Attention-based CNNs for Image Classification: A Survey

Attention-based CNNs for Image Classification: A Survey

Menghua Zheng[1,2], Jiayu Xu[3], Yinjie Shen[4], Chunwei Tian[1,2,*], Jian Li[1,2], Lunke Fei[5], Ming Zong[6], Xiaoyang Liu[7]

[1] School of Software, Northwestern Polytechnic University, Xi'an, China

[2]Yangtze River Delta Research Institute of NPU, Taicang, China

[3] Shenzhen Research Institute, Guangdong Databeyond Technology Co., Ltd, Shenzhen, Guangdong, China

[4] China Mobile (Suzhou) Software Technology Co., Ltd., China

[5] School of Computers, Guangdong University of Technology, Guangzhou, 510006, China

[6] School of Electronics Engineering and Computer Science, Peking University, Beijing, China

[7]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China

[*]email: chunweitian@nwpu.edu.cn

Abstract. Deep learning techniques as well as CNNs can learn power context information, they have been widely applied in image recognition. However, deep CNNs may reply on large width and large depth, which may increase computational costs. Attention mechanism fused into CNNs can address this problem. In this paper, we summary an attention mechanism acts a CNN for image classification. Firstly, the survey shows the development of CNNs for image classification. Then, we illustrate basis of CNNs and attention mechanisms for image classification. Next, we give the main architecture of CNNs with attentions, public and our collected datasets, experimental results in image classification. Finally, we point out potential research points, challenges attention-based for image classification and summary the whole paper.

1. Introduction

Deep networks can mine more accurate information to express images, AlexNet [1] has obtained huge success on the ImageNet in 2012. Subsequently, due to flexible architectures, convolutional neural networks (CNNs) are extended to image recognition [2]. Besides, with the improvement of computer power from graphics processing unit (GPU), varying network architectures are presented in image recognition, which is divided into two kinds, increasing depth and width of CNNs [3]. In terms of increasing the depth of CNNs, VGG [4] used stacked small convolutional kernels to increase perception field for extracting more accurate information in image classification. Alternatively, GoogLeNet [5] utilized convolutional kernels of different types to increase the width for improving the performance in image classification. Although mentioned methods are competitive in image recognition, they still suffered from the drawbacks[6]. Firstly, deeper CNNs are easily faced with gradient explosion or gradient vanishing[6]. Secondly, wider CNNs may cause overfitting

phenomenon[6]. To prevent the phenomenon, a ResNet was proposed[7]. That is, it fused outputs of current layer and previous layers as input of next layer to enhance the memory ability in image classification, which improved the performance at no extra cost in image classification[7]. To reduce the computational costs, CNNs based attention are used in image classification[8]. There are less literatures to summary CNNs based attention for image classifications. In this paper, we conduct a survey to classify these literatures, which can make readers easily understand their principles. Firstly, it gives the development of CNNs for image classification. Subsequently, we introduce basis of popular CNNs and attention techniques in image classification. Then, we show the main architectures of CNNs with attentions, experimental settings, experimental results in image classification. Finally, we point out potential research points, challenges attention-based for image classification and sum up the whole paper.

2. Related work

It is known that CNNs with attentions have obtained excellent performance in image classification. Thus, understanding CNNs and attentions techniques are essential to improve these methods for improving the classification results. Thus, we introduce popular CNNs and main attentions techniques in image classification in this section.

2.1 Popular CNNs for image classification

According to previous illustrations, it is known that residual networks are very effective in image classification[7]. Also, wider CNNs can extract complementary information to boost the classification results [9]. Inspired by that, a residual architecture is used to expend the width for obtaining robust information to recognizes images[9]. For instance, fusing ResNet and branch of convolutions with 3×3 and 1×1 to mine representative information in image classification[9]. ResNeXt used homogeneous and multi-branch architecture to represent the classified image[10]. Besides, using residual network as a component gathers into a CNN can improve generalization ability of a classifier[11]. Utilizing multi-scale and residual blocks to fuse different semantic information was a good tool for image classification[12]. Using residual learning techniques to fuse hierarchical features can enhance memory ability of a deep CNN in image classification[13]. To deal with insufficient samples, generative adversarial networks (GAN) employed a generative network to generate similar samples, according to given training samples[14]. Then, GAN used a discriminative network to judge truth of all the training samples for training robust classifier[14]. Besides, graph convolutional networks are very effective in multi-label image recognition[15].

