Attention graph: Learning effective visual features for large-scale image classification

Xuelian Cui1, Zhanjie Zhang2, Tao Zhang2 D, Zhuoqun Yang2 and Jie Yang3

Abstract
In recent years, the research of deep learning has received extensive attention, and many breakthroughs have been made in various fields. On this basis, a neural network with the attention mechanism has become a research hotspot. In this paper, we try to solve the image classification task by implementing channel and spatial attention mechanism which improve the expression ability of neural network model. Different from previous studies, we propose an attention module consisting of channel attention module (CAM) and spatial attention module (SAM). 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. Besides, this module is lightweight and can be easily integrated into image classification algorithms. In the experiment, we combine the deep residual network model with the attention module and the experimental results show that the proposed method brings higher image classification accuracy. The channel attention module adds weight to the signals on different convolution channels to represent the correlation. 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. Through a large number of comparative experiments, our proposed model achieved outstanding performance.

Keywords
Image classification, channel attention module, spatial attention module, deep learning, residual network

Received 7 April 2021; Revised received 26 August 2021; accepted 22 November 2021

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 detection1–5, face recognition6–10, pose estimation11–13, population density estimation14,15, image segmentation16–19, etc. Therefore, image classification20–26 has always been a hot topic. 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.

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.

1Wuxi XINJE Electric Co.,Ltd, Wuxi, China;
2School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China;
3Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China

Corresponding author:
Zhanjie Zhang, School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China.
Email: 6191910041@stu.jiangnan.edu.cn
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\textsuperscript{27,28} for image classification tasks. The experimental result of VGGNet\textsuperscript{29-31} shows that the block with the same shape can obtain better classification accuracy by constructing deeper convolutional neural network model. With the same idea, the deep residual network is constructed by cross-layer connection method and it achieves higher classification accuracy. GoogLeNet\textsuperscript{32} 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\textsuperscript{33} and Xception\textsuperscript{34} 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 human vision. When we look at the scene around us, we always focus our attention on the main things to get key information. 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.

Wang fei et al. propose a novel model based on the attention mechanism\textsuperscript{35} 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\textsuperscript{36-43}. HU J et al. proposed the SENet\textsuperscript{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. The contributions of this paper are mainly in the following two aspects.

(i) We propose a large-scale image classification algorithm. By combining RestNet and attention module, our proposed algorithm obtains lower Top-1 error and Top-5 error performance on Image-1K and Cifar-100 respectively.

(ii) We introduce channel attention module and spatial attention module. To verify the order in which module is used first, we experiment much ablation study. The result confirms that the channel attention module should be used before the spatial attention module.

**Channel and spatial attention module**

For the convolutional neural network model, depth, width and attention mechanism\textsuperscript{45-49,51} are the main factors affecting the accuracy of image classification. At present, attention mechanism contains CAM and SAM. CAM acts on the channel domain, weighting different channel features. For a $C \times H \times W$ feature graph, the C weight of channel attention is different, while the weight of $H \times W$ is the same. For CAM, the weight of each C on different channel dimensions needs to be learned. To reduce the amount of computation and improve classification accuracy, the pooling layer in the general convolutional neural network directly uses the maximum pooling method or average pooling method to compress the image information. For SAM, only the key information in the spatial features is extracted. We first introduce the general framework of proposed model with CAM and SAM in this section. Finally, we describe how to pull them together.

After convolution operation, the intermediate feature map, $F_{\text{in}} \in \mathbb{R}^{C \times H \times W}$, is obtained. $F_{\text{in}}$ represents the input of model. After implementing the channel attention module, one-dimensional channel map $M_{c} \in \mathbb{R}^{C \times 1 \times 1}$ is inferred. Similarly, after implementing the spatial attention module, two-dimensional spatial map $M_{s} \in \mathbb{R}^{1 \times (H \times W) \times (H \times W)}$ is inferred. The proposed model is shown in Figure 1. The entire attention calculation process can be summarized as follows.

\begin{equation}
F_{1} = M_{c}(F_{\text{in}}) \odot F_{\text{in}} \tag{1}
\end{equation}
\begin{equation}
F_{\text{out}} = M_{s}(F_{1}) \odot F_{1} \tag{2}
\end{equation}

where $\odot$ stands for element-wise multiplication. After $F_{\text{in}}$ passing the channel attention module, $F_{1}$ is achieved. $F_{\text{out}}$ is the final output and the attention value is broadcasted during multiplication process. Attention and feature are the
multiplication of element levels, which will be propagated automatically, that is, channel attention broadcasts along spatial dimension, while spatial attention broadcasts along channel dimension.

