Large-scale Image Classification with Explainable deep learning scheme

Bin Wu[a], Yuhong Fan[b]*, Li Mao[c]

[a]*School of Internet of Things Engineering, Jiangnan University, Wuxi, 214122, China*

[b]*Department of Computer Engineering, LangFang YanJing Vocational Technical College,Sanhe,065200,China*

[c]*School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, 214122, China*

Abstract: For the uncertainty and complexity in object decision making and the differences of decision makers' reliabilities, an object decision making method based on deep learning theory is proposed. However, traditional deep learning approaches optimize the parameters in an "end-to-end" mode by annotating large amounts of data to propagate the errors backwards. The learning method could be considered to be as a "black box", which is weak in explainability. Explainability refers to an algorithm that gives a clear summary of a particular task and connects it to defined principles or principles in the human world. This paper proposes an explainable attention model consisting of channel attention module and spatial attention module. The proposed module derives attention graphs from channel dimension and spatial dimension respectively, then the input features are selectively learned according to the importance of the features. For different channels, the higher the weight, the higher the correlation which required more attention. The main function of spatial attention is to capture the most informative part in the local feature graph, which is a supplement to channel attention. We evaluate our proposed module based on the ImageNet-1K and Cifar-100 respectively. Experimental results show that our

algorithm is superior in both accuracy and robustness compared with the state of the arts.

Keywords: Large-scale image classification; Spatial attention module; Object decision making; Channel attention module; Explainable AI

1 Introduction

In computer vision, image classification is a subject worthy of long-time research, and it is an important foundation in some fields such as object detection (Diba et al., 2017; Ren et al., 2017), face recognition (Girshick et al., 2014), pose estimation (Chen et al., 2019; Chen et al., 2021), population density estimation(Zhang et al., 2012; Chen et al., 2021), image segmentation (Chen et al., 2020 and Huang et al., 2019) etc. Therefore, image classification has always been a hot topic (Lin et al., 2019 and Lin et al., 2021). At present, the image classification algorithm mainly relies on the deep neural network model. Besides, a lot of high quality images are needed. After training the deep neural network model, an input image can be correctly recognized. However, due to the large number of image categories and the limitation of computing resources, it is difficult for traditional classification algorithms to achieve satisfactory accuracy. In fact, deep learning is an extension of machine learning and many scholars have done a lot of research on it. Different from the traditional image classification method, it does not require complex feature decomposition of the target image. Deep learning uses deep neural network models and a large number of images to learn features. Thus, the deep learning algorithm is suitable for image classification task. The unexplained black box in AI system makes users can only get decision results, and cannot distinguish the specific reasons, processes and action logic of the system's decision-making (Florian et al., 2015 and Leung et al., 2013). This kind of

AI system is difficult to achieve the trust and understanding of different decision-makers, because we fail to know what is controlling the design, operation and decision-making, especially for autonomous decision-making with uncontrollable risks. In this regard, the Explainable AI (XAI) is widely developed in handling these issues (Parkhi et al., 2015). Generally, different decision-makers possess different goals in XAI to ensure the high performance of intelligent decision-making and provide a reasonable explainable model synchronously, so that different stakeholders could effectively understand, trust and manage AI systems (Jamie et al., 2011). In this process, multi-attribute group decision-making has been widely used, which can effectively evaluate different decision schemes through extracting and fusing the cognitive judgment information of multiple decision-makers (Alejandro 2016; Alex et al., 2017), ensuring that multiple decision-makers can achieve group consensus and trust for the Explainable AI.

In daily life, when we look at a image, the most basic task is what the image is, whether it is a landscape image or a figure image, whether it describes a building or food. For computer vision, this is an image classification task. The main difficulty of image classification is the step of feature extraction. Once a distinguishable feature is found, the image classification becomes very easy. The so-called feature extraction refers to constructing an algorithm to extract features in the target image, such as the edge feature of the face, the color feature of the skin etc. The extracted features are used to distinguish the target object as much as possible from other objects. For example, what you need to distinguish is a black cat from a white cat, so the color feature is definitely a good feature. However, the difficulties encountered in life are often difficult to extract features, such as detecting pedestrians and vehicles on noisy streets. The task requires high accuracy of the detection algorithm to avoid a car

accident.

At present, the best approach is to use deep neural network model for image classification tasks (Cires et al., 2012). The experimental result of VGGnet shows that the block with the same shape can obtain better classification accuracy by constructing deeper convolutional neural network model. Inspired by this idea, the deep residual network is constructed by cross-layer connection method and it achieves higher classification accuracy. GoogLeNet increases the adaptability of the network to different scales, showing that adjusting the width of model is also an important method to obtain better classification accuracy. ResNeXt and Xception added cardinality to the network model, proving that the cardinality can not only reduce the overall parameters of the model, but also has a strong ability of representation.

The attention mechanism is similar to that of human vision. When we look at the scene around us, we always focus our attention on the main things to get key information (Pentina et al., 2015). The main purpose of the attention mechanism is to make the system focus its attention on key information in the scene. Attention mechanics can be used in a wide range of scenarios. The neural network captures key information with the help of the attention mechanism and we can take use of attention mechanism to observe things in the environment better. In the traditional neural network model, adding convolution channels and operating multiple convolutions on features in the same channel usually bring a certain degree of accuracy improvement. The attention mechanism make neural network model know how to pay attention on channel dimension and spatial dimension. In order to verify the role of attention mechanism in computer vision more clearly, the CAM and the SAM are analyzed from the point of attention domain. The attention domain mainly consists of three types: spatial domain, channel domain and mixed domain. In our experiments, it is

found that better performance is obtained with using the CAM before the SAM.

