ISSN: 2664-9640

Application of Improved Residual Attentional Network Model in Fundus Images Recognition System

Meiyuan Xie [1,*], Yaoping He [2]

[1] School of Artificial Intelligence, Wenzhou Polytechnic, Wenzhou 325000, China

[2] Zhejiang Wangxin Medical Technology CO.,LTD, Hangzhou 310000, China.

* Corresponding Author

Abstract

According to the demand of medical image analysis platform in one medical applications, this paper proposes a attentional neural network model based on the architecture of the Residual Attention Network (RAN) model, For the new model, serveral modifications are made to it to improve the overall network performance. For the architecture, the soft mask branch is optimized so that the number of parameters is fewer, and the generalization ability is improved under the same training protocol. In addition, mixup technology is also introduced in the training method for the target task. As a result, this network model is experimented for the first time to aid diagnosis of medical fundus image data, along with experiments on commonly used image task of CIFAR10 / 100, ImageNet datasets to test its performance, The new model TRAN showed performance improvement in terms of accuracy on these image discrimination tasks, which justifies its values in pruning and simplification of RAN-type neural networks with attention mechanism. This algorithm can be adopted in medical industrial applications such as intelligent fundus image camera. The research also has useful implications for the network model design in environments with lower computing capacity such as mobile terminals.

Keywords

Diabetic retinopathy, CNN, attention, mixup.

1. Introduction

Diabetes is a chronic disease that affects wide populations throughout the world. Patients have a certain probability of suffering diabetic retinopathy, which may lead to sight degeneration and even enucleation of eyeball. Therefore, regular fundus screening is necessary as a precaution to alarm patients with risk to take therapeutical actions. There are many patients with diabetic disease in China, and thus huge demand for fundus screening. On the other hand, the number of ophthalmologists is seriously insufficient, which is unlikely to be solved in the short term. Many in the field view automatic identification of diabetic retinopathy as a key to the fundus screening workload dilemma. Clinically, severity of diabetic retinopathy is divided into 2 categories according to international clinical criteria, which is non-proliferatavie diabetic retinopathy (NPDR) and proliferatavie diabetic retinopathy (PDR). NPDR is further divided into stage I (mild NPDR), stage II (moderate NPDR), and stage III (severe NPDR); and PDR is divided into stage IV (early PDR), stage V (fibrous proliferation), and VI (advanced PDR). Earlier methods [1,2] used manual features to represent the fundus image, which do not generalize well and is easily affected by image quality, photo angle, and noise. Recently, CNN-based methods [3,4,5] have significantly improved the ability to detect diabetic retinopathy. However, most methods use general CNN architectures which impose same weighting for all locations in the fundus image. This sometimes leads to unsatisfying results, for example, Microaneurysm, a key

ISSN: 2664-9640

lesion feature to grading the severity of diabetic retinopathy, is small in size and is usually shown in less pixels than other type of nidus such as hemorrhage and hard exudation. It is thus more likely to be neglected in usual CNN architectures. Therefore, this paper introduces the attention mechanism model RAN (Residual Attention Network) that adopts different weights for different type of nidus pixels, which helps to balance the size of the nidus area and nidus importance.

Attention mechanism is a very representative example of information flow control in networks. The point is to imitate the attention mechanism of the human brain--When the human brain receives external information, it often does not process all of it, but only focuses on significant or interesting objects. In this way, the information is processed in a dynamic manner, making the model more adaptive to different objects in more challenging tasks, and the efficiency of the information processing is improved. In the field of deep learning technology, early research on attention mechanisms often used recurrent neural network (RNN) [6], or in joint mission of natural language applications and image recognition, such as image title generation. Since 2017, the attention mechanism has also gained popularity in the feedforward network for visual recognition tasks. PengYuxin [7] improved accuracy of object recognition by focusing attention on the parts of the object and the relationship between them. HuJie [8] added weights to the data in units of channels, and ZhangHang [9]made further efforts to by providing supervision of attention itself by separating spatial coverage and category judgments in segmentation tasks.

The work of Wang Fei [10] provided a new idea to use Attention to improve the performance of the model in classification tasks. Similar to residual network which sets up a direct connection path to allow the convolutional layer to automatically decide the degree of processing of the image according to the goal, residual attention sets parallel path outside the attention path to retain the original signal beside attention weighted result, and passed them altogether to the next layer. The network determines the extent to which attention-weighted results are used according to the task optimization goals. Therefore, the number of attention layers stacked in the overall network can theoretically be arbitrary large. At the same time, the attention module of the Residual attention network adopts the architecture of "encoding-decoding" [11,12], which integrates the global features of the space by downsampling and then upsampling, and this integration of global and Local information improves the accuracy of attention.

