A Technical Report for VIPriors Image Classification Challenge

Xinran Song，Chang Liu，Wenxin He

Xidian University, Xi'an, China

15232388156@163.com

***Abstract*- Image classification has always been a hot and challenging task. This article is a brief report we submitted to VIPriors Image Classification Challenge. In this challenge, the difficulty lies in the small amount of data available for training, and the use of pre-training is prohibited. Following this idea, we focused on model optimization and data processing. We tried a large number of classic image classification models and performed extraction and enhancement training on several low-precision image categories. In our method, several strong backbones and multiple loss functions are used to learn more representative features. In order to improve the generalization and robustness of the model, effective image enhancement strategies such as autoaugment and cutmix are adopted. Finally, the method of model fusion is used to screen models with high reference value. Our team jzdjjz got the Top-1 accuracy rate of 0.69, next to 0.75, 0.74, 0.72 and 0.70.**

I.INTRODUCTION

The VIPriors Image Classification Challenge is one of "Visual Inductive Priors for Data-Efficient Computer Vision" challenges. The main goal of this challenge is to obtain the Top-1 Accuracy on the dataset. According to the official baseline to generate the dataset, the training and test sets are extracted from Imagenet 2012. Each dataset contains 1000 categories, and each category contains 50 pictures; the test set is taken from the verification segmentation of the Imagenet 2012 dataset. We can't use any pre-training during the process.

In recent years, many works have achieved great results on image sets and data sets. The groundbreaking ResNet model [1] introduced in 2016 has completely changed the world of deep learning. Then ResNeXt is an improvement of ResNet. It studies the influence of cardinality on CNN performance and the number of packets for ordinary convolution. In 2018, senet introduced the channel attention mechanism to weight the channels. Instead of introducing a new spatial dimension to fuse feature channels, it adopted a new "feature recalibration" strategy. Plug and play, the amount of parameters introduced is small, but the improvement is huge. The basic idea of DenseNet [2] is the same as ResNet, but it establishes the dense connection between all the front layers and the back layers, and realizes feature reuse through the connection of features on the channel, which has better performance than ResNet; PyramidNet [3] uses additive pyramid to gradually increase dimensions, and uses zero filling direct identity mapping, which makes the network wider and more accurate; Finally, with the same computational efficiency, the performance of ResNeSt [4] is better than all existing

ResNet variants, which can achieve a better speed accuracy tradeoff. This also gives us a lot of reference. In addition to the optimization of the model, the processing of data is another direction that can not be ignored. Data augmentation is an essential technology to improve the generalization ability of deep learning model. In addition to several strategies that have been proposed to automatically search for augmentation strategies from the data set, the performance of image recognition task is significantly improved by using AutoAugment[5], RandAugment [6] and other methods.

In a word, although the image classification model has performed well in Imagenet2012, the number of images in this challenge is greatly reduced compared with the original Imagenet2012 data set. Considering these factors, in our method, based on strong backbone, we use multiple loss function, data augmentation strategy and ensemble learning to improve classification performance.

II. METHOD

Model Architecture

We use PyramidNet，DenseNet161，and ResNeSt200 as our backbones. These models are introduced as follows.

1.PyramidNet:

As our backbone network, PyramidNet uses additive pyramid to gradually increase the dimension, and adds SE module. It adopts zero filling direct connected identity mapping, which makes the network wider and more accurate.

Compared with the traditional residual module, the dimension of each unit of the pyramid residual unit gradually increases until the down sampled residual unit appears. When the number of parameters is small, the performance of additive and multiplicative pyramid networks is basically the same, because there is no significant structural difference between the two network architectures. However, with the increase of the number of parameters, they begin to show more significant differences in the dimension configuration of the feature graph, and the additive pyramid will show better performance. Figure 1 shows the depth residual pyramid network structure. On the left of the figure below is a directly connected residual element with zero filled identity mapping; The figure on the right is an expanded representation of the figure on the left, which constitutes a residual network mixed with direct connection and ordinary network.