Daniel Wolpert believed that the reason for the evolution of the brain is not for thinking and feeling, but for controlling motion, which is the core idea of the Deep Reinforcement Learning (Lin et al., 2017 and Holzinger et al., 2017). The problem model studied by it includes an environment and an agent interacting with the environment. The goal of reinforcement learning is to design a behavioral strategy for the agent that maximizes its benefits in interacting with the environment. Google DeepMind used this strategy in 2016 to get computers to go beyond the level of top professionals. However, the development of object detection and behavior recognition in complex environment is slow, that is mainly because :(1) the efficiency of model optimization algorithm is not high; (2) Lack of interpretability of constructed model; (3) Small sample with label; (4) Complex and unrestricted scenes; (5) High-dimensional video data leads to difficulty in parameter adjustment. Curriculum Learning and self-paced Learning represent the recently proposed Learning strategy. Their core idea is to simulate the cognitive mechanism of human beings, they first learn simple and general knowledge structure, then gradually increase the difficulty degree and transition to more complex and professional knowledge. These two methods have similar conceptual learning paradigms, but differ in their specific learning schemes.

This paper proposes a novel image classification framework based on multi-perspective deep transfer learning. Attention mechanism is used to describe varied image characteristics, the self-paced learning strategy is adopted to solve the problems of small number of labeled samples and model interpretation. A nonlinear model based on deep transfer learning is proposed to solve the problem that it is difficult to distinguish the perspective-related features from image-related features,

the interpretability of the model is further improved.

The main structure of this paper is as follows: Section 2 starts with a brief presentation of necessary related concepts. Section 3 proposes a channel and spatial attention module based on explainable machine learning. Designed experiments and discussion are depicted in Section 4. Finally, some conclusions are presented in Section 5.

2 Related work

Explainable AI (XAI) refers to those Artificial Intelligence techniques aimed at explaining, to a given audience, the details or reasons by which a model produces its output (Arrieta et al., 2020). To this end, XAI borrows concepts from philosophy, cognitive sciences and social psychology to yield a spectrum of methodological approaches that can provide explainable decisions for users without a strong background on Artificial Intelligence. Therefore, XAI targets at bridging the gap between the complexity of the model to be explained, and the cognitive skills of the audience for which explainability is sought. Interdisciplinary XAI methods have so far embraced assorted elements from multiple disciplines, including signal processing, adversarial learning, visual analytics or cognitive modeling, to mention a few. Although reported XAI advances have risen sharply in recent times (Russakovsky et al., 2020 and Deng et al., 2015), there is global consensus around the need for further studies around the explainability of ML models. A major focus has been placed on XAI developments that involve the human in the loop and thereby, become human-centric. This includes interpretable reasoning of models, neurosymbolic reasoning or systems based on fuzzy rules, etc. At present, explanations provided by different algorithms are fragmented and independent, which makes it difficult to determine reasonable decisions and explain model structures. In addition, in the

149 design of interpretable classifier, the selection of optimal training set, correlation
150 selection of heat graph, semantic analysis, model visual interpretation and error
151 analysis cannot be combined compulsively (Wang et al., 2017 and Yang et al., 2020).
152 Moreover, the text interpretation can't match the characteristics of a certain layer of
153 deep learning network, and it lacks of continuous interpretibility. In general, the text
154 explanation generated for classification comes from training data based on model
155 annotations. Up to now,data labels are set manually and are very subjective, and it
156 doesn't take into account the differences between the different elements. Therefore, it
157 is not possible to determine the relevant region of the image that is most useful for
158 classification. In most cases, experts are encouraged to use their attribute label data as
159 interpretable evidence. Existing interpretable artifical intelligence models can provide
160 the basis behind the classification. However, in the existing classification model,there
161 is no mechanism for identifying potential misclassification of classifiers. Warning
162 users about misclassification will help prevent errors from entering the system. One of
163 the reasons for misclassification is the reduction of distance between classes. Some
164 outliers or edge elements of a class can share the common characteristics of adjacent
165 classes. However, there is no mechanism to ensure the number of subclasses of a
166 given class and whether it makes sense to merge closely related subclasses of two
167 adjacent classes into a new class and implement the correct classification. Xu (Xu et
168 al., 2017) proposes a solution based on database transaction model interpretation,
169 whose explanation is on the basic of logical structure or reasoning. The static structure
170 makes it unsuitable for the deep network classifier. In the constructed model, the
171 system dynamically give appropriate explanations from stored vocabularies, which in
172 turn are generated based on model learning. It provides a consistent view of models
173 and interpretations beyond the scope of existing technology. Reference (Sachan et al.,

2020) proposes an interpretable model in which they are interactively validated through visual features and similarity. Moreover, k-means clustering was used to analyze the similar features, so that the average features obtained had greater robustness and relatively low time complexity. In addition, it does not consider the importance or relevance of model, nor does it cluster with respect to output classes. He et al. (He et al., 2016) conducts the explanatory demonstration of the model mainly from the aspect of correlation calculation. By observing the change of the connection mode of the network, the hidden layer is explained visually. However, the feature learned by the network layer are not described in detail. Simonyan et al. (Simonyan et al., 2020) proposed an image interpretation and generation method based on visual features. Take an image signature with a fixed length of 8000 to generate a caption. In this model, the correlation of features is firstly determined and the signature generation is carried out on this basis. Since eigenvalues can be of any length, strategies that follow highly correlated features are interpretable. As the power of interpretation becomes more important in intelligent decision-making, AI systems are no longer there to serve as black boxes (Xie et al., 2016) . Decision makers of AI services have the right to know the reasons behind their decisions so that they can better play to their strengths.

Abhronil et al. (Abhronil et al., 2019) propose a novel model based on the attention mechanism and it is named the residual attention network. As the network layer deepens, attention modules can extract key information from different layers. Finally, it got 4.8% Top-5 error rate on ImageNet. HU J et al. (Hu et al., 2019) proposed the SENet 44 network model. In the training process, the model can distinguish the importance of different channels, then enhances useful features and inhibits useless features according to the importance of feature. Finally, it won the

ILSVRC2017 classification task championship with a 2.25% Top-5 error rate. This paper mainly verifies the interpretability of the model through image classification scenarios in real applications.

3 The proposed channel and spatial attention model

This paper makes an deep study of image classification and recognition from the most cutting-edge technical perspectives such as deep reinforcement learning, self-paced learning and transfer learning. In view of the existing problems in the field of image classification in complex environments, this paper considers the solutions based on deep network, and further studies popular algorithms such as deep reinforcement learning, self-paced learning and transfer learning, aiming at the urgent model explanation problems in the framework of deep network. On this basis, object detection technology and image recognition technology in complex environments are specifically studied.The overall technical framework of the paper and the relationship between the research contents of each part are shown in Figure 1, and the research of this paper is also carried out according to these contents one by one.