2.2 Attention mechanisms for image classification

Because deep CNNs depend on deep or wide architectures, they may have huge computational cost. To address this issue, an attention method is developed[16]. That is, the attention method uses obtained features of different parts of a network as weights to act other parts for learning more substantial sequential information. Current attention methods can be divided into two kinds: channel attention[17] and spatial attention methods[8]. Specifically, channel attention method emphasizes effects of channel features on the whole CNN. The second attention method treats pixels of all dimensions at the same location as a whole and its weight is learned via each pixel at each location. All weights from a spatial attention matrix. The mentioned attention mechanisms can solve the problem from different perspectives, which can give readers inspiration.

3. Attention-based CNNs for image classification

According to illustrations of Section 2.1 and Section 2.2, we can see that residual network and attention methods are effective for image classification. Inspired by that, scholars combined an attention mechanism into a residual network [18]. That is divided into five kinds: single-path attention [18], multi-path attention [19], channel attention [17], spatial attention [8] and the combination of channel attention and spatial attention [20]. Specifically, a single-path attention method mainly uses its

current layer to guide previous for extracting salient features in image classification. Multi-path attention method used a path to adjust other a path for image classification. Channel attention method (CA) [17] enhanced the effects of different channels to improve the classification results. Spatial attention method [8] used mean and max pooling operations to extract useful information for image classification as illustrated in Fig.1. The combination of channel attention and spatial attention (CBAM) [20] inherited the merits of spatial attention and channel attention in image classification.

3.1 Datasets and experimental settings

VOC dataset [21] includes 9,963 natural images and is composed of 20 categories. Specifically, training dataset includes 5,011 images. Test dataset contains 4,952 images.

WIDER [22] attribute dataset includes 50,574 wider natural images. And the training dataset contains 5,509 natural images, validation dataset makes up of 1,362 natural images, test dataset includes 6,918 images.

Collected waste bottle dataset by Guangdong Databeyond Technology Co., Ltd is composed of three categories. The mentioned three categories include 8,500 images with sizes of 640×480. Specifically, 7,000 images are used as training images and others are employed as test images. Also, some visual waste bottle images are shown in Figure 1.

Fig. 1 Nine waste bottle images

To test the robustness of the mentioned CNNs with attentions in image classification under certain scene. We use residual networks with attentions on collected waste bottle dataset to test classification results. Also, the experimental settings are shown as follows.

Firstly, we crop given images to size of 224×224 via randomly horizontally and scaled operations to accelerate the training speed of classifier. Then, probabilities of ColorJitter, Gaussian blurs, and grayscale are set to 0.8, 0.5 and 0.2, which can enhance the training dataset. Finally, we normalize training images via channel mean and standard deviations to unify distributions of training samples. Besides, all the networks are optimized by Adam [23]. Also, initial learning rate parameter is 0.05, which it is varied by a cosine annealing strategy [24] with a minimum learning rate limit of 1e-4. The batch size is 128 and the number of epoch is 200.

3.2 Experimental Results

We discuss the effects of CBAM and CA on ResNet18 [7] at different locations. As shown Table 1, we can see that the ResNet18+CBAM and ResNet18+CA outperform ResNet18 on collected waste bottle image dataset, which shows the effectiveness of CBAM and CA on CNNs for image classification. Besides, ResNet18+CBAM(L=in), ResNet18+CBAM(L=pre) and ResNet18+CBAM(L=post) obtain different accuracy in waste image classification, which illustrates importance of different locations of CBAM is different. Besides, to make a tradeoff between performance and complexity, we reduce the four blocks in ResNet18 as well as Net to discuss the effects of CBAM and CA at different locations on waste bottle image classification. In Table 1, we can see that Net+CA obtains the best performance than that of other methods in terms of accuracy, flops and parameters. Specifically, ResNet18+CBAM

(L=in) denotes the CBAM is fused into each block of ResNet18. Also, ResNet18+CBAM (L=pre), and ResNet18+CABRM (L=post) denote CBAM are set to head and tail of the ResNet18, respectively. ResNet18+CA represents the CA is fused into each block in the ResNet18. Besides, name manners of Net+CABM (L=in), Net+CBAM (L=pre), Net+CABM (L=post) and Net+CA are the same as the mentioned illustrations.

Tab. 1 Accuracy and complexity of different methods on collected waste image dataset for image classification. "L" denotes the location of the attention module embedded into the network

Methods	Accuracy (%)	Flops	Params
ResNet18	94.47	1.82 G	11.18M
ResNet18+CBAM (L=in)	94.91	1.82 G	11.27M
ResNet18+CBAM (L=pre)	95.52	1.82 G	11.18M
ResNet18+CBAM (L=post)	94.87	1.82 G	11.21M
ResNet18+CA	95.33	1.82 G	11.27M
Net	95.07	1.36 G	6.38M
Net+CBAM (L=in)	94.33	1.36 G	6.47M
Net+CBAM (L=pre)	95.13	1.36 G	6.38M
Net+CBAM (L=post)	95.33	1.36 G	6.42M
Net+CA	95.47	1.36 G	6.44M

3.3 Potential research points and challenges

Potential research points: 1) Fusing the Transformer into CNN addresses image classification. 2) How to use the combination of CNN and attention mechanism to deal image classification under complex scenes. 3) How to use the CNN and attention mechanism to address image classification with insufficient samples. Challenges: 1) How to address the unstable training of the combination of CNN and attention mechanism. 2) How to reduce the high complexity and huge computational cost of Transformer.

