional layers, and the operation of these convolutional layers usually loses some information, such as the angle and orientation of the target object. If you just want to cluster the overall image, the loss of position and posture data will not affect the results, but these missing data are decisive for tasks such as image segmentation (which requires more precise position and posture). of.

(2) The routing protocol algorithm has a good effect on dealing with complex scenes. The routing tree can also map the hierarchy of objects, assigning each part to a capsule. Moreover, the capsule network has strong robustness to changes in angle and position. And because its output is a vector, its result is better interpretable.

After the successful construction of the capsule network, scholars in various fields have a keen interest and have tried a variety of different tasks. Among them, Lu Chunyan [6] and others applied the capsule network to the image generation task. By improving the capsule network, they proposed a method of parameter sharing, which further reduced the amount of parameters of the capsule layer; Ren Qiang [7] and others proposed innovatively A variable-dimensional capsule network is used. Because the number of dimensions of the capsule is closely related to the amount of information contained in the capsule, the amount of data contained in the capsule can be increased through variable dimensions. After understanding the characteristics of the capsule and the structure of the capsule network itself, this paper adds an attention mechanism to the capsule network and conducts analysis, research and performance evaluation on the modified network.

II. Capsule Network Structure and Principle

The core idea of the capsule network is inverse rendering, which comes from computer graphics. During the rendering process, we need to provide geometric information, such as telling the computer where to draw the object, such as the scale, angle, and spatial information of the object. Inverse rendering is based on the image (rendering result) to infer the information of the object, including spatial geometric information. The capsule network needs to learn how to render the image in reverse—by observing the image, predict the instance parameters of the image. Vector encoding form of capsule network features

The capsule network encodes the spatial information and the existence probability of the object at the same time, and encodes it in the Capsule vector.

Capsule vector: The modulus of the vector indicates the probability of the feature's existence; the direction of the vector indicates the posture information of the feature; moving the feature will change the Capsule vector without affecting the probability of the feature's existence.

The original capsule network for recognizing digital images consists of three layers, the initial convolutional layer, the primary capsule layer, and the digital capsule layer. The initial convolutional layer is a conventional convolutional layer; the primary capsule layer is a convolutional layer using the Squash activation function, and its output is reshaped to [Batch Size, num caps, dim caps]; the third layer is a digital capsule layer, routing The algorithm works in the digital capsule layer, and transmits the output of the primary capsule to the digital capsule layer in an optimal way. Finally, the digital capsule obtains a number between 0-1 through the modulus, and this number is the digital capsule the credibility of the corresponding number.

The network is shown in Figure 1.The first layer is the input layer. The data set used in this paper is MNIST, so all images are 28*28 singlechannel grayscale images (the value of each pixel is between 0-255), so the input layer It is 28*28.The second layer is to perform convolution with a step length of 1 through 256 convolution kernels with a size of 9*9.The formula for calculating the output size of the convolutional layer: The third layer is convolution with a step size of 2 through 256 9*9 convolution kernels, and the output is calculated by formulas as 6*6*256 data. Some data can be collected through two consecutive convolution layers.

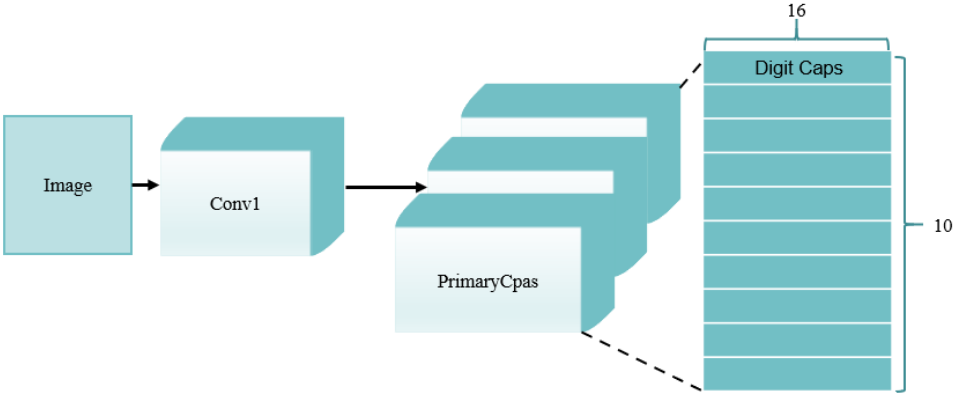

Figure 1. The original capsule network

The fourth layer is the primary capsule layer. This layer rearranges the data of the second convolutional layer and transfers data with the digital capsules in the lower layer through a dynamic routing algorithm. The fifth layer is the output layer. First, calculate the length of the vector output by each digital capsule, and find the vector with the largest modulus. The number corresponding to vector that largest modulus is the number corresponding to the output of the input capsule network with this picture.