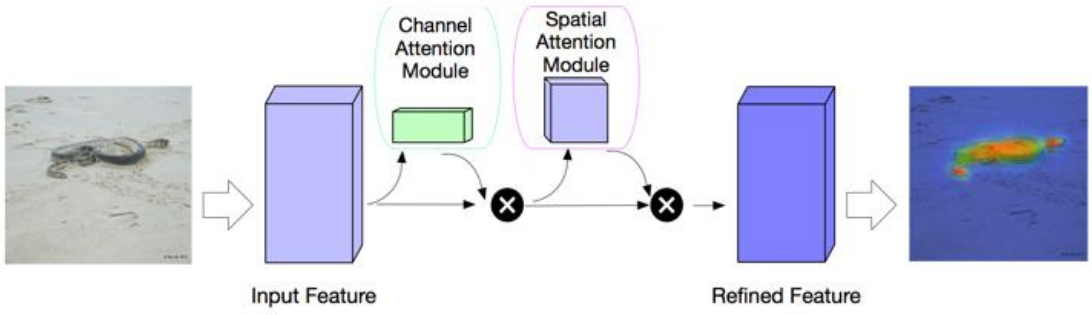

Figure 1. The overview of model

For the convolutional neural network model, depth, width and attention mechanism are the main factors affecting the accuracy of image classification. At present, attention mechanism contains CAM and SAM. CAM acts on the channel

219 domain, weighting different channel features. For a C × H × W feature graph, the
220 C weight of channel attention is different, while the weight of H × W is the same.
221 For CAM, the weight of each C on different channel dimensions needs to be learned.
222 To reduce the amount of computation and improve classification accuracy, the pooling
223 layer in the general convolutional neural network directly uses the maximum pooling
224 method or average pooling method to compress the image information. For SAM,
225 only the key information in the spatial features is extracted. We first introduce the
226 general framework of proposed model with CAM and SAM in this section. Finally,
227 we describe how to pull them together.

228 After convolution operations, an intermediate feature map $\mathrm{F_{in}} \in \mathbb{R}^{\mathrm{C\times H\times W}}$ is
229 obtained. $\mathrm{F_{in}}$ is the input of model. One-dimensional channel map $\mathrm{M_c} \in \mathbb{R}^{\mathrm{C\times 1\times 1}}$ is
230 inferred after the channel attention module and two-dimensional spatial map $\mathrm{M_s} \in$
231 $\mathbb{R}^{1\times(\mathrm{H}\times\mathrm{W})\times(\mathrm{H}\times\mathrm{W})}$ is inferred after the spatial attention module as shown in Fig. 1.
232 The entire attention calculation process can be summarized as：

233 $$\mathrm{F_1} = \mathrm{M_c}(\mathrm{F_{in}}) \otimes \mathrm{F_{in}} \qquad (1)$$

234 $$\mathrm{F_{out}} = \mathrm{M_s}(\mathrm{F_1}) \otimes \mathrm{F_1} \qquad (2)$$

235 where $\otimes$ denotes element-wise multiplication. After F_{in} passing the channel
236 attention module, F1 is achieved. F_{out} is the final output and the attention value is
237 broadcasted during multiplication process. Attention and feature are the multiplication
238 of element levels, which will be propagated automatically, that is, channel attention
239 broadcasts along spatial dimension, while spatial attention broadcasts along channel
240 dimension
241

3.1 Channel attention module

The CAM captures the relationship between channel features. It pays attention to the key channel information and weakens the influence of the useless channel information (Russakovsky et al., 2010 and Deng et al., 2009). It uses an attention mechanism similar to the self attention mechanism (query, key, value) to get the similarity between channel graphs, and then use the weight of channel graphs to update. Finally, the matrix of computation attention is obtained, which can enhance the key features. The CAM makes the neural network model to pay attention to the channel features with the key information. On the basis of convolution, we first extrude the feature graph to obtain the global feature of each channel. Then, we use the global feature to get the relationship between different channels and the get the weight of different channels. Finally, we multiply the weights to get the features on the basis of the original feature graph.

In a convolutional neural network, the convolution operation only performs on image feature space, it is difficult for the convolution module to get the relationship between different feature channels. To get an eigenmatrix, an image needs go through several convolutional layers and the number of channels represents the number of cores in the convolutional layers. In a normal neural network, the number of convolution kernels is usually as high as 1024 or 2048. Therefore, not every channel is useful for feature extraction. The CAM will help the neural network model to select more informative channels. Besides, We encode spatial features using a global average pool that provides feedback for each pixel on the feature map. The following formula shows the global average pool calculation process.

$$F_{avg} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{in}(i, j) \quad (3)$$

where $F_{avg} \in \mathbb{R}^{C\times1\times1}$ represents the result after implementing the global average pooling on the input feature map F_{in}. In order to capture the relationships between different channels when obtaining the global description features, two conditions need to be met for CAM: firstly, it must be flexible, because it needs to learn the nonlinear relationship between different channels; Secondly, the learning relationship is not mutually exclusive, because it allows for a multichannel feature instead of a hot spot form. We describe the channel attention map Mc as follows.

$$M_c = \sigma\left(W_2 ReLU\left(W_1 F_{avg}\right)\right) \quad (4)$$

where $W_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $W_2 \in \mathbb{R}^{C\times\frac{C}{r}}$, σ denotes Sigmoid function. W_1 and W_2 are fully connected layers. To improve the explanatory ability of the model, we construct two fully connected layers, namely the bottleneck structure. W_1 layer is a way of dimensionality reduction and the dimensionality reduction factor r is a super parameter. Then the Relu function is used and the W_2 layer restores the dimension to the original finally. The process is shown in Figure 2.