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. 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 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.

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

where $F_{avg} \in \mathbb{R}^{C \times 1 \times 1}$ 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 multi-channel feature instead of a hot spot form. We describe the channel attention map $M_c$ as follows.

$$M_c = \sigma(W_2ReLU(W_1F_{avg}))$$

where $W_1 \in \mathbb{R}^{C \times C}$, $W_2 \in \mathbb{R}^{C \times C}$ and $\sigma$ represents 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.

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.
Figure 3. The details of spatial attention module.

Figure 4. Diagram of the visualization results.
A ∈ R^{C×H×W} represents the input of the SAM. After conducting convolutional layers, A generates feature maps B, C and D where \( B, C, D ∈ R^{C×H×W} \). We reshape B and C to \( R^{C×(H×W)} \) and \( H \× W \) represents pixels in spatial module. Then we get \( R^{H×W} \times (H × W) \) by operating a matrix multiplication between the transpose of B and C. We get the spatial attention map \( M_s \) when a softmax layer is applied. \( M_s \) is computed as follows.

\[
M_{sij} = \frac{\exp(B_i \times C_j)}{\sum_{i=1}^{H×W} \exp(B_i \times C_j)}
\]  

(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.

**Experiments**

The dataset is used in the experiment including ImageNet-1K and Cifar-100\(^5\). The Cifar-100 contains 100 classes and every class has 600 color images. But every images is only a size of \( 32 \times 32 \). 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 5, we extracted the images from Cifar-100 as the visual example.

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 6 shows the visual examples of the ImageNet-1K.

**Ablation studies about different attention module**

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 7.
From Figure 7, it is easily concluded that the ResNet-50 model with CSM (channel attention module and spatial attention module) 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.

Ablation studies about the order of attention module

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 8 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.

Experiment condition and computational complexity

In this paper, we verify our proposed algorithm on a server with 8 NVIDIA Tian X. With the deeper of the residual network, the better our proposed algorithm performs. ResNet-50 obtains the

### Table 1. Comparison of different module combination order on the ImageNet-1K dataset.

| Architecture          | Top-1 error(%) |
|-----------------------|----------------|
| ResNet-50             | 24.55          |
| ResNet-50+SAM         | 23.47          |
| ResNet-50+CAM         | 23.21          |
| ResNet-50+CSM (CAM+SAM) | 22.78        |
best performance. Thus, we give the time consumption of ResNet-50. For a single pass forwards and backwards, ResNet-50 takes 190 ms with a training minibatch of 256 images. SE-ResNet-50 takes 209 ms. BAM-ResNet-50 takes 195 ms. CBAM-ResNet-50 takes 210ms and our proposed method takes 215 ms. Furthermore, it can be found that the larger the batch size, the higher the stability of the experiment data. To avoid the uncertainty of the parameters in the training model, the experimental data (classification error and classification accuracy) was recorded by mean.

**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.

**Image Classification on CIFAR-100**

Based on Cifar-100, we carry out image classification experiment to verify the effectiveness of the CSM. ResNet and

---

Table 2. Results of the experimental on the ImageNet-1K dataset.