The capsule network is a machine learning neural network developed to solve the problem that the convolutional neural network will lose the location characteristics of the data. Capsule network is mainly composed of the convolutional layer and primary capsule layer and the digital capsule layer, and a dynamic routing algorithm is used between the primary capsule layer and the digital capsule layer. The dynamic routing algorithm is explained below.

```
**Procedure** Routing algorithm
1: procedure ROUTING ($\hat{\boldsymbol{u}}_{i|j}, i, l$)
2:    for all capsule $i$ in layer $l$ and capsule $j$ in layer ($l+1$): $b_{ij} \leftarrow 0$
3:    for $r$ iterations do
4:       for all capsule $i$ in layer $l: c_i \leftarrow \text{softmax}(b_i)$
5:       for all capsule $j$ in layer $(l+1): s_j \leftarrow \sum_i c_{ij}\hat{\boldsymbol{u}}_{i|j}$
6:       for all capsule $j$ in layer $(l+1): v_j \leftarrow \text{squash}(s_j)$
7:       for all capsule $i$ in layer $l$ and capsule $j$ in layer $(l+1)$
8:    return $v_j$
```

The dynamic routing algorithm is used to update the capsules in the front and back layers of the capsule network, that is, the capsule vector in the back layer and the capsule connection coefficient between the previous layer are the same, and the capsule vector in the previous layer is multiplied by the weight matrix. A prediction vector is obtained, and if there is a large scalar product between this prediction vector and the vector calculated later, the coupling coefficient of the capsule vector between the two layers before and after will increase. This consistent routing method is more effective than the maximum pooling used in traditional convolutional neural networks

Through (1) function, the length of an infinite vector can be compressed to close to 1, and a short vector can be compressed to close to 0, without affecting the direction of the vector. This function can make the vector easier to propagate and prevent excessively long noise vectors Affect the output of the entire network.

$$v_j = \frac{\left\|s_j\right\|^2}{1+\left\|s_j\right\|^2}\frac{s_j}{\left\|s_j\right\|} \quad (1)$$

Capsule network has a unique loss function, because the output of the capsule network is a vector:

Among them, if there is a number K , then T_k =1, and $m^+ = 0.9$, $m^- = 0.1$. We use $\lambda = 0.5$. The sum of the losses of all digital capsules is the total loss of this training.

It can be seen from formula (2) that $\max\left(0,m^+ - V_k\right)^2$ is the probability of the existence of the number K , that is, when the length of V_k is greater than 0.9, the loss function of this part is 0, when it is less than 0.9 When this part of the loss function is not 0. When the number K does not exist, you need to use the second part to calculate $\max\left(0,V_k - m^-\right)^2$, when the length of V_k is less than 0.1, the loss function of this part is 0, when it is greater than 0.1 When this part of the loss function is not 0.

$$L_k = T_k \max\left(0, m^+ - V_k\right)^2 + \lambda\left(1-T_k\right)\max\left(0, V_k - m^-\right)^2 \quad (2)$$

In the code, we calculate this function as a matrix operation. First calculate the length of the digital capsule output (16*10 matrix) as V_k (1*10 matrix), when the correct number is 1, T_k is in the form of one hot encoding (1*10 vector [0,1, 0,0,0,0,0,0,0,0]), that is, the position values except 1 are all 0. Bring the two matrices into the loss function, and the final L_k is also a 1*10 The total loss value can be obtained by summing L_k in the 0 dimension.

III. Attention mechanism CBAM

A. CBAM

Convolutional Block Attention Module (CBAM) represents the attention mechanism module of the convolution module. CBAM is an attention mechanism module that combines spatial and channel. Compared with Senet is attention mechanism that only focuses on channels, CBAM also uses an attention mechanism to enhance the spatial dimension of the feature map, so CBAM can achieve better results than SEnet.

CBAM is a simple and effective attention module for feedforward convolutional neural networks. Given an intermediate feature map, the CBAM module sequentially infers the attention map along two independent dimensions (channel and space), and then multiplies the attention map by the input feature map for adaptive feature modification. Since CBAM is a lightweight general-purpose module, CBAM can be seamlessly integrated into any CNN architecture with negligible overhead, and CBAM can be trained end-to-end together with the basic CNN.

These characteristics of CBAM make it easy to be integrated into each network, attach attention mechanism to each special network, and improve the performance of the network.