This paper proposes several improvements to the residual attention architecture, including adjusting the information processing ratio of the module trunk and attention weight channels, simplifying the compression depth of a single module for the overall architecture of multi-layer overlay, etc. The model training method was adjusted. The study used independent collected fundus retinal image data to analyze the ability of the new model design to recognize diabetic retinopathy through a series of experiments. It was also applied to different types of natural image data sets to verify its comprehensive performance.

2. Text Residual Attention Network

In the field of computer vision, attention models can be mainly classified as spatial locations-oriented (pixels, feature units) and channel-oriented. For example, the Squeeze-and-Excitation Network (SENet) works on channels, By modeling the dependencies between channels, it learns the weight (attention) of each channel, and then multiplies the weight to the features of the corresponding channel to get new features. For the attention model that works on spatial position , for the same channel, each pixel has different weights. For example, the Residual Attention Network proposed by Wang Fei et al (referred to as RAN in this paper) is one such example.

ISSN: 2664-9640

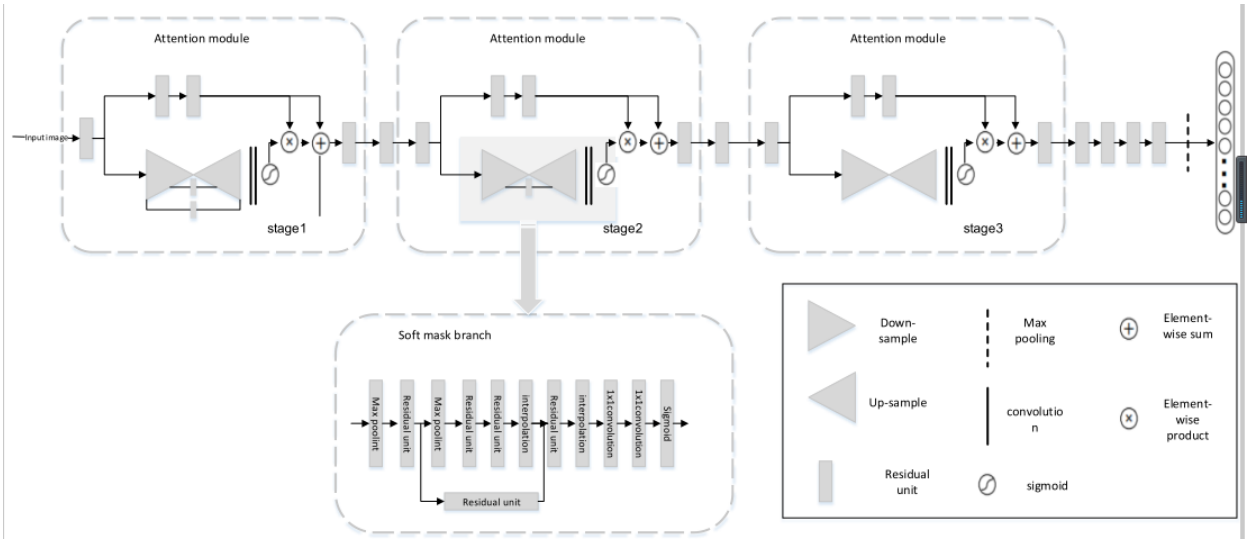

Figure 1:Architecture of the RAN

The RAN is composed of multiple attention modules as in Figure 1, and each attention module has two branches: a soft mask branch and a trunk branch. The trunk branch is used for feature extraction, and the soft mask branch is used to obtain feature importance, which is where the attention mechanism lies. the soft mask branch in the RAN adopts a bottom-up top-down architecture and is implemented in a "convolution-downsampling, then convolution-upsampling " process. Let the input of the attention module be x, the output of the trunk branch be T (x), and the output of the mask branch be M (x). The shapes of the two tensor matrices T (x) and M (x) are the same. The model is capable of giving high weight to "interesting" or important features, and giving lower weight to non- "interesting" or unimportant features to improve the efficiency of information processing. The attention on the result is implemented by dot multiplication (multiplying pixel by pixel) of the two tensors, so the output of the attention module H is:

$$H_{i,c}(x) = M_{i,c}(x) * T_{i,c}(x) \tag{1}$$

On the other hand, residual unit is used in the module. The residual architecture was proposed by He Kaiming et al. in 2015. It introduces shortcut connections between different layers in the forward feedback network. Shortcut connection is equivalent to performing identity mapping without adding additional parameters. For a residual unit, assuming that the input is x, the output is F (x) + x. F (x) is defined as the residual mapping, and the shortcut part is defined as the identity mapping. The parameters of such architectured network are easier to train because they are now more sensitive to loss. Even if the network stacked more layers than the optimal level, its performance will not decrease as the depth increases. In view of the advantages of identity mapping, the author modified formula (1) accordingly, so that the output of an attention module is

$$H_{i,c}(x) = (1 + M_{i,c}(x)) * T_{i,c}(x) \tag{2}$$

3. Tiny Residual Attention Network

The soft mask branch of the original version of RAN has a lot more learning parameters than the trunk branch, resulting in too much emphasis on learning the soft mask part instead of the trunk part for feature extraction. Therefore, this paper proposes the Tiny Residual Attention Network (TRAN) Model, the overall architecture is shown in Figure. 2. This model deletes the bottom-up top-down architecture, so that the soft mask branch consists of only two residual