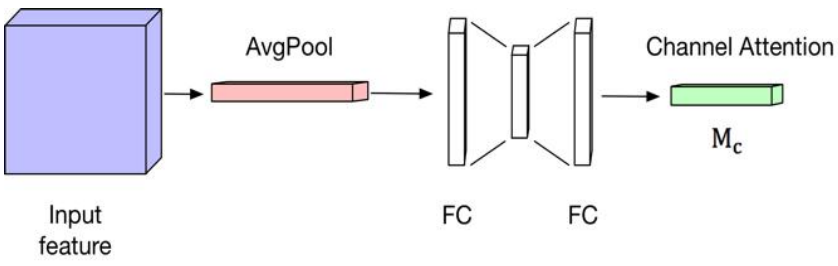

Figure 2. The details of channel attention module

3.2 Spatial attention module

Unlike CAM, SAM only plays the role of distinguishing key information within a single image feature map. First of all, we use average pooling and max pooling to compress the input feature and then we use mean and max operations on

the input feature at channel dismensions. Finally, we get two two-dimensional features. Considering different channel size, the two two-dimensional features are spliced together to obtain the feature with channel number of 2. And they are convolved to ensure that the resulting features are consistent with the input feature in spatial dimension.

Spatial attention map is generated by paying attention to the internal relationship, as shown in Figure 3. $A \in \mathbb{R}^{C\times H\times W}$ represents the input of the SAM. After consducting convolutional layers, A generates feature maps B, C and D where $\{B, C, D\} \in \mathbb{R}^{C\times H\times W}$. We reshape B and C to $\mathbb{R}^{C\times(H\times W)}$ and H $\times$ W represents pixels in spatial module.

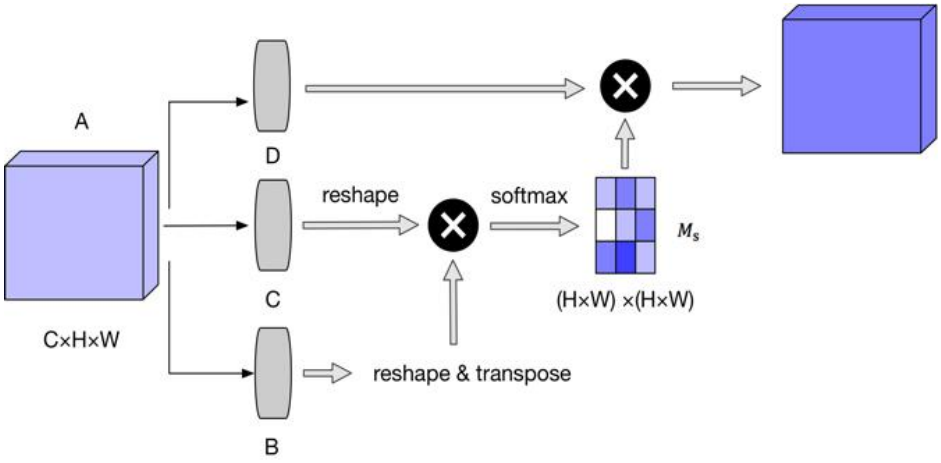

Figure 3. The details of spatial attention module

Then we get $\mathbb{R}^{(H\times W)\times(H\times W)}$ by operating a matrix multiplication between the transpose of B and C. We get the spatial attention map Ms when a softmax layer is applied. Ms is computed as follows.

$$M_{s_{ij}} = \frac{exp(B_i \times C_j)}{\sum_{i=1}^{H\times W} exp(B_i \times C_j)} \tag{5}$$

where the stronger the correlation, the more similar the features of the two sites. Figure 4 shows the visual result of images. It is observed that our proposed model focuses limited attention on key information, saves resources, and quickly obtains the most effective information.

Figure 4. Diagram of the visualization results

3.3 Explainable object detection model of combing self-paced learning and deep reinforcement learning

The deep network model is designed for learning across model characteristics.This paper proposes a new application in deep network, using deep network to learn multi-mode. In particular, this paper demonstrates that cross-modal feature learning - if multimodal features are present during feature learning, better features can be learned for a modal (multimodal learning, monomodular-tested). In addition, the paper designed how to learn a shared feature between multiple modes and evaluate it on a particular task -- the classifier was trained with only audio data but tested on video-only data. Multi-mode explainable deep network model can be seen in Figure 5. This model consists of two streams, one for video information and the other for audio information. The structure of the two streams is identical, each consisting of eight layers (including the input layer).

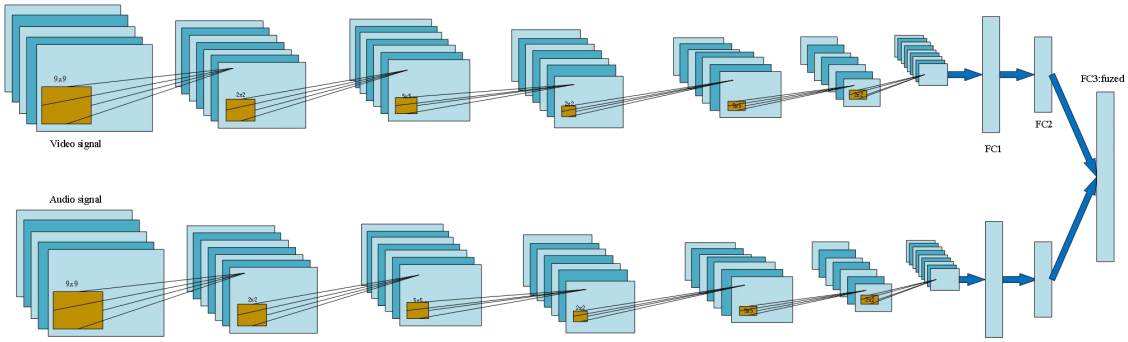

Fig. 5. Proposed multi-mode explainable deep network model

There are two problems with the traditional multimodal model. First, there is no clear goal for the model to find correlations across modes. Some hidden layer units adjust the parameters only for voice, and others adjust the parameters only for video, so that it is possible for the model to find the desired feature. Second, there is only one mode for supervised training and testing in the cross-modal learning arrangement, which makes the model unexplainable. If there is only one modal representation, it is necessary to integrate observable variables that are not observed. Therefore, this paper proposes a deep self-coding model to solve the above problems.Inspired by the noise-reducing self-coding model, this paper proposes a training two-mode deep self-coding model (Figure 5), which uses an extended (extended single-mode input) but noisy data set. In fact, the model is still required to reconstruct the two modes when one mode uses zero as input and the other uses the original value as input when expanding.Therefore, one-third of the training data is input only by video, one-third of the training data is input only by voice, and the last third has both video and voice. This model can be viewed as an example of multitasking learning.