| Architecture               | Top-1 error(%) | Top-5 error(%) | average(%) | total(%) |
|----------------------------|----------------|----------------|------------|----------|
| ResNet-18                  | 29.62          | 10.56          | 20.09      | 40.18    |
| ResNet-18+SE44             | 29.43          | 10.24          | 19.8       | 39.67    |
| ResNet-18+CSM(Ours)        | 29.29          | 10.12          | 19.7       | 39.41    |
| ResNet-18+BAM51           | 28.88          | 10.01          | 19.45      | 38.89    |
| ResNet-18+CBAM46          | 29.27          | 10.09          | 19.68      | 38.36    |
| ResNet-34                  | 26.71          | 8.62           | 17.66      | 35.33    |
| ResNet-34+SE44            | 26.16          | 8.37           | 17.26      | 34.53    |
| ResNet-34+CSM(Ours)       | 26.03          | 8.28           | 17.15      | 34.31    |
| ResNet-34+BAM51           | 26.02          | 8.33           | 17.12      | 34.35    |
| ResNet-34+CBAM46          | 25.99          | 8.24           | 17.11      | 34.23    |
| ResNet-50                  | 24.55          | 7.52           | 16.0       | 32.07    |
| ResNet-50+SE44            | 23.21          | 6.74           | 14.97      | 29.95    |
| ResNet-50+CSM(Ours)       | 22.78          | 6.57           | 14.67      | 29.35    |
| ResNet-50+BAM51           | 24.02          | 7.18           | 15.6       | 31.2     |
| ResNet-50+CBAM46          | 22.66          | 6.31           | 14.49      | 28.97    |
| WideResNet18 (widen=1.5)  | 26.86          | 8.91           | 17.88      | 35.77    |
| WideResNet18 (widen=1.5)+SE46 | 26.23      | 8.51           | 17.37      | 34.74    |
| WideResNet18 (widen=1.5)+CSM(Ours) | 26.14 | 8.49           | 17.31      | 34.63    |
| WideResNet18 (widen=1.5)+BAM51 | 27.31 | 9.65           | 18.48      | 36.96    |
| WideResNet18 (widen=1.5)+CBAM46 | 26.10 | 8.43           | 17.26      | 34.53    |
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.

Conclusion

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 dimension. 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 deep neural network’s ability of expression. Besides, the width and depth of the neural networks are also worth considering.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is supported by the National Natural Science Foundation of China (NO. 61702226); the 111 Project (B12018); the Natural Science Foundation of Jiangsu Province (NO. BK20170200); the Fundamental Research Funds for the Central Universities (NO. JUSRP11854, NO. JUSRP11851).

ORCID iD

Tao Zhang https://orcid.org/0000-0003-3966-4691

References

1. 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.
2. Shaqing Ren, He Kaiming, Girshick Ross et al. and Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Analysis And Machine Intelligence 2017; 39: 1137–1149.
3. Ross Girshick, Donahue Jeff, Darrell Trevor et al., Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In: IEEE conference on computer vision and pattern recognition (CVPR), 2014, pp. 580–587.
4. Shu Zhibiao and Guo Yangqing. Algorithm on Contourlet Domain in Detection of Road Cracks for Pavement Images. In: International symposium on distributed computing and applications to business, engineering and science (DCABES), 2010, pp. 518–522.
5. Xiao Ke and Guolong Chen. Automatic Image Annotation Based on Co-Training. Journal of Algorithms and Computational Technology 2014; 8(1): pp. 1748–3018.8.1.1.
6. Hu g, Yang YX, Yi D, et al. When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition. In: International conference on computer vision (ICCV), December 2015, pp. 142–150. Santiago, ST: IEEE.
7. Wright J, Yang AY, Ganesh A et al. Robust face recognition via sparse representation. IEEE Trans Pattern Analysis And Machine Intelligencetext 2009; 31: 210–217.
8. Schroff Florian, Kalenichenko Dmitry and Philbin James. FaceNet: A Unified Embedding for Face Recognition and Clustering. In: IEEE conference on computer vision and pattern recognition (CVPR), 2015, pp. 815–823.
9. Parkhi Omkar M, Vedaldi Andrea and Zisserman Andrew. Deep Face Recognition. In: British machine vision conference (BMVC), 2015, pp. 516–529.
10. Tian Wu, Qi Sun Xiao and Xiaojun Wu. A Face Recognition Algorithm Based on Contextual Constraints Generalized Two-Dimensional FLD. Journal of Algorithms and Computational Technology, 2014, pp. 8.193-202.10.1260/1748-3018.8.2.193.
11. CAO Z, SIMON T, WEI S, et al. Realtime multi-person 2D pose estimation using part affinity fields. In: IEEE conference on computer vision and pattern recognition (CVPR), July 2017, Honolulu, HI, USA, pp. 2181–2188.