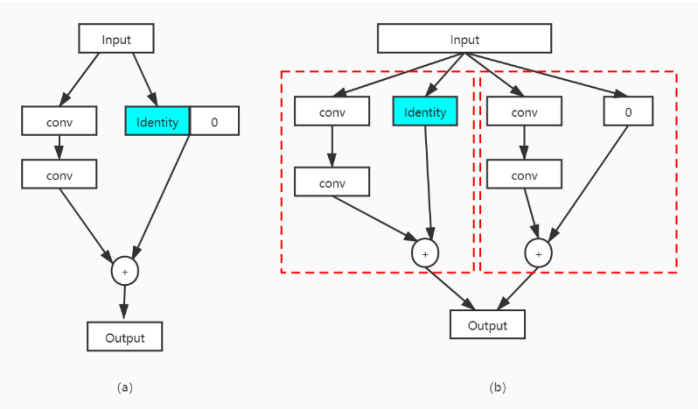

Figure 1.PyramidNet network structure

2.DenseNet:

A more radical dense connection mechanism is proposed based on densenet: connect all layers to each other. Specifically, each layer will accept all the layers in front of it as its additional input. For an L-layer network, densenet contains $\frac{(L+1)L}{2}$ connections. Moreover, our densenet directly concatenates feature maps from different layers, which can realize feature reuse and improve efficiency. Figure 2 shows DenseNet network structure.

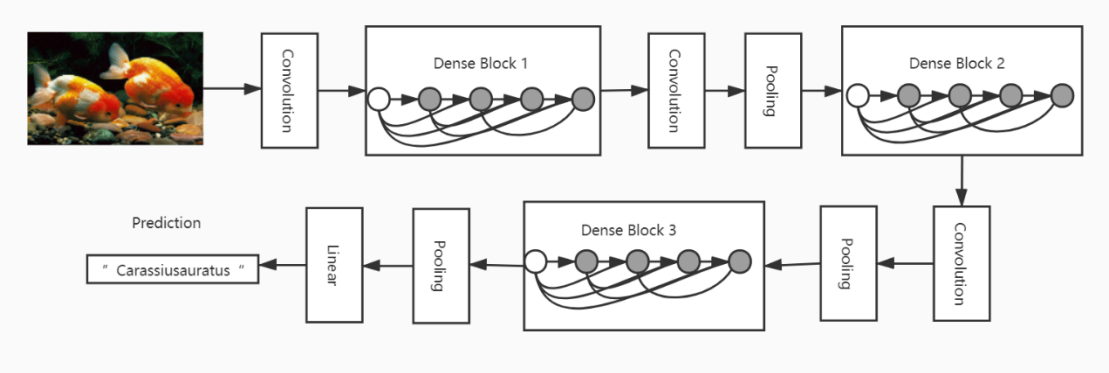

Figure 2.DenseNet network structure

3.ResNeSt:

The key part of ResNeSt is Split-Attention block. Split-Attention block is a computational unit consisting feature-map group and split attention operations.In order to enhance the compatibility of the model with small sample data, the width and depth of the network are increased. Grouping convolution makes the feature extraction more efficient. Figure 3 shows ResNeSt network structure.

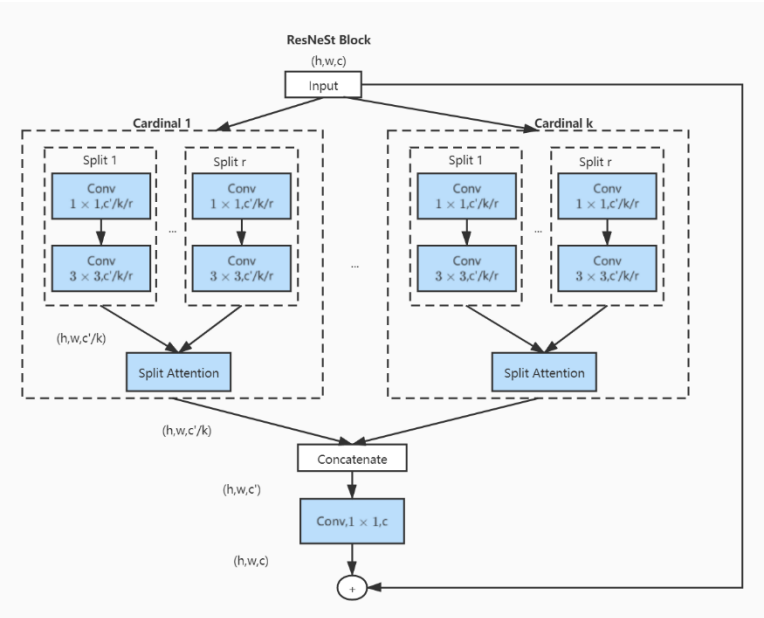

Figure 3.ResNeSt network structure

III. EXPERIMENT

Data Augmentation

Data augmentation can effectively prevent overfitting. Most of the enhancement channels used for training models are quite standard. Random rotation, horizontal flip, blur and scale change are all used, and these methods also improve the verification score. However, through accuracy comparison, it is found that there are categories that are difficult to classify, which greatly reduces the score. We call them stubborn images. The model needs a lot of time to learn to distinguish these categories, but obviously this is not enough. We noticed early that images with high color saturation or single category are really easy to recognize. And we want to increase more image accuracy. Therefore, the final model, which is also the model with the best performance, pays more attention to the stubborn image. The data augmentation strategies we used is described as follows.