When designing the intensification strategy, the paper uses the Q network to interact with its environment during the data generation phase. The system looks at the current scene, which consists of audio and video frames, and takes actions using the -greedy strategy. This environment in turn provides scalar rewards.Interaction

experiences are stored in replay memory M. Replaying M preserves N recent experiences, which are then used to update the network parameters during the training phase. In the training stage, the network structure will use the data stored in replay memory M to train the network.Assume that the superparameter n represents the number of experiences replay, and for each experience replay, a mini-cache B containing several interactions is randomly sampled from the finite size replay memory M. The model will be trained by sampling from cache B, and the parameters of the network will be updated iteratively in the direction of The Behrman target.The algorithm is divided into two phases to avoid latency. Therefore, this paper divides the algorithm into two stages: in the first stage, the robot collects data through limited time interaction with human beings; In the second stage, it enters the stage. During this rest phase, the training phase is activated to train the multimodal depth Q network.

In this paper, for the sake of the regularization model and make it thin, it is needed to make each unit has a hidden layer using the regularized the expected activation function of punishment, the form of the regularized punishment is the need to focus on research, it determines the cell activation function of hidden layers on the sparse sex (whether the function of the activation of the hidden layer unit is activated or not).

In order to avoid non-convex optimization problems from falling into poor local solutions, the proposed network optimization method adopts multiple random initializations to train the model, and then chooses the initialization network with the best effect to construct the model. However, this method is too adhoc and the calculation cost is too high. Self-learning is just the best solution to non-convex optimization problems. The curriculum learning is to simulate the cognitive

mechanism of human beings by first learning simple and universal knowledge structure and then gradually increasing the difficulty to learn more complex and specialized knowledge. However, self-learning has been improved in course learning. Instead of assigning prior knowledge to sample learning sequence in advance, the learning algorithm itself determines the next learning sample in each iteration.

4 Experimental results and analysis

The dataset is used in the experiment including ImageNet- 1K and Cifar-10050 (Hu et al., 2018; Zhang et al., 2010; Sanghyun et al., 2010; Chen et al., 2017; Zhao et al., 2018). The Cifar-100 contains 100 classes and every class has 600 color images. But every images is only a size of 32 × 32. Five hundred images in each class serve as the training set and the rest as test set. For each image, it has two labels, fine-labels and coarse-labels, which represent the fine-grained and coarse-grained labels of the image respectively and Cifar-100 is hierarchical. In Figure 6, we extracted the images from Cifar-100 as the visual example.

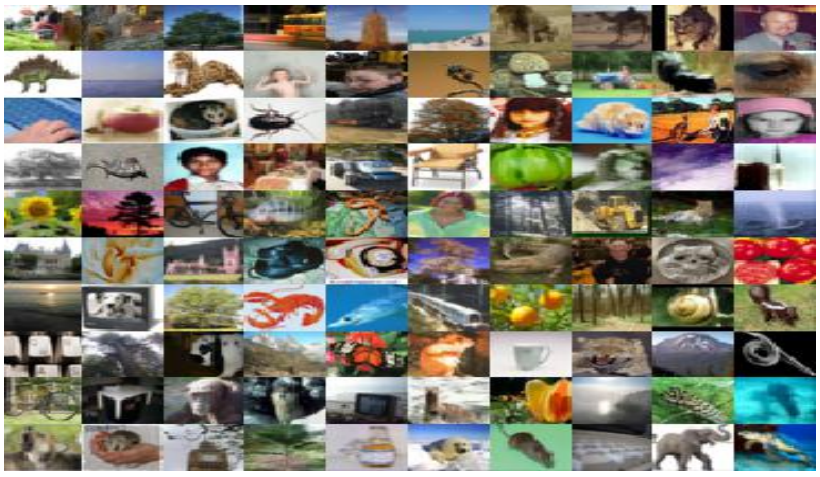

Figure 6. Visual examples of the Cifar-100

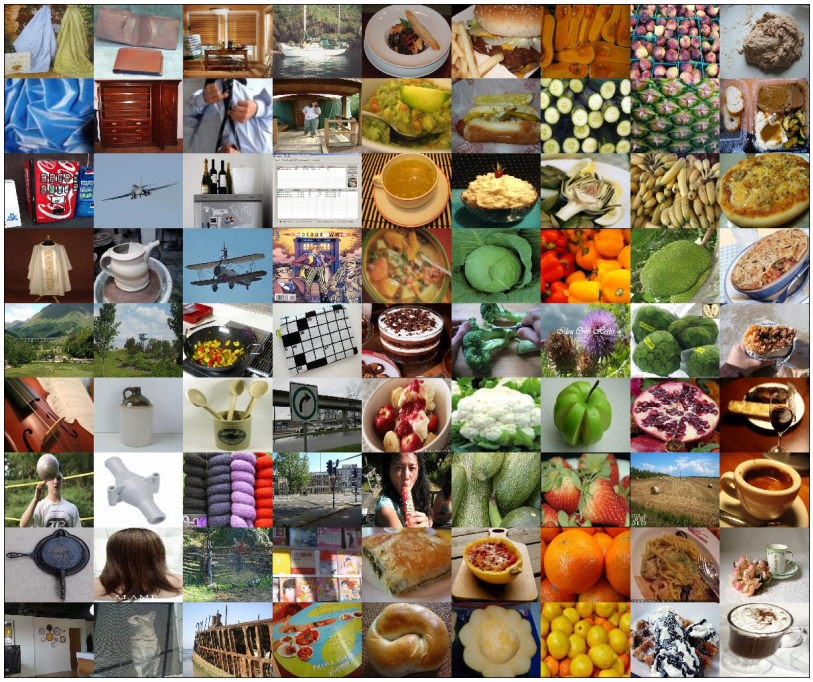

Figure 7. Visual examples of the ImageNet-1K

ImageNet-1K is an image dataset and each concept image is quality-controlled and manually tagged. At present, ImageNet-1K consists of 1,4197,122 images. The major categories include: animal, bird, fish, flower etc. In Figure 7 shows the visual examples of the ImageNet-1K.