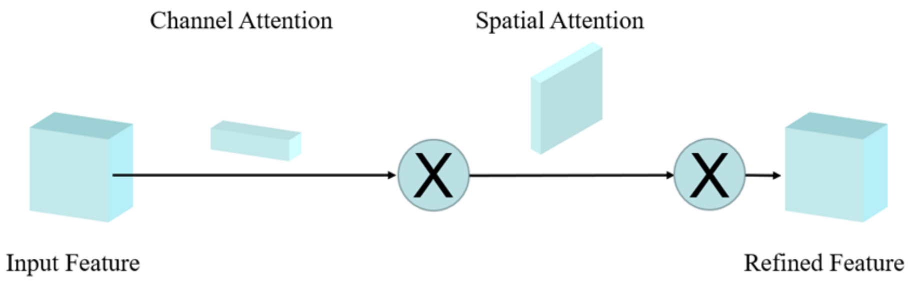

Figure 2. CBAM Module

Based on the traditional convolutional neural network module, channel and spatial attention mechanism modules are added after convolution. The CBAM module is shown in Figure 2. First, the convolution output feature map of the convolutional layer is added to the feature map with channel attention characteristics through the Channel Attention module, and then the spatial attention feature is added to the feature map through the Spatial Attention module, and finally Refined is obtained Feature.

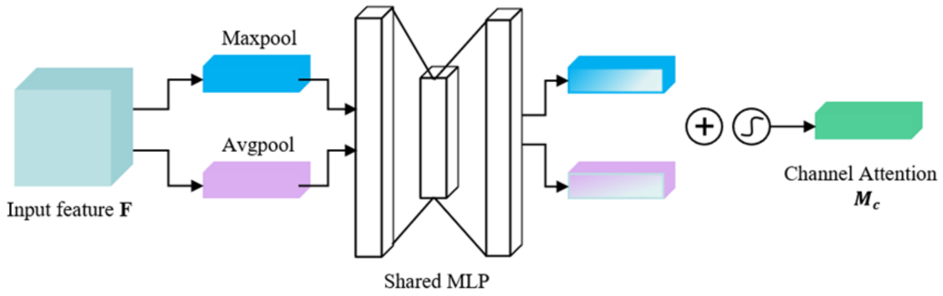

Figure 3. Channel Attention Module

B. *Channel Attention Module*

The channel attention module is shown in Figure 3. After processing by the channel attention module, a feature map that compares the importance of all channels will be obtained. All channels in the feature map are multiplied by the corresponding channels. As a result, related channels will be multiplied by a larger value to get a reward, and irrelevant channels will be multiplied by a smaller value and be punished.

The formula for channel attention is(3) and (4) function.

$$M_c(F) = \sigma(\mathrm{MLP}(\ \mathrm{AvgPool}\ (F)) + \mathrm{MLP}(\mathrm{MaxPool}(F))) \tag{3}$$

$$M_c(F) = \sigma(W_1(W_0(F^c_{\mathrm{avg}})) + W_1(W_0(F^c_{\mathrm{max}}))) \tag{4}$$

$$W_0 \in R^{c/r\times c}, W_1 \in r^{c/r\times c}$$

Is the channel attention probability distribution of the final output, is the sigmoid operation, and is the shared network. The operation makes a probability distribution.

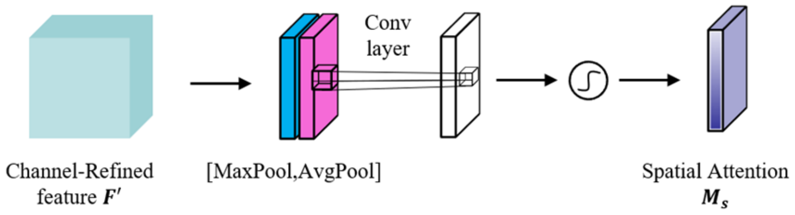

Figure 4. Spatial Atttion Module

C. Spatial Atttion Module

The spatial attention module is shown in Figure 4. After processing by the spatial attention module, a feature map will be obtained. This feature map is processed by the sigmoid function and contains the probability distribution of the information contained in each region. This probability distribution is compared with the original feature By multiplying the graphs, you can increase the value of the relevant feature area and decrease the value of the irrelevant area.

Through this method, the key areas in the image can be numerically enlarged, so that the influence of this area on the output of the network is enlarged.

