An Improved PeleeNet Algorithm with Feature Pyramid Networks for Image Detection

Bai Yangfan*, Inwhee Joe*
*Dept. of Computer Software, Hanyang University
e-mail : baiyangfan08@hanyang.ac.kr

Abstract

Faced with the increasing demand for image recognition on mobile devices, how to run convolutional neural network (CNN) models on mobile devices with limited computing power and limited storage resources encourages people to study efficient model design. In recent years, many effective architectures have been proposed, such as mobilenet_v1, mobilenet_v2 and PeleeNet. However, in the process of feature selection, all these models neglect some information of shallow features, which reduces the capture of shallow feature location and semantics. In this study, we propose an effective framework based on Feature Pyramid Networks to improve the recognition accuracy of deep and shallow images while guaranteeing the recognition speed of PeleeNet structured images. Compared with PeleeNet, the accuracy of structure recognition on CIFA-10 data set increased by 4.0%.

1. Introduction

In this paper, We redesigned the original architecture of Feature Pyramid Networks (FPN). The method of feature fusion is simplified, and the feature is fused by up-sampling for prediction. Under the condition of guaranteeing the speed of PeleeNet structure image recognition, the deep and shallow image features are fused together with the redesigned FPN structure for image recognition. Compared with PeleeNet, the accuracy of structure recognition on CIFA-10 data set increased by 4.0%.

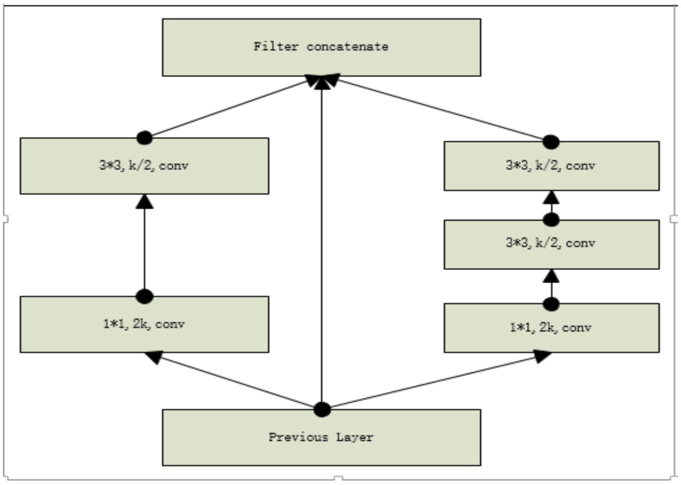
2. PeleeNet architecture

The architecture of the PeleeNet is shown as follows in Table 1. The entire network consists of a stem block and four stages of feature extractor. Except the last stage, the last layer in each stage is average pooling layer with stride 2. A four-stage structure is a commonly used structure in the large model design. ShuffleNet uses a three stages structure and shrinks the feature map size at the beginning of each stage. Although this can effectively reduce computational cost, it is very important for early stage features especially vision tasks, and that premature reducing the feature map size can impair representational abilities. Therefore, PeleeNet still maintain a four-stage structure. The number of layers in the first two stages are specifically controlled to an acceptable range.

<Table 1> PeleeNet architecture

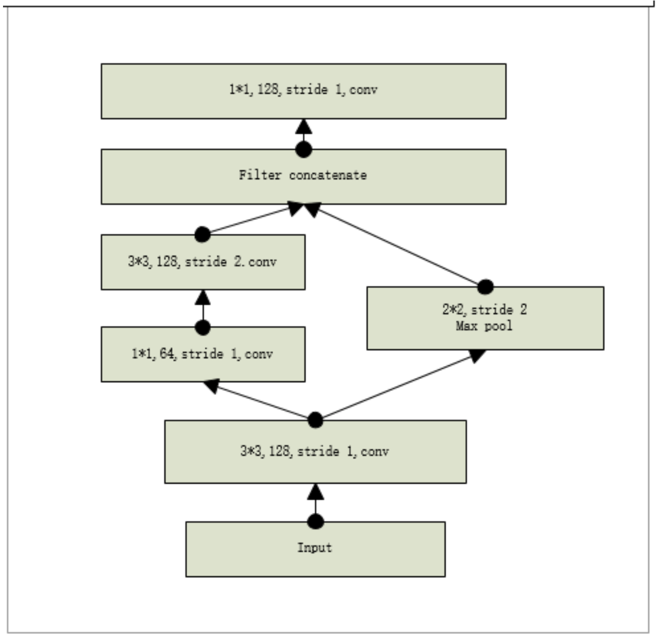
Stage		Layer	Output Shape
	Input		32*32*3
Stage0	Steam Block		16*16*128
Stage1	Dense Block	Dense Layer*4	
	Transition Layer	1*1 conv,stride1	8*8*256
		2*2average pool ,stride2	
Stage2	Dense Block	Dense Layer*8	
	Transition Layer	1*1 conv,stride1	4*4*512
		2*2average pool ,stride2	
Stage3	Dense Block	Dense Layer*6	
	Transition Layer	1*1 conv,stride1	4*4*704
		2*2average pool ,stride2	
Classification Layer		7*7 global average pool	1*1*704
		1000D fully-connect, softmax	

PeleeNet uses a 2-way dense layer to get different scales of receptive fields. The layer uses two stacked 3x3 convolution to learn visual patterns for large objects. The structure is shown on Fig.1.