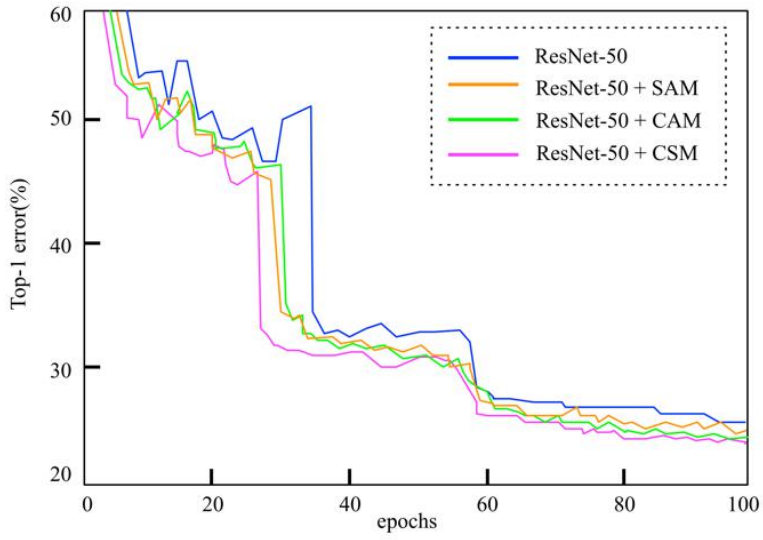

Figure 8. Comparison of different network models

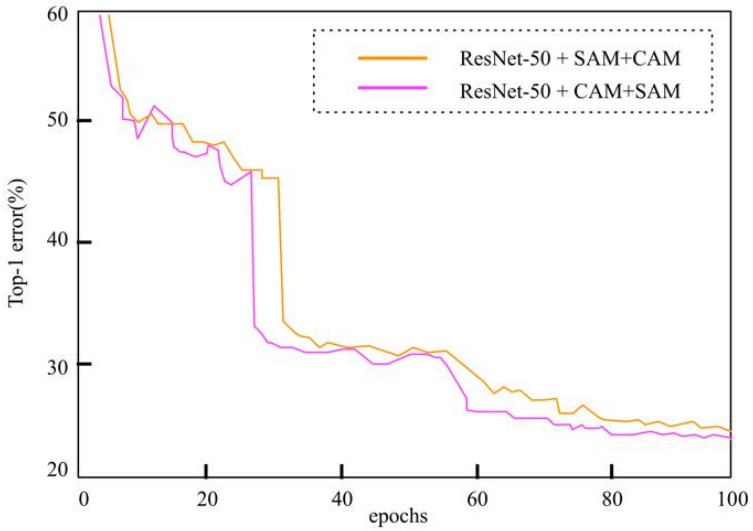

Figure 9. Comparison of different network models

In experiment, we compare the effectiveness between the CAM and the SAM. We compared 4 different network models: baseline network, baseline network with CAM, baseline network with SAM, baseline network with CSM. The result of experiment is shown in Figure 8. As is shown in Figure 8, it is easily concluded that the ResNet-50 model with CSM has achieved higher accuracy. We can observe that CAM perform better than SAM. Besides, combination of two attention modules can bring better performance. The experiment shows that it is effective to conduct CAM and SAM at the same time. The experimental results are shown in Table 1.Based on the above part of the experiment, we find that it is effective to conduct CAM and SAM at the same time for improving the expression ability of neural networks. We want to know if the CAM should be used before SAM. So, in the ablation experiment, we compare the effect of CAM and SAM in different order of use. The baseline network with CAM and SAM, baseline network with SAM and CAM are conducted respectively. Figure 9 shows the result of experiment.

Consistent with the above experiment, adding the attention module still bring improvement on image classification accuracy from Figure 8. We can observe that the

CAM-first combination method achieves better performance. The experimental result shows that CAM-first combination method is more effective.

TABLE I. COMPARISON OF DIFFERENT NETWORK MODELS

Architecture	Top-1 error (%)
ResNet-50	24.55
ResNet-50+SAM	23.47
ResNet-50+CAM	23.21
ResNet-50+CSM	22.78

4.1 Image Classification on ImageNet-1K

In this part, we use ResNet and WideResNet as the baseline model. On this basis, we add attention mechanism for comparison. The extensive image classification experiments are conducted based on the ImageNet-1K. The structure of ResNet with adding SE module is shown in Figure 9 and the results of the experiment is shown in Table 2. The experiment still prove that networks with CSM performs better than the baseline module, indicating that attention mechanism can be well used on the various network models. Besides, the depth and width of the neural network also greatly affect image classification accuracy. SENet won the ILSVRC2017 classification task championship. But CSM fuses channel features with spatial features for better representation capabilities and CSM performs better than SENet.

TABLE II. RESULTS OF THE EXPERIMENT ON THE IMAGENET-1K DATASET

Architecture	Top-1 error (%)	Top-5 error (%)
ResNet-18	29.62	10.56
ResNet-18+SE	29.43	10.24
ResNet-18+CSM	29.29	10.12
ResNet-34	26.71	8.62
ResNet-34+SE	26.16	8.37
ResNet-34+CSM	26.03	8.28
ResNet-50	24.55	7.52
ResNet-50+SE	23.21	6.74
ResNet-50+CSM	22.78	6.57

WideResNet18(widen=1.5)	26.86	8.91
WideResNet18(widen=1.5) +SE	26.23	8.51
WideResNet18(widen=1.5) +CSM	26.14	8.49

4.2 Image Classification on CIFAR-100

Based on Cifar-100, we conduct image classification experiment to verify the effectiveness of the CSM. ResNet and WideResNet are used as baseline model. Table 3 shows the experimental result. The experimental results prove that the combination of CAM and SAM can improve classification accuracy. Besides, the depth and width of the neural network also greatly affect image classification accuracy.