4 Conclusion

In this paper, we summary CNNs with attention mechanisms for image classification. The survey introduces development of CNNs, basis of CNNs and attention mechanism in image classification. Subsequently, we give the main architecture of CNNs with attentions, public and our collected datasets, experimental results in image classification. Finally, we point out potential research points, challenges attention-base for image classification and summary the whole paper.

Acknowledgments

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant D5000210966, in part by the Basic Research Programs of Taicang.

Reference

[1] Krizhevsky A, Sutskever I and Hinton G E 2012 Imagenet classification with deep convolutional neural networks *Advances in Neural Inform*ation *Processing Systems* **25** pp 1097-1105

[2] Wang J, Yang Y, Mao J, Huang Z, Huang C and Xu W 2016 Cnn-rnn: A unified framework for multi-label image classification *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition* pp 2285-2294

[3] Wang W, Yang Y, Xin W, Wang W and Li J 2019 Development of convolutional neural network and its application in image classification: a survey *Optical Engineering* vol 58

[4] Simonyan K and Zisserman, A 2014 Very deep convolutional networks for large-scale image recognition *Preprint* 1409.1556

[5] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A 2015 Going deeper with convolutions. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition* pp 1-9
[6] Joshi S, Verma D K, Saxena G and Parwye A 2019 Issues in training a convolutional neural network model for image classification *Int. Conference on Advances in Computing and Data Sciences* pp 282-293
[7] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition* pp 770-778
[8] Chen B, Huang Y, Xia Q and Zhang Q 2020 Nonlocal spatial attention module for image classification *Int. Journal of Advanced Robotic Systems* vol 17 p 5
[9] Mohammed A I, Tahir A A K 2020 A new optimizer for image classification using wide resnet *Academic Journal of Nawroz University* **9** pp 1-13
[10] Hitawala S 2018 Evaluating resnext model architecture for image classification *Preprint* 1805.08700
[11] Mahmood A, Bennamoun M, An S and Sohel F 2017 Resfeats: residual network based features for image classification *IEEE International Conference on Image Processing* pp 1597-1601
[12] Li G, Li L, Zhu H, Liu X and Jiao L 2019 Adaptive multiscale deep fusion residual network for remote sensing image classification *IEEE Transactions on Geoscience and Remote Sensing* **57** pp 8506-8521
[13] Zhang K, Guo Y, Wang X, Yuan J and Ding Q 2019 Multiple feature reweight DenseNet for image classification *IEEE Access* **7** pp 9872-9880
[14] Zhu L, Chen Y, Ghamisi P and Benediktsson J A 2018 Generative adversarial networks for hyperspectral image classification *IEEE Transactions on Geoscience and Remote Sensing* **56** pp 5046-5063
[15] Chen Z M, Wei X S, Wang P and Guo Y 2019 Multi-label image recognition with graph convolutional networks *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp 5177-5186
[16] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I 2017 Attention is all you need *Advances in Neural Information Processing Systems* pp 5998-6008
[17] Hu J, Shen L and Sun G Squeeze-and-Excitation Networks 2018 *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition* pp 7132-7141
[18] Tian C, Xu Y, Li Z, Zuo W, Fei L and Liu H 2020 Attention-guided CNN for image denoising *Neural Networks* **124** pp 117-129
[19] Wang M 2015 Multi-path convolutional neural networks for complex image classification *Preprint* 1506.04701
[20] Woo S, Park J, Lee J Y and Kweon I S 2018 Cbam: convolutional block attention module *Proc. of the European Conference on Computer Vision* pp 3-19
[21] Everingham M, Gool L V, Williams C K I, Winn J and Zisserman A 2010 The pascal visual object classes (voc) challenge *Int. Journal of Computer Vision* vol 88 pp 303–338
[22] Li Y, Chen H, Loy C C and Tang X 2016 Human attribute recognition by deep hierarchical contexts 2016 *Proc. of European Conference on Computer Vision* vol 9910 pp 684–700
[23] Kingma D P and Ba J 2014 Adam: A method for stochastic optimization *Preprint* 1412.6980
[24] Loshchilov I and Hutter F 2016 Sgdr: Stochastic gradient descent with warm restarts *Preprint* 1608.03983