ISSN: 2664-9640

units, which greatly reduces the number of network parameters. The optimized branch is named the tiny mask branch.

According to [1], each stage takes a different number of attention module stacks, and the corresponding network model can be subtyped to attention-56, attention-92, attention-128, attention-164. This paper mainly uses attention-92 among them for comparison. The improved network is named tiny-attention-92. For tiny-attention-92, p, q, and r are respectively 1, 2, and 3. The specific design composition of the network is shown in Table 1, which lists the types and attributes of each layer of the paper's use model.

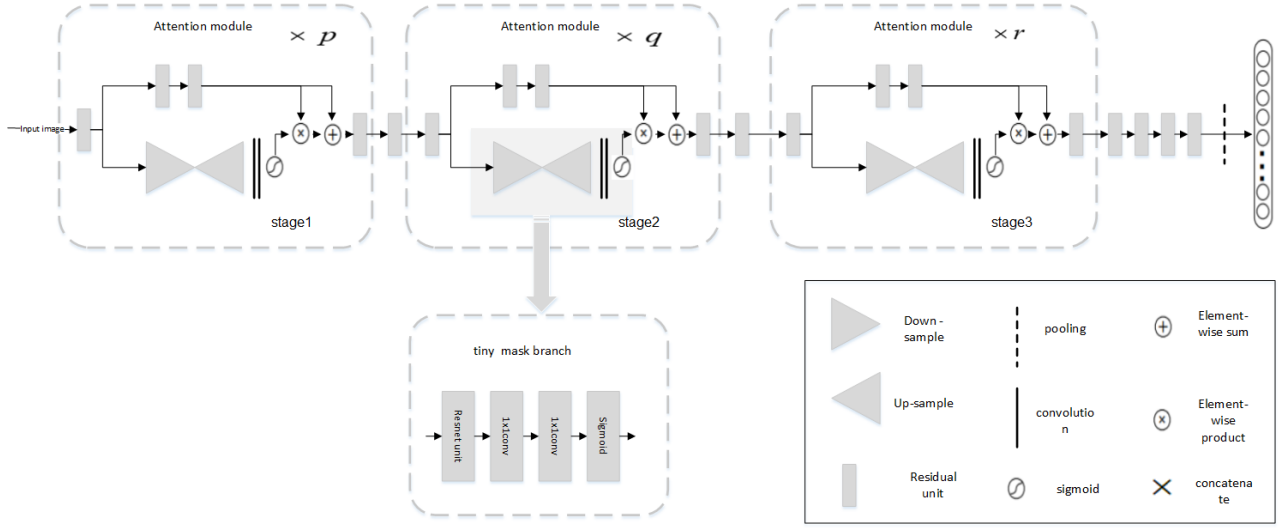

Figure 2: Architecture of the TRAN

Table 1: TRAN network architecture with 224x224 and 32x32 input sizes. From top to bottom, the layers of the network are listed from input to output.

zNetwork layer type	224x224 input size		32x32 input size	
	Output size	Attributes	Output size	Attributes
Conv1	112x112	7x7,64, stride 2	32x32	3x3, 32, stride 1
Max pooling	56x56	3x3, stride 2	/	/
Residual unit	56x56	$\begin{pmatrix}1\times1,\ 64\\3\times3,\ 64\\1\times1,\ 256\end{pmatrix}\times1$	32x32	$\begin{pmatrix}1\times1,\ 32\\3\times3,\ 32\\1\times1,\ 128\end{pmatrix}\times1$
Attention module	56x56	Attention x 1	32x32	Attention x 1
Residual unit	28x28	$\begin{pmatrix}1\times1,\ 128\\3\times3,\ 128\\1\times1,\ 512\end{pmatrix}\times1$	16x16	$\begin{pmatrix}1\times1,\ 64\\3\times3,\ 64\\1\times1,\ 256\end{pmatrix}\times1$
Attention module	28x28	Attention x 2	16x16	Attention x 2
Residual unit	14x14	$\begin{pmatrix}1\times1,\ 256\\3\times3,\ 256\\1\times1,\ 1024\end{pmatrix}\times1$	8x8	$\begin{pmatrix}1\times1,\ 128\\3\times3,\ 128\\1\times1,\ 512\end{pmatrix}\times1$
Attention module	14x14	Attention x 3	8x8	Attention x 3
Residual unit	7x7	$\begin{pmatrix}1\times1,\ 512\\3\times3,\ 512\\1\times1,\ 2048\end{pmatrix}\times3$	8x8	$\begin{pmatrix}1\times1,\ 256\\3\times3,\ 256\\1\times1,\ 1024\end{pmatrix}\times3$
Average pooling	1x1	7x7, stride 1	1x1	8x8, stride 1
FC, Softmax	Number of output categories			