The formula for spatial attention is as follows:

$$M_c(F) = \sigma(f^{7\times 7}([AvgPool(F), MaxPool(F)])) \quad (5)$$

$$M_c(F) = \sigma\left(f^{7\times 7}\left[F^s_{\text{avg}}, F^s_{\max}\right]\right) \quad (6)$$

Among them is the spatial attention probability distribution of the final output, and 7*7 represents the size of the convolution kernel. Experiments show that the 7*7 convolution kernel has a larger amount of parameters, but its effect is very better. M_c presents a probability distribution,Because σ is a sigmoid function.

IV. Capsule Network Based on Attention Mechanism

The capsule network based on the attention mechanism is shown in Figure 5. After the first layer of the convolutional layer, a layer of CBAM module is added to enhance the channel and space attention of the feature map output by the convolutional layer. The main function is to extract the features in the feature layer obtained by the first layer of convolution, and transfer many features to the digital capsule of the next layer through the routing algorithm. The CBAM module makes it easier to obtain features that are beneficial to the results in PrimaryCaps, and then transfer the enhanced beneficial features to the next layer of digital capsules through the routing algorithm, so that the network has a better result. After adding the CBAM module, the network has the ability to learn the importance of feature channels and the importance of feature spaces, and the overall learning performance of the network has been improved. Therefore, the CBAM module can greatly improve the convergence speed of the network with a small training cost.

The attention mechanism can better help the capsule network to collect information more accurately, and to recognize the required patterns more quickly, and this method is in line with biology. When humans focus on observing an object, they will naturally adjust the interpupillary distance. Then pay more attention to the object under observation and ignore the objects around the observation object. This can bring better efficiency, reduce the waste of information resources, and improve efficiency. This is also the optimal choice for biological evolution. Adjusting the interpupillary distance is similar to finding the best it conforms to the interpupillary distance corresponding to the observed object distance, and the CBAM attention mechanism uses the micro network inside the module to calculate the channel that the network pays attention to. Therefore, the attention mechanism is used to try to solve the problem of the capsule network with large parameters and slow convergence.

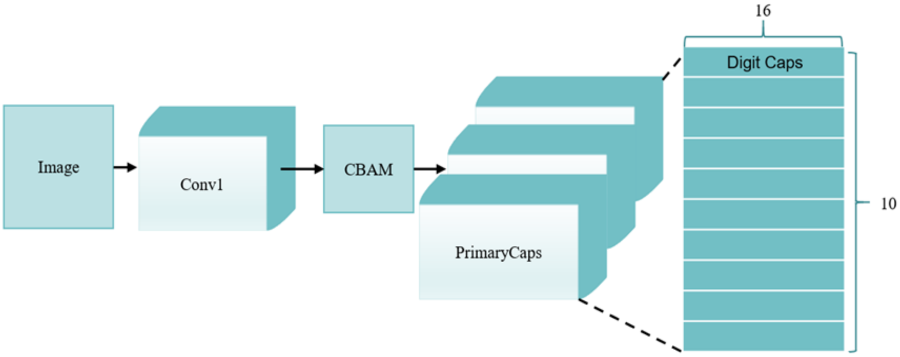

Figure 5. Capsule network based on attention mechanism

V. Experiment and Result Analysis

The Python environment used in the experimental environment used in this article is version 3.8, Pytorch uses version 1.8, the GPU used is Titan V, the optimizer used is Adam optimizer, the learning rate optimizer is ReduceLROn Plateau, the initial learning rate is 0.001, and the number of iterations is set to 250 times.

Table 1 contains the parameter configuration of each layer in the network, which is mainly divided into the Conv layer, the CBAM layer, and the PrimaryCaps layer and the DigitCaps layer.

TABLE I. Network Model and Parameters

Layer	Parameters
Conv	inputChannel:1; outPutChannel:256 kernel size= 9; stride=1
CBAM	Channel:256
PrimaryCaps	InputChannel:256; OutputCaps:32*6*6 output_dim:8;kernel_size:9,stride:2
DigitCaps	Inputcaps:32*6*6; out_put_caps:10

TABLE II. Network Model Accuracy

Net Work Name	Routing Number	CBAM Module	Max Accuracy	First time
CapsNet1	1	False	99.70999908%	133
CapsNet1_CBAM	1	True	99.66999817%	100
CapsNet2	2	False	99.65000153%	111
CapsNet2_CBAM	2	True	99.61000061%	42
CapsNet3	3	False	99.69000244%	131
CapsNet3_CBAM	3	True	99.62002324%	82
CapsNet4	4	False	99.59999847%	141
CapsNet4_CBAM	4	True	99.55999756%	86
CapsNet5	5	False	99.65000153%	125

The data set in this article uses the MNIST data set, and the accuracy of Table 2 is calculated by calculating 10,000 images in the test set of the MNIST data set. The training set contained in the MNIST handwritten digits data set contains 60,000 example pictures, and the test set contains 10,000 examples. These numbers have been processed into a 28*28 picture and centered in the image. This data set is widely used in experiments in the field of machine learning. It was created by recombining samples from the original NIST dataset. The creator believes that because MNIST's training data set comes from US Census Bureau employees, and the test data set comes from American high school students, this is not an experiment suitable for machine learning, so the MNIST data set was born. In addition, the black and white images in MNIST are also normalized and processed into 28*28 images.