1.Cutmix

After comparing mixup, cutout and cutmix, we found that cutmix can increase the accuracy to the greatest extent, so we finally used cutmix.

Cutmix cuts out part of the area, but does not fill in 0 pixels, but randomly fills in the area pixel values of other data in the training set, and the classification results are distributed according to a certain proportion; Cutmix enables the model to recognize two targets from the local view of an image and improve the efficiency of training. Figure 4 shows the original images and the image after cutmix.

Figure 4. The Original images and the image after cutmix.

2.Auto Augmentation

Auto Augmentation [7] is a method to search for data enhancement methods suitable for current problems. This method creates a search space for data enhancement strategy, and uses the search algorithm to select the data enhancement strategy suitable for a specific data set. A search which tries various candidate augmentation policies returns the best 24 best combinations. One of these 24 policies is then randomly chosen and applied to each sample image during training.

3. Stubborn Image Augmentation(SIA)

This is done separately for each color channel, assuming that it is Gaussian (not Gaussian in practice), and the mean and standard deviation are modified accordingly - basically reducing the red and blue channels. Then we also randomly stretch the contrast of each color channel. The stubborn image itself may be very diverse, and the fixed transformation cannot reflect this change. Because the performance of this model is very good, we also added a model that does not generalize and enhance all images, which greatly reduces the training time, so the result of this method looks very good.

4.Other tricks

(1) Linear scaling learning rate

With the increase of batch size, the speed of processing the same amount of data will be faster and faster, but more and more epochs are required to achieve the same accuracy. In other words, when using the same epoch, the verification accuracy of the model trained with large batch size will be reduced compared with the model trained with small batch size.

In mini batch SGD training, the value of gradient descent is random, because the data of each batch is randomly selected. Increasing the batch size will not change the expectation of the gradient, but will reduce its variance. In other words, large batch size will reduce the noise in the gradient, so we can increase the learning rate to speed up the convergence.

When the batch size is 256, the selected learning rate is 0.1. When we change the batch size to a large number B, the learning rate should be 0.1 × b/256。

(2) Label Smoothing

Full probability and 0 probability encourage the gap between the category and other categories to increase as much as possible. From the bounded gradient, it is difficult to adapt to this situation, which will cause the model to believe in the predicted categories too much. Label smoothing regulation is one of the effective methods to deal with this problem. Its specific idea is to reduce our trust in labels. For example, we can slightly reduce the target value of loss from 1 to 0.9, or slightly increase from 0 to 0.1. By adding noise to the label, the model constraint is realized and the over fitting degree of the model is reduced.

$$q_i = \begin{cases} 1-\varepsilon & \text{if } i = y, \\ \varepsilon/(K-1) & \text{otherwise,} \end{cases}$$

(3)CE loss

Cross entropy can measure the difference between two different probability distributions in the same random variable. In machine learning, it is expressed as the difference between the real probability distribution and the predicted probability distribution. The smaller the value of cross entropy, the better the prediction effect of the model.

$$H(p,q) = -\sum_{i=1}^{n} p(x_i) \log(q(x_i))$$

(4) Random image cropping and patching

Random image clipping and patching method(RICP), randomly cuts the middle parts of four pictures, then splices them into one picture, and mixes the labels of the four pictures at the same time.

As shown in the figure below, IX and IY are the width and height of the original picture. W and H are called boundary position, which determines the size of the four cropped small pictures. Distribution of W and H from beta（β， β）randomly generated

in, β is also a hyperparameter of RICAP. The size of the final spliced picture is consistent with the size of the original picture.