ISSN: 2664-9640

4. Training method based on amplification of mixup data

In addition to optimization of the network architecture, this paper also explore the introduction of the mixup technology [13] to the training of this particular architecture for the target task. Mixup is a method of data amplification.The general purpose of data amplification is to build much more samples that can be actually used on the basis of available training data, thus increasing the diversity of data, reducing the possibility of over-fitting. From the perspective of sample space, data amplification can increase the coverage of sample distribution and make the sample-feature manifold recognized by machine learning more complete. Consider the case that distribution of test data and training data is not completely aligned, test data may fall in the niche that manifold distribution around training samples does not cover. Mixup can expand this distribution and lower the possibility of prediction bias caused by mismatch.

Different from ordinary augmentation methods that adjust non-semantic variables such as brightness, sharpness, and image angle, mixup instead attempts to augment the semantic axis by overlapping 2 samples, thus to strengthen the model's ability to identify the transition region of the points represented by the samples in the high-dimensional space, and the continuity and consistency of the sample manifolds modeled by the model [14]. Specifically, it mix the two samples on the input image and output label at the same time. The mixing is achieved by a simple linear superposition method, and the proportion of the two samples in the mixed result is controlled by the proportional coefficient. In theory, this method is equivalent to increasing the model's coverage of the connected area of two samples in the sample space, which can greatly increase the coverage of the full sample space, so that the model can better predict outside the training set,This is especially useful for samples that are ambiguous in appearance.

The specific process of Mixup is shown in Table 2.

Table 2: mixup process

Input: certain image pair$[(x_i, y_i), (x_j, y_j)]$
Onput: Fused sample$(\hat{x}, \hat{y})$,Used as training sample
1.Randomly generated weight coefficient:λ, $\lambda \sim B\text{eta}(\alpha, \alpha), \alpha \in (0, \infty)$
2. Fusion images, $\hat{x} = \lambda x_i + (1 - \lambda)x_j$
3. Fusion Notes, $\hat{y} = \lambda y_i + (1 - \lambda)y_j$
Xi represents the i-th picture; Yi represents the corresponding annotation, and is encoded by onehot

The weighting coefficient λ is drawn from a beta distribution at each training step, and the alpha coefficient of the beta distribution determines whether the sample is biased towards the two extremes of 0,1 or the middle position of 0.5. When α = 1, it corresponds to λ obeying uniform distribution.

5. Experiment

In order to verify the performance of the model, this paper evaluates the performance and effectiveness of the new network on four datasets, including three natural image datasets CIFAR-10, CIFAR-100 and ImageNet, and a medical image dataset.

We first evaluated the performance of TRAN architecture and mixup technology on the CIFAR-10 / 100 public dataset [15], and compared the results of RAN. Then, we compared CIFAR with other common network models such as resnet [17] and inception [18] on the ImageNet dataset [16], and explored its performance on large-scale image classification tasks. The number of classification categories in these three data sets is in different orders of magnitude (the number of output categories at the end of the network is 10, 100, 1000 respectively), which can cover different representative recognition task scenes.

ISSN: 2664-9640

Finally, this study also tested the performance of the network on the medical fundus image dataset to explore the pros and cons of TARN in different types of image types. To compare with attention-92 in RAN, this paper uses tiny-attention92 network for experiments. The overall network framework for the experiment is shown in Figure. 2. The network architecture details for the input picture sizes of 32x32 and 224x224 are shown in Table.1..

5.1. CIFAR10/100 DataSet

CIFAR10 and CIFAR100 are image data sets of objects in natural images. They contain 10 and 100 types of objects, respectively, each containing 60,000 pictures and corresponding category tags, of which 50,000 are for training sets and the remaining 10,000 are test sets. Pictures are 32x32 RGB-colored images.

5.1.1. Experimental scheme

This paper sets up two experimental schemes, using a tiny-attention-92 network to verify the effectiveness of the tiny mask branch proposed in this paper; tiny-attention-92_mixup is based on tiny-attention-92 and introduces mixup for data augmentation during training to verify the effect of mixup combined with tiny-attention-92. In order to make a relatively fair comparison, this paper retains most of the experimental settings of RAN [10]. The hyper-parameter settings of the two schemes are shown in Table 3.