The Routing Number in the table is the routing times of the capsule network, CBAM Module is whether the CBAM module is added, Max accuracy is the highest accuracy rate in 100 training sessions, and First time is the highest accuracy rate of Epoch for the first time.

It can be seen from Table 2 that the highest accuracy point of the network after adding the CBAM module is about 50% ahead, and the accuracy of the loss is within the error range, and after the CBAM module is added, the complexity of the network is not Improve and strengthen the performance of the network. Therefore, we believe that the CBAM attention module can improve the training efficiency of the network and reduce the training time and cost of the network.

VI. CONCLUSION

This paper conducts theoretical research and comparison on the introduction of the CBAM module in the capsule network, and chooses to use Python and Pytorch frameworks to conduct experiments on the network, and uses the MNIST data set for verification research. Conclusions are drawn through 10 experiments under the same hardware environment. After adding the attention module, the training convergence speed has been greatly improved, and the accuracy of the network has not decreased, which proves the feasibility of the attention module in the capsule network. Taking into account the errors in the experiment, the number of network layers is small, and we will use a larger amount of data and deeper networks for research in future research.

REFERENCES

[1] Sara Sabour. Dynamic Routing Between Capsules. http s://arxiv.org/abs/1710.09829, 2017-11-07J.

[2] A-reum Lee, Yongwon Cho, Seongho Jin, Namkug Ki m. Enhancement of surgical hand gesture recognition us ing a capsule network for a contactless interface in the o perating room [J]. Elsevier B.V., 2020, 190.

[3] Amlan Basu, Lykourgos Petropoulakis, Gaetano Di Cat erina, John Soraghan. Indoor Home Scene Recognition Using Capsule Neural Networks [J]. Elsevier B.V., 202 0, 167.

[4] Yujia Wu, Jing Li, Jia Wu, Jun Chang. Siamese capsule networks with global and local features for text classifi cation [J]. Elsevier B.V., 2020, 390.

[5] Technology - Ambient Intelligence and Humanized Co mputing; Findings from Aliah University in the Area of Ambient Intelligence and Humanized Computing Repo rted (Handwritten Arabic numerals recognition using co nvolutional neural network) [J]. Journal of Engineering, 2020.

[6] Lu Chunyan. Improvement of Capsule Network and Its Application in Image Generation [D]. Southwest Unive rsity, 2019.

[7] Ren Qiang, He Lianghua. Research on Variable Dimens ion Capsule Based on Capsule Network [J]. Computer k nowledge and technology, 2020, 16(02):204-205I.

[8] Jaiswal Amit Kumar, Tiwari Prayag, Garg Sahil, Hossai n M. Shamim. Entity-aware capsule network for multi-c lass classification of big data: A deep learning approach [J]. Future Generation Computer Systems, 2021, 117.

[9] Ren Haohao, Yu Xuelian, Zou Lin, Zhou Yun, Wang X uegang, Bruzzone Lorenzo. Extended convolutional cap sule network with application on SAR automatic target recognition [J]. Signal Processing, 2021, 183(prepublis h).

[10] Long Fei, Peng Jing-Jie, Song Weitao, Xia Xiaobo, San g Jun. BloodCaps: A capsule network based model for t he multiclassification of human peripheral blood cells [J]. Computer Methods and Programs in Biomedicine, 2 021, 202.

[11] Hongwang Xiao, Yun Yang, Ke Yu, Jiao Tian, Xinyi C ai, Usman Muhammad, Jinjun Chen. Sign language digi ts and alphabets recognition by capsule networks [J]. Jo urnal of Ambient Intelligence and Humanized Computi ng, 2021.

[12] Tiwari Shamik. Dermatoscopy Using Multi-Layer Perc eptron, Convolution Neural Network, and Capsule Net work to Differentiate Malignant Melanoma From Benig n Nevus [J]. International Journal of Healthcare Inform ation Systems and Informatics (IJHISI), 2021, 16(3).

[13] Akshi Kumar, Nitin Sachdeva. Multimodal cyberbullyi ng detection using capsule network with dynamic routin g and deep convolutional neural network [J]. Multimedi a Systems, 2021.