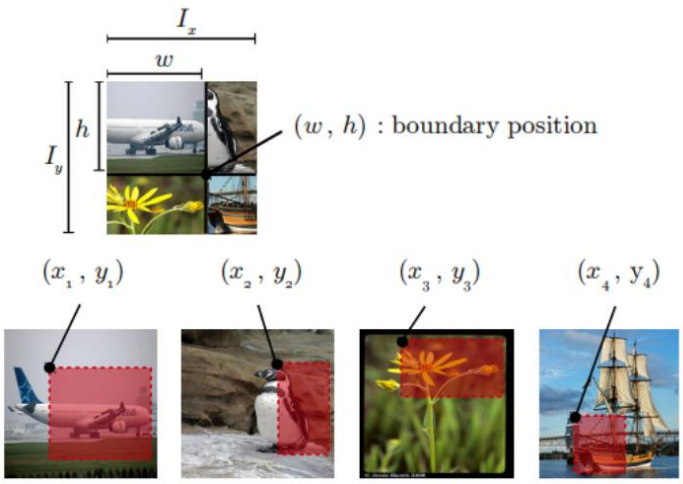

Figure 5. image after RICAP

5.Training Strategy

For the final submission of the competition, 6 models are used, which are DenseNet161+CE， PyramidNet+CE， Pyramid+SENet+CE， ResNeSt101+CE， ResNeSt101+CE+RICP，ResNeSt200e+CE+RICP.

SGD and Adam optimizers are used on different models, Batchsize of pyramid + senet, resnest101, resnest200e are 128, pyramidnet and densenet161 are 256, The learning rate is cosine learning rate decay, This warm-up [13] strategy is applied over the first epoch, gradually increasing the learning rate linearly from 0 to the initial value for the cosine schedule. The weight decay is set to 0.0001. All the models were implemented in pytorch. The models with CE loss was trained for 250 epochs. The models with Label Smoothing loss were trained for 200 epochs base on the weights trained using CE loss. This is because we found that it is hard to converge from scratch using Label Smoothing+CEloss loss.The data are the results with the Top-1 accuracy after many experiments.

6. Experimental Results

To reduce the overfitting phenomenon, The training set and the validation set are merged to train our model. In addition, no pretrained-models and any other data were used. The experiment results on test set are shown as Table 1 and Table 2.

backbone	Data Augmentation	Top-1 Acc（%）
DenseNet161	CutM+ AuA	54.16
PyramidNet	CutM+ AuA	54.21
SEPyramidNet	CutM+ AuA	59.61
SEPyramidNet	CutM+ AuA +SIA	62.31
ResNeSt101	CutM+ AuA	63.17
ResNeSt101	CutM+ AuA +SIA	65.23
ResNeSt200e	CutM+ AuA	64.03
ResNeSt200e	CutM+ AuA +SIA	65.71

Table 1. The experimental results of backbones.

backbone	Data Augmentation	lr	Label Smoothing	CE	RICP	Top-1 Acc
Densenet161	√			√		54.16
Densenet161	√	√		√		55.09
Densenet161	√	√	√			55.12
Densenet161	√	√	√	√		55.98
Densenet161	√	√	√	√	√	56.21
SEPyramidNet	√			√		62.31
SEPyramidNet	√	√		√		63.24
SEPyramidNet	√	√	√			63.29
SEPyramidNet	√	√	√	√		64.52
SEPyramidNet	√	√	√	√	√	64.54
ResNeSt101	√			√		65.23
ResNeSt101	√	√	√	√		66.83
ResNeSt101	√	√	√	√	√	66.91
ResNeSt200e	√			√		65.71
ResNeSt200e	√	√	√	√		65.89
ResNeSt200e	√	√	√	√	√	66.99

Table 2. Experimental results of adding tricks to each backbone network

By adding different tricks to the backbone network, the results show that the combination of label smoothing and CE loss can improve the accuracy of the model better. The five strategies are all effective strategies, and the accuracy increases in varying degrees. In the first mock exam, we combine 5 tricks strategies to train the same model to ensure the best accuracy and prepare the basic model for the latter model integration.

7.Ensemble Learning

Model ensemble learning can improve the accuracy of the results. We use the soft voting model fusion method to save the prediction array of each picture during the operation. Finally, our submission top-1 accuracy is 0.69.

IV. CONCLUSION

During the competition, based on the backbone networks DenseNet, PyramidNet and ResNeSt, try different tricks such as linear scaling learning rate, label smoothing + CE loss and RICP. Finally, for the characteristics of stubborn images, soft voting model ensemble is adopted to integrate all models, which effectively improves the final score.

References

[1] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[2]Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1492-1500.

[3] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132- 7141.

[4] Zhang H, Wu C, Zhang Z, et al. Resnest: Split-attention networks[J]. arXiv preprint arXiv:2004.08955, 2020.

[5] Cubuk E D, Zoph B, Mane D, et al. Autoaugment: Learning augmentation policies from data[J]. arXiv preprint arXiv:1805.09501, 2018.

[6] Cubuk E D, Zoph B, Shlens J, et al. Randaugment: Practical automated data augmentation with a reduced search space[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020: 702-703.

[7]Deng J, Guo J, Xue N, et al. Arcface: Additive angular margin loss for deep face recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 4690-4699.