Table 3: TRAN experimental hyperparameter settings on CIFAR10/100 datasets

parameter	attention92 Parameter value	TRAN92_mixup Parameter value
batch size	64	64
Epoch	320	320
Beta distribution	no	Set α in Beta distribution to 1.0
mixup	no	When the number of epochs is less than 300 times, new samples are obtained using the mixup technology for training. When the number of epochs is greater than 300 times, the mixup is canceled.
Optimizer	Using Nesterov SGD as the optimizer, the momentum is 0.9 and the weight decay coefficient is 0.0001	
Learning rate	The initial learning rate is 0.1, which is reduced to 1/10 at 96, 192, and 288 epoch.	
Data augmentation	Fill the four pixel layers around the image, with a fill value of 0, so as to get an image with a size of 40x40, randomly pick 32x32 image blocks from the image; flip around randomly with a 50% probability	

5.1.2. Experimental results

The performance of different network architectures on the CIFAR10/100 dataset is shown in Figure 3,Figure 4.

5.2. ImageNet

In order to explore the generalization ability of the proposed improved model TRAN, this paper uses the LSRRC 2012 dataset [13] for experiments. This data set is filtered from the ImageNet full data set and contains 1,000 classes, of which 12 million are in the training set, 50,000 are in the validation set, and 100,000 are in the test set. This paper conducts experiments on the validation set to evaluate the effectiveness of the TRAN proposed in this paper.

ISSN: 2664-9640

5.2.1. Experimental scheme

Because the ImageNet image is more complicated and the size is larger, the batch size here is changed to 40, and the total number of epoch is changed to 10 because the number of images in the LSVRC dataset is extremely large compared to the other two tasks so each epoch will impose a lot of training and it does not require too many epochs. The learning rate and mixup scheme were adjusted accordingly. For the parameter settings that are adjusted relative to Table. 3, see Table 4.

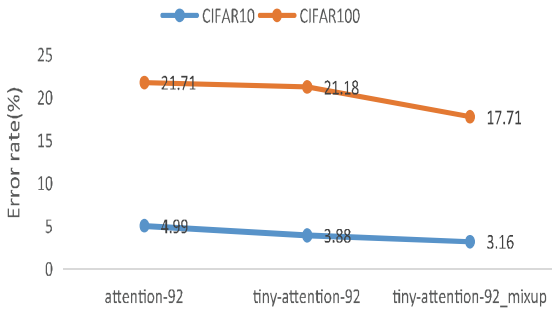

Figure 3: Different network architectures at CIFAR10/100 error rate

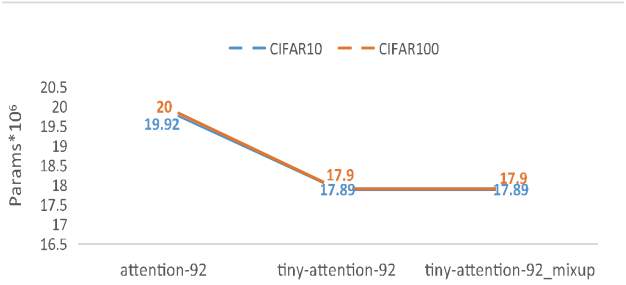

Figure 4: Number of parameters for different network architectures in CIFAR10/100

Table 4:. Experimental setup of LSRRC2012 data

Parameter	Parameter value
batch size	40
Epoch	10
Learning rate	The initial learning rate is 0.01, divided by 10 when 1, 2, 5 epochs are completed
Mixup (For TRAN92_mixup)	When epoch number is less than 5, new samples are obtained by using the mixup technology for training. When epoch number is more than 5, the mixup is cancelled.
Data augmentation	no

5.2.2. Experimental results

The performance of different network architectures on the ImageNet dataset is shown in Figure 5,Figure 6.