| Architecture | Accuracy (%) |
|--------------|--------------|
| ResNet-18    | 91.7         |
| ResNet-18+SE | 91.9         |
| ResNet-18+CSM (Ours) | 93.1 |
| ResNet-18+BAM | 93.4         |
| ResNet-18+CBAM | 93.2         |
| ResNet-34    | 92.4         |
| ResNet-34+SE | 92.7         |
| ResNet-34+CSM (Ours) | 93.8 |
| ResNet-34+BAM | 94.0         |
| ResNet-34+CBAM | 94.0         |
| ResNet-50    | 92.9         |
| ResNet-50+SE | 93.2         |
| ResNet-50+CSM (Ours) | 94.3 |
| ResNet-50+BAM | 94.5         |
| ResNet-50+CBAM | 94.4         |
| WideResNet18 (widen=1.5) | 92.8 |
| WideResNet18 (widen=1.5)+SE | 93.2 |
| WideResNet18 (widen=1.5)+CSM (Ours) | 94.2 |
| WideResNet18 (widen=1.5)+BAM | 94.5 |
| WideResNet18 (widen=1.5)+CBAM | 94.4 |
12. 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. 1297–304.

13. Newell, Alejandro, Kaiyu Yang, and Jia Deng. Stacked Hourglass Networks for Human Pose Estimation. In: Computer Vision-ECCV, 2016, pp. 483–9.

14. Li Zhang. Population Density Estimation Method Based on Convolutional Neural Network, 2016.

15. Do C, Waples RS, Peel D et al. Re-implementation of soft anomal detection based on wavelet denoising and support vector regression. Journal of Algorithms and Computational Technology, 2016; 10(4): 244–264. 10.10.1177/1748301816660406.

16. Kaiming He, Gkioxari Georgia, Dollar Piotr et al. Mask R-CNN. IEEE Trans Pattern Analysis And Machine Intelligence 2020; 42: 386–387.

17. Ronneberger, Olaf, Philipp Fischer and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: International conference on medical image computing and computer-Assisted intervention (MICCAI), 2015, pp. 234–1.

18. Jianbo Shi and Jianbo J. Normalized Cuts and Image Segmentation. In: IEEE Trans Pattern Analysis And Machine Intelligence 2000; 22: 888–905.

19. Williams Bryan, Spencer Jack, Chen Ke, Zheng Yalin and Harding Simon. An effective variational model for simultaneous reconstruction and segmentation of blurred images. Journal of Algorithms and Computational Technology, 2016; 10(4): 244–264. 10.10.1177/1748301816660406.

20. Muhammad, Usman, Weiqiang Wang, Shahbaz Pervaiz Chatttha and Sajid Ali. Pre-Trained VGGNet Architecture for Remote-Sensing Image Scene Classification. In: International conference on pattern recognition (ICPR), 2018. pp. 1622–1627.

21. Alex Krizhevsky, Sutskever Ilya and Hinton Geoffrey E. ImageNet classification with deep convolutional neural networks. Commun ACM 2017; 60: 80–84.

22. Kipf Thomas N and Welling Max. Semi-Supervised Classification with Graph Convolutional Networks. In: International conference on learning representations (ICLR), 2016, pp. 1812–1821.

23. Cireşan Dan, Meier Ueli and Schmidhuber Juergen. Column Deep Neural Networks for Image Classification. In: IEEE conference on computer vision and pattern recognition (CVPR), 2012, pp. 3642–3649.

24. Redmon Joseph and Farhadi Ali. YOLO9000: Better, Faster, Stronger. In: IEEE conference on computer vision and pattern recognition (CVPR), 2017. pp. 6517–6525.

25. Cireşan Dan, Meier Ueli and Schmidhuber Juergen. Multi-Column Deep Neural Networks for Image Classification. In: IEEE conference on computer vision and pattern recognition (CVPR), 2012. pp. 3642–3649.

26. Jian Wu, Zhiming Cui, Shi Yujie et al. Traffic flow anomaly detection based on wavelet denoising and supporting vector regression. Journal of Algorithms and Computational Technology 2013; 7(2): 209–225.

27. LeCun Y, Bottou L, Bengio Y et al. Gradient-based learning applied to document recognition. Intelligent Signal Processing 2001; 86(11): 2278–2324.

28. He K, Zhang X, Ren S et al. Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition (CVPR), 2016, pp. 770–778.

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

30. Sengupta Abhrionl, Ye Yuting, Wang Robert et al. Going deeper in spiking neural networks: VGG and residual architectures. Neuroscience 2009; 13: 95–105.