TABLE III. RESULTS OF THE EXPERIMENT ON THE CIFIR-100 DATASET

Architecture	Accuracy (%)
ResNet-18	91.7
ResNet-18+CSM	93.1
ResNet-34	92.4
ResNet-34+CSM	93.8
ResNet-50	92.9
ResNet-50+CSM	94.3
WideResNet18(widen=1.5)	92.8
WideResNet18(widen=1.5) +CSM	94.2

As can be seen from Table 3., the prediction density map we get after adding feature self-learning is closer to the true value, so we can see from Fig.8 and Fig.9 below that we can more accurately describe the feature after adding feature self-learning Loss reduction and corresponding changes in MAE accuracy, we also give the performance of net similarity.

5 Conclusions

Different from the previous research on image classification, we propose an attention module based on spatial dimension and channel dimension. This module derives the attention map by CAM and SAM respectively. Then it multiplies the attention map into the input feature map. The experiment shows that adding the

attention module to some image classification algorithms is an effective method. In order to make image classification perform better, combination of CAM and SAM can improve classification accuracy and the CAM should be used before SAM. The CAM selectively enhances some feature channels and suppresses some feature channels by learning the relational mapping. The SAM aggregates features by weighting features at spatial dismension. We conducted a lot of image classification experiments for comparison based on the ImageNet-1K and Cifar-100. In fact, the attention module can be well embedded in different deep neural networks and it improves the explainable deep neural network's ability of expression. Besides, the width and depth of the neural networks are also worth considering.

In the following research, we will pay more attention to the improvement of real-time algorithm and effectiveness of explainable deep learning, also be able to obtain more accurate data collection develop an intelligent prediction machines for image classification based on more effective machine learning approaches.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NO. 61702226); the 111 Project (B12018); open Fund of Jiangsu Key Laboratory of Image and Video Understanding for Social Safety, Nanjing University of Science and Technology, Nanjing (J2021-7).

Conflict of interest: The authors declare that they have no conflict of interest.

Data availability: The data that support the findings of this paper are available from the corresponding author.

Authorship contributions: **Bin Wu**: Investigation, Methodology, Writing- original

draft, Supervision.

Yuhong Fan: Writing- Reviewing and Editing, Methodology, Visualization.

Li Mao: Validation, Visualization, Data curation.

References

Diba A, Sharma V, Pazandeh A et al. Weakly super-vised cascaded convolutional networks. In: IEEE conference on computer vision and pattern recognition (CVPR), July 2017, pp. 5131-5139. New York, NY: IEEE.

Ren, Shaoqing, Kaiming He, Ross Girshick and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans Pattern Analysis And Machine Intelligence 2017; 39: 1137–49.

Girshick, Ross, Jeff Donahue, Trevor Darrell and Jitendra Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In: IEEE conference on computer vision and pattern recognition (CVPR), 2014, pp. 580–87.

Xing Chen, Ming Li, Hao Zhong*, Yun Ma, Ching-Hsien Hsu. DNNOff: Offloading DNN-based Intelligent IoT Applications in Mobile Edge Computing. IEEE Transactions on Industrial Informatics, Publish Online, DOI: 10.1109/TII.2021.3075464.

Xing Chen, Shihong Chen, Yun Ma, Bichun Liu, Ying Zhang, Gang Huang*. An Adaptive Offloading Framework for Android Applications in Mobile Edge Computing. SCIENCE CHINA Information Sciences, 2019, 62(8): 82102.

Ying Zhang, Gang Huang, Xuanzhe Liu, Wei Zhang, Hong Mei, Shunxiang Yang. Refactoring android Java code for on-demand computation offloading. ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages, and Applications. 2012.

Xing Chen, Junqin Hu, Zheyi Chen*, Bing Lin*, Naixue Xiong, Geyong Min. A Reinforcement Learning Empowered Feedback Control System for Industrial Internet of Things. IEEE Transactions on Industrial Informatics, Publish Online, DOI: 10.1109/TII.2021.3076393.

Bing Lin, Yinhao Huang, Jianshan Zhang, Junqin Hu, Xing Chen*, Jun Li*.

Cost-Driven Offloading for DNN-based Applications over Cloud, Edge and End Devices. IEEE Transactions on Industrial Informatics, 2020, 16(8): 5456-5466.

Bing Lin, Fangning Zhu, Jianshan Zhang, Jiaqing Chen, Xing Chen*, Naixue Xiong, Jaime Lloret Mauri. A Time-driven Data Placement Strategy for a Scientific Workflow Combining Edge Computing and Cloud Computing. IEEE Transactions on Industrial Informatics, 2019, 15(7): 4254-4265.

Xing Chen, Fangning Zhu, Zheyi Chen*, Geyong Min*, Xianghan Zheng, Chunming Rong. Resource Allocation for Cloud-based Software Services Using Prediction-Enabled Feedback Control with Reinforcement Learning. IEEE Transactions on Cloud Computing, Publish Online, DOI: 10.1109/TCC.2020.2992537.

Gang Huang, Chaoran Luo, Kaidong Wu, Yun Ma and Ying Zhang, Xuanzhe Liu. Software-Defined Infrastructure for Decentralized Data Lifecycle Governance: Principled Design and Open Challenges. IEEE International Conference on Distributed Computing Systems, 2019.

Schroff, Florian, Dmitry Kalenichenko, and James Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. In: IEEE conference on computer vision and pattern recognition (CVPR), 2015, pp. 815 – 23.

Leung Y., Ji N.N., Ma J.H., An integrated information fusion approach based on the theory of evidence and group decision-making, Information Fusion, (2013),14(4):410-422.

Parkhi, Omkar M, Andrea Vedaldi, and Andrew Zisserman. Deep Face Recognition. In: British machine vision conference (BMVC), 2015.