ISSN: 2664-9640

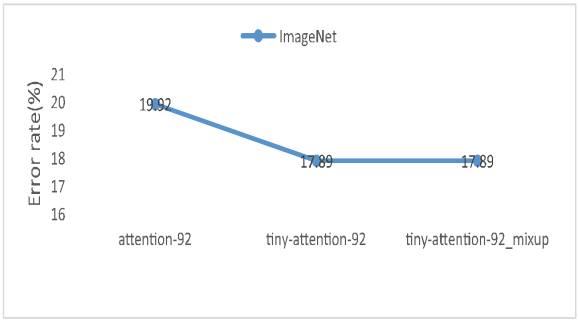

Figure 5: Different network architectures at ImageNet error rate

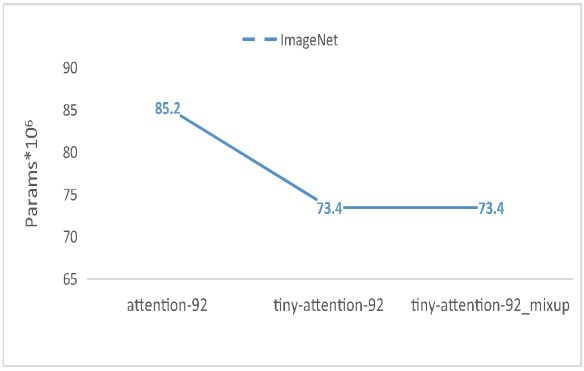

Figure 6: Number of parameters for different network architectures in ImageNet

5.3. Fundus image

This paper applies TRAN to the fundus image recognition of diabetic retinopathy. The performance of RAN on this data set is also evaluated as a comparison.

Fundus image is a common diagnostic image in ophthalmology. Special fundus camera is used to image the retina. Fundus imaging is an effective screening and diagnosis tool for common endocrine, blood, and physiological related diseases such as diabetic retinopathy, macular degeneration, and high myopia, so it has great diagnostic needs. Intelligent automatic recognition of fundus images is of great value in field ophthalmological screening. The anonymous fundus image data set related to diabetic retinopathy used in this paper is provided by Beijing Tongren Eye Hospital as a partner. The dataset includes 65599 fundus pictures, and the classification marks are divided into 5 types, which correspond to 0 (normal), 1 (mild NPDR), 2 (moderate NPDR), 3 (severe NPDR), 4 (PDR). the number of pictures of each type is shown in Figure 7

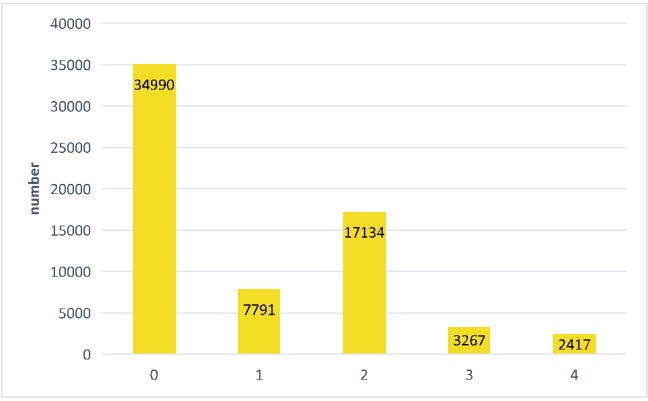

Figure 7 Distribution of diabetic retinopathy data

ISSN: 2664-9640

5.3.1. Experimental scheme

Firstly, the samples are randomly divided into training set, validation set and test set by the ratio of 0.8:0.1:0.1. Models were trained on the training set, and evaluated on the validation set every epoch, and select the model with the least loss on the validation set as the final model. The result of evaluating the model on the test set is the final experimental result.

Because of the highly uneven number of samples in each category, this paper adopts the following resampling strategy. First, the sample weight is defined according to the sample size of each category, and then the oversampling of the rare category is gradually reduced. Define the initial weight as $w_0 \in \Re^5$, The final weight is $w_n \in \Re^5$, Weight decay coefficient is $r \in [0,1]^1$, Therefore, when the e-th cycle is specified, the sampling weight of each sample is

$$w_e = r^{e-1}w_0 + (1 - r^{e-1})w_n \tag{3a}$$

In this experiment, we set:

$$w_0 = [\frac{max\_s}{s_0}, \frac{max\_s}{s_1}, \frac{max\_s}{s_2}, \frac{max\_s}{s_3}, \frac{max\_s}{s_4}] \tag{3b}$$

$$w_n = [1,2,2,2,2] \tag{3c}$$

$$\text{r=0.95} \tag{3d}$$

Where s_i is the number of samples of the i-th category,

$$max\_s = max(s_0, s_1, s_2, s_3, s_4) \tag{3e}$$

Considering the fact that misclassification of different types of sample is of different seriousness in medical scenario, for example, misclassifying level 3 diabetic retinopathy into level 0 is more serious than misclassifying 1 into 0, this paper introduces weight kappa as an evaluation index in addition to accuracy. The definition of weight kappa is shown in the formula:

$$wkap = 1 - \frac{\sum_{i=0}^{k}\sum_{j=0}^{k} w_{ij}x_{ij}}{\sum_{i=0}^{k}\sum_{j=0}^{k} w_{ij}m_{ij}} \tag{4}$$

Here, $w_{ij} = \frac{(i-j)^2}{(N-1)^2}$,N is the number of categories,here is 5, x_{ij} represents the number of categories i judged as j in all samples, m_{ij} represents the expectation that category i is judged as j.