31. Zhou Shuren, Liang Wenlong, Li Junguo et al. Improved VGG model for road traffic sign recognition. Cmc-computers Materials Continua 2018; 57: 11–14.

32. Szegedy C, Liu W, Jia Y et al. Going deeper with convolutions. In: IEEE conference on computer vision and pattern recognition (CVPR), 2015, pp. 2182–2190.

33. Xie S, Girshick R, Dollar P et al. Aggregated residual transformations for deep neural networks. 2016. arXiv preprint arXiv: 1611.05431.

34. Chollet Francois. Xception: Deep Learning with Depthwise Separable Convolutions. In: IEEE conference on computer vision and pattern recognition (CVPR), 2017, pp. 1800–807. arXiv preprint arXiv: 1610.02357.

35. Wang Fei, Jiang Mengqing, Qian Chen etc. Residual Attention Network for Image Classification. In: IEEE conference on computer vision and pattern recognition (CVPR), 2017, pp. 6450–6458.

36. Yang Kaiyu, Qinami Klint, Fei-Fei Li et al. Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy. In: Conference on fairness, accountability and transparency (FAT), 2020. pp. 1021–1027.

37. Russakovsky Olga, Deng Jia, Su Hao et al. ImageNet large scale visual recognition challenge. International journal of computer vision (IJCV) 2015; 115: 211–252.

38. Deng J, Russakovsky O, Krause J et al. Scalable multi-label annotation. In: ACM conference on human factors in computing (CHI), 2014, pp. 811–817.

39. Russakovsky O, Deng J, Huang Z et al. Detecting avocados to zucchinis: what have we done, and where are we going, In: IEEE international conference on computer vision (ICCV), 2015. pp. 906–914.

40. Deng J, Berg A, Li K et al. What does classifying more than 10, 000 image categories tell us? In: Computer Vision-ECCV 2010, pp. 3103–3110.

41. Russakovsky O and Fei-Fei L. Attribute Learning in Large-scale Datasets. In: Computer Vision-ECCV, 2010. pp. 708–719.

42. Deng J, Dong W, Socher R et al. ImageNet: A Large-Scale Hierarchical Image Database. In: IEEE international conference on computer vision and pattern recognition (CVPR), 2009, pp. 1042–1050.

43. Deng J, Li K, Do M et al. Construction and Analysis of a Large Scale Image Ontology. In: Vision Sciences Society (VSS), 2009; 19: 711–719.

44. HU J, SHEN L and SUN G. Squeeze-and-excitation networks. In: IEEE conference on computer vision and pattern recognition (CVPR), June 2018, pp. 7132–7141. New York, USA. NY: IEEE.

45. Zhang Yulun, Li Kunpeng, Li Kai et al. Image Super-Resolution based on wavelet denoising and support vector regression. Journal of Algorithms and Computational Technology 2013; 7(2): 209–225.

46. Tran Sanghyun, Park Jongchan, Lee Joon-Young et al. Using Very Deep Residual Channel Attention Networks. In: Computer Vision-ECCV 2010, pp. 294–310.

47. Shin Sanghyun, Park Jongchan, Lee Joon-Young et al. CBAM: Convolutional Block Attention Module. In: Computer Vision-ECCV, 2010. pp. 3–9.
47. Chen Long, Zhang Hanwang, Xiao Jun et al. SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning. In: IEEE international conference on computer vision and pattern recognition (CVPR), 2017. pp. 6298–6306.

48. Zhao Hengshuang, Zhang Yi, Liu Shu et al. PSANet: Point-Wise Spatial Attention Network for Scene Parsing. In: Computer Vision-ECCV, 2018. pp. 267–273.

49. Mitchell Jude F, Sundberg Kristy A and Reynolds John H. Spatial attention decorrelates intrinsic activity fluctuations in macaque area V4. Neuron 2009; 63: 879–888.

50. Krizhevsky and Alex. Learning Multiple Layers of Features from Tiny Images, 2009.

51. Park J, Woo S, Lee JY, et al. BAM: Bottleneck Attention Module[C]//British Machine Vision Conference (BMVC). British Machine Vision Association (BMVA), 2018, pp. 981–996.