Shotton, Jamie, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman and Andrew Blake. Real-Time Human Pose Recognition in Parts from Single Depth Images. In: IEEE conference on computer vision and pattern recognition (CVPR), 2011, pp. 1297 – 1304.

Newell, Alejandro, Kaiyu Yang, and Jia Deng. Stacked Hourglass Networks for Human Pose Estimation. In: Computer Vision-ECCV 2016, pp. 483 – 99.

Krizhevsky, Alex, Ilya Sutskever, Geoffrey E and Hinton. ImageNet Classification with Deep Convolutional Neural Networks. Communications of The ACM 2017; 60: 84 – 90.

Cires¸an, Dan, Ueli Meier and Juergen Schmidhuber. Multi Column Deep Neural

Networks for Image Classification. In: IEEE conference on computer vision and pattern recognition (CVPR), 2012, pp. 3642–49.

Cires¸an, Dan, Ueli Meier and Juergen Schmidhuber. Multi Column Deep Neural Networks for Image Classification. In: IEEE conference on computer vision and pattern recognition (CVPR), 2012. pp. 3642–49.

A.Pentina, V.Sharmanska, and C.H.Lampert. Curriculum learning of multiple tasks. In Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition, pages. 2547–2554, (2015).

L.Lin, K.Wang, D.Meng, et al. Active Self-Paced Learning for Cost-Effective and Progressive Face Identification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, PP(99):7-19, (2017).

A.Holzinger, C.Biemann, S.P.Constantinos, and B.K.Douglas, What do we need to build explainable ai systems for the medical domain?, arXiv:1411.1784,(2017).

A. Barredo Arrieta, N. Diaz-Rodriguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”, Information Fusion, vol. 58, pp. 82-115, June (2020).

D. Xu, W. l. O. Yang, Xavier. Alameda-Pineda, E. Ricci, X. g. Wang, and N.cu. Sebe. Learning deep structured multi-scale features using attention-gated crfs for contour prediction. In NIPS, (2017), pages. 3961–3970.

Sachan, S., Yang, J. B., Xu, D. L., Benavides, D. E., Li, Y. An explainable AI decision-support-system to automate loan underwriting. Expert Systems with Applications, (2020), 144, 113100.

He K, Zhang X, Ren S and Sun J. Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition (CVPR), 2016.

Simonyan K and Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv preprint arXiv: 1409.1556.

Sengupta, Abhronil, Yuting Ye, Robert Wang, Chiao Liu and Kaushik Roy. Going Deeper in Spiking Neural Networks: VGG and Residual Architectures. Neuroscience, 2009; 13: 95–95.

Xie S,Girshick R, Dollar P, Tu Z and He K. 2016. Aggregated residual

562 transformations for deep neural networks. arXiv preprint arXiv: 1611.05431.
563 Wang, Fei, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang,
564 Xiaogang Wang and Xiaoou Tang. Residual Attention Network for Image
565 Classification. In IEEE conference on computer vision and pattern recognition
566 (CVPR), 2017, pp. 6450 – 58.
567 Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng and Olga Russakovsky. Towards
568 Fairer Datasets: Filtering and Balancing teh Distribution of the People Subtree in
569 the ImageNet Hierarchy. In: Conference on fairness, accountabiility and
570 transparency (FAT), 2020.
571 Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma,
572 Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander
573 C. Berg and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge.
574 International journal of computer vision (IJCV), 2015.
575 J Deng, A Berg, K Li and L Fei-Fei, What does classifying more than 10,000 image
576 categories tell us? In: Computer Vision-ECCV 2010.
577 O Russakovsky and L Fei-Fei, Attribute Learning in Large scale Datasets. In:
578 Computer Vision-ECCV 2010.
579 J Deng, W Dong, R Socher, L -J Li, K Li and L Fei-Fei, ImageNet: A Large-Scale
580 Hierarchical Image Database. In: IEEE international conference on computer
581 vision and pattern recognition (CVPR) , 2009.
582 HU J, SHEN L and SUN G. Squeeze-and-excitation net works. In: IEEE conference
583 on computer vision and pattern recognition (CVPR), June 2018, pp. 7132-7141.
584 New York, USA. NY: IEEE.
585 Zhang, Yulun, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong and Yun Fu. Image
586 Super-Resolution Using Very Deep Residual Channel Attention Networks. In:
587 Computer Vision ECCV 2010. pp. 294 – 310.
588 Woo, Sanghyun, Jongchan Park, Joon-Young Lee and So Kweon. CBAM:
589 Convolutional Block Attention Module. In: Computer Vision-ECCV 2010. pp. 3
590 – 19.
591 Chen, Long, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and
592 Tat-Seng Chua. SCA-CNN: Spatial and Channel-Wise Attention in
593 Convolutional Networks for Image Captioning. In: IEEE international
594 conference on computer vision and pattern recognition (CVPR), 2017. pp. 6298

– 6306.

Zhao, Hengshuang, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. PSANet: Point-Wise Spatial Attention Network for Scene Parsing. In: Computer Vision-ECCV 2018. pp. 267 – 83.

BIN WU received the B.Sc. degree from Jiangnan University, Wuxi, China, in 1996, and the M.S. degree from Jiangnan University, Wuxi, China, in 2005. He is currently a lecturer with the School of Internet of Things Engineering, Jiangnan University. His major research interests include visual surveillance, object detection, integrated circuit design and application of embedded system.

Yuhong Fan graduated from Shandong Agricultural University, Tai'an, China in 2008. In 2011, she obtained a master's degree from Xihua University, Chengdu, China. She is currently a lecturer in the Department of Computer Engineering， LangFang YanJing Vocational Technical College, Sanhe, China. She has studied many topics and written several high-quality journal papers and conference papers. His current research interests include data mining, information systems, wireless networks, artificial intelligence, Internet of things and security, medical data analysis, visual monitoring, scene understanding, behavior analysis, target detection and pattern analysis.

Li Mao received the B.Sc. degree from Southeast University (Nanjing, China) in 1990 and M.S. degree from Donghua University (Shanghai, China) in 2003. He is currently an associate professor at the department of computer science, Jiangnan University. His major research interests include visual surveillance, object detection, and data mining.