For the sake of fairness, the training of attention-92 and tiny-attention-92 both use the hyper-parameter settings shown in Table 5 during the experiments in this paper.

Table 5: Parameter settings for training attention-92 and tiny-attention-92

Parameter	Parameter value
batch size	20
Optimizer	Using Nesterov SGD as the optimizer, the momentum is 0.9 and the weight decay coefficient is 0.0001
Epoch	100
Learning rate	Initialization learning rate is 0.1, divided by 10 at 30, 60, 9 epochs
Beta distribution	Set α in the Beta distribution to 1.0
mixup	When the epochs number is less than 85, new samples are obtained by using the mixup technology for training. When the epochs number is greater than 85, the mixup is cancelled.
Data augmentation	Scale the picture to 224x224; randomly rotate it at any angle; fill the outer periphery of the picture with 26 pixel layers, the value is 0 to get 276x276, and then randomly pick the 224x224 image block; flip left and right randomly

ISSN: 2664-9640

5.3.2. Experimental results

The performance of different network architectures in the fundus dataset is shown in Figure 8.

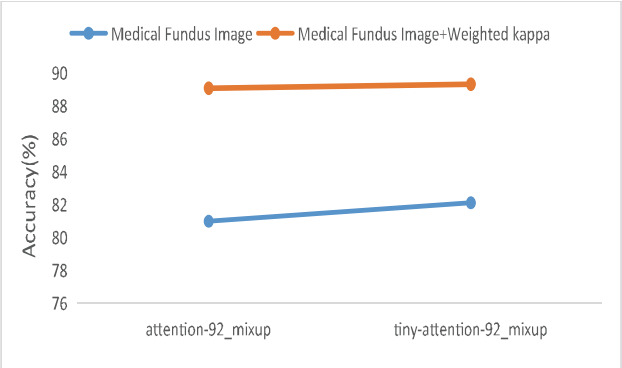

Figure 8:Different network architectures evaluation indicators

5.4. Comparison of experimental results

As can be seen from Figure 3, on the CIFAR-10 / 100 dataset where image size is relatively small (32 × 32) , the TRAN network is a standard res-attention model for comparison, and its error rate has decreased significantly, especially in CIFAR -10 dataset. The decrease in the number of model parameters may be one of the reasons of the performance improvement, as fewer parameters represent smaller learning burden and lower risk of overfitting. On the other hand, this advantage is not as big on CIFAR-100 as it is on CIFAR-10. One possible explanation is that tasks with a large number of classifications require models with greater learning capabilities, and the reduced learning capacity of TRAN offsets some of the benefit.

In addition to the improvement of the network architecture, the effect of mixup on the improvement of the results is also very obvious. There is a huge extra improvement in addition to the already optimized network architecture. This shows that mixup does achieve the expected increasing of sample space coverage to test samples.

Fig. 8 shows the performance of the new model on the fundus image dataset. The accuracy index is used here according to medicine practice. It can be seen that in the case where the mixup method is used, the accuracy of the new model TRAN has a significantly better, and the weighted kappa coefficient is also superior to the ordinary RAN model. The original size of the fundus image is larger than that of the CIFAR and ImageNet datasets. Even after compression to size of 224x224, it still has more detailed features. This result may indicate that the simplified architecture of TRAN does not affect the identification of image detail information. At the same time, compared with ImageNet data, the results show that even without the bottom-up top-down architecture of the original RAN model, the attention module is capable of obtaining the information necessary for classification, so the new architecture is proper for the task. It also implies that the key feature for diagnosis of diabetic retinopathy in the fundus image is relatively localized.

6. Conclusion

This paper proposes improvements to the RAN, specifically the corresponding improvements to the soft mask branch, which simplifies the model architecture. Relevant experiments on CIFAR10 / 100, ImageNet, and fundus image datasets show that the improved model TRAN has better performance over RAN. The new model has fewer parameters, and the generalization ability of the model obtained under the same training time and training scheme is better. For image classification tasks, it surpasses common image models and the original res-attention network in terms of recognition accuracy, although its advantages are not obvious on the

ISSN: 2664-9640

ImageNet dataset which has relatively larger image sizes and number of classification categories (hence requires larger network capacity) On the other hand, the experiment also explored the boundary of the res-attention network paradigm's capability, that the new model can handle larger input images, its advantages are only offset when the number of target classes increases. This can be confirmed by TRAN's performance on the fundus image dataset - new model is doing better than RAN on clssification task of professional images of similar size to the ImageNet samples. In addition, the reuslts also provide insight for the pruning and simplification of neural networks with attention mechanism, and it might be helpful for the application of network models in environments with small computing capacity such as mobile terminals, for example it can be used in medical industry applications such as smart fundus recognition cameras.

In the follow-up work, the authors will mainly explore the reduction of the amount of attention module calculation in structural design without reducing the receptive field range (using dilated convolution), increasing the attention mechanism on the channel domain, and so on, while continuing to explore The effectiveness of the TRAN network-based model in medical industry applications such as segmentation and detection of medical images.

Acknowledgements

Natural Science Foundation: Horizontal Scientific Research Project of Wenzhou Polytechnic in 2021(No:H2021111).

REFERENCES

[1] E. Chaum, T. P. Karnowski, V. P. Govindasamy, M. Abdelrahman, and K. W. Tobin: Automated diagnosis of retinopathy by content-based image retrieval, Retina, Vol. 28(2008) No. 10, p. 1463-147.

[2] M. D. Abràmoff, J. M. Reinhardt, S. R. Russell, J. C. Folk, V. B. Mahajan, M. Niemeijer, and G. X. Quellec: Automated early detection of diabetic retinopathy, OPHTHALMOLOGY-ROCHESTER AND HAGERSTOWN-, Vol. 117(2010) No. 6, p. 1147-1154.

[3] M. D. Abràmoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk, and M. Niemeijer:Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning, Investigative Opthalmology & Visual Science, Vol. 57(2016) No. 13, p. 5200-5206.

[4] T. Chandrakumar, and R. Kathirvel:Classifying diabetic retinopathy using deep learning architecture, International Journal of Engineering and Technical Research, Vol. 5(2016) No. 6, p. 19-24.

[5] V. Gulshan, L. Peng, M. Coram , M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, JAMA, Vol. 316(2016) No. 22, p.2401-2410.

[6] V. Mnih, N. Heess, A. Graves ,and K. Kavukcuoglu:Recurrent models of visual attention, Advances in Neural Information Processing Systems, Vol. 2(2014) No. 6, p. 2204–2212.

[7] Y. Peng, X. He ,and J. Zhao:Object-part attention model for fine-grained image classification, IEEE Transactions Image Process, Vol. 27(2018) No. 3, p. 1487-1500.

[8] J. Hu, L. Shen, and G. Sun:Squeeze-and-Excitation Networks, Conference on Computer Vision and Pattern Recognition (Salt Lake City, Utah, USA, Jun 2018), p. 7132-7141.

[9] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, T. Ambrish, and A. Amit:Context Encoding for Semantic Segmentation, Conference on Computer Vision and Pattern Recognition (Salt Lake City, Utah, USA, Jun. 2018), p.7151-7160.

ISSN: 2664-9640

[10] F. Wang,M. Jiang , C. Qian , S. Yang, C. Li, and H. Zhang:Residual Attention Network for Image Classification, Conference on Computer Vision and Pattern Recognition(Honolulu, Hawaii, USA, Jul. 2017), p.6450-6458.

[11] H. Noh, S. Hong , and B. Han:Learning Deconvolution Network for Semantic Segmentation, International Conference on Computer Vision(Santiago, Chile, Dec. 2015),p.1520-1528.

[12] O. Ronneberger, P. Fischer, and T. Brox:U-Net: Convolutional Networks for Biomedical Image Segmentation, Medical Image Computing and Computer Assisted Intervention Society,(Munich, Germany, Oct. 2015),p. 234-241.

[13] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. "mixup:Beyond Empirical Risk Minimization, " in Proc. ICLR, Vancouver, BC, Canada, May.2018, pp.1-13. [Online]. Available: https:// openreview.net/forum?id=r1Ddp1-Rb.

[14] V.Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, and Y. Bengiol:Manifold Mixup:Better Representations by Interpolating Hidden States, International Conference on Machine Learning,(Long Beach, California, USA, Jun. 2019), p.6438-6447.

[15] A. Krizhevsky, and G. Hinton:Learning multiple layers of features from tiny images,Computer Science Department, University of Toronto, Tech. Rep, Vol. 1(2009) No. 4, p. 126-133.

[16] J. Deng, W. Dong, R. Socher,L .Li, K. Li, and F. Li: ImageNet:A large-scale hierarchical image database,Conference on Computer Vision and Pattern Recognition(Miami, Florida, USA, Jun. 2009), p.248-255.

[17] K. He, X. Zhang, S. Ren, and J. Sun. :Deep Residual Learning for Image Recognition, Conference on Computer Vision and Pattern Recognition(Las Vegas, Nevada, USA, Jun. 2016), p. 770-778.

[18] C. Szegedy, S. Ioffe, V. Vanhoucke: Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning, American Association for Artificial Intelligence(San Francisco, CA, USA, Feb. 2017), p.4278-4284.