(Fig.1) 2-way dense layer

PeleeNet designs a cost efficient stem block before the first dense layer. The structure of stem block is shown on Fig. 2. This stem block can effectively improve the feature expression ability without adding computational cost too much - better than other more expensive methods, increasing channels of the first convolution layer or increasing growth rate.



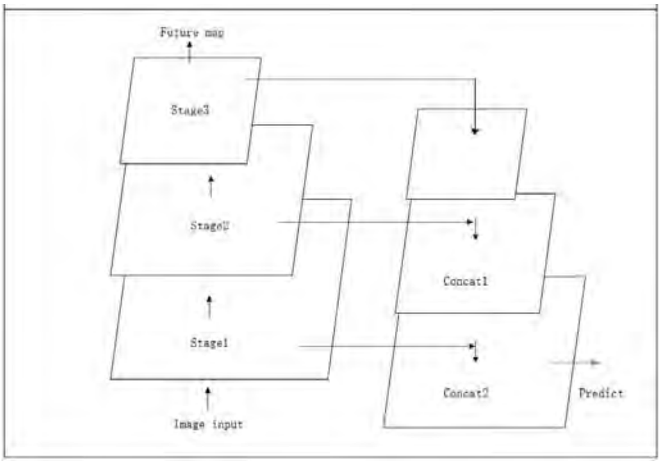
(Fig.2) Stem Block

3. Feature pyramid network architecture

Recognizing objects at vastly different scales is a fundamental challenge in computer vision. Feature pyramids built upon image pyramids (for short we call these featured image pyramids) form the basis of a standard solution. These pyramids are scale-invariant in the sense that an object' s scale change is offset by shifting its level in the pyramid. Intuitively, this property enables a model to detect objects across a large range of scales by scanning the model over both positions and pyramid levels.

Most object detection algorithms only use top-level features to predict, but we know that the low-level feature semantic information is relatively small, but the target location is accurate. The high-level feature semantic information is rich, but the target location is rough. In this paper, multi-scale feature fusion method is used for feature prediction. With shallow location features and deep semantic information preserved, the calculation parameters are reduced to ensure the real-time operation of the mobile terminal.

The goal of this paper is to make use of the pyramidal shape of a FPN' s feature hierarchy while creating a feature pyramid that has strong semantics at all scales. To achieve this goal, we rely on an architecture that combines low-resolution, semantically strong features with high-resolution, semantically weak features via a top-down pathway. The result is a feature pyramid that has rich semantics at all levels and is built quickly from a single input image scale.Fig.3 below is the FPN architecture designed in this paper. Table 2 is the architecture design of specific FPN.



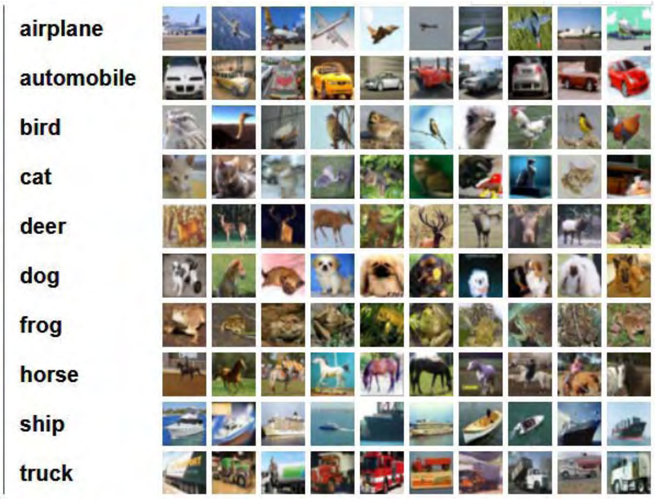
(Fig.3) Improved FPN

<Table 2> Improved FPN architecture

Concatenate	Layer	Output shape
	Stage3_input	4*4*704
	Conv 1*1*512	4*4*512
Concat1		4*4*1024
	Conv 1*1*256	4*4*256
	Conv 3*3*512	4*4*512
	Conv 1*1*256	4*4*256
	resize	8*8*256
Concat2		8*8*512
	Conv 1*1*128	8*8*128
	Conv 3*3*256	8*8*256
	Conv 1*1*128	8*8*128
Adjust size	Conv 1*1*256 Stride 2*2	4*4*256
	Conv 1*1*512	4*4*512
	Conv 1*1*704	4*4*704
Classification layer	7*7 global average pool	1*1*704
	1000D FC, softmax	

4. Experiments on Object Detection

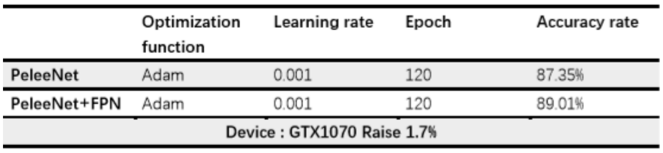
The CIFA-10 data set is used in this test. There are 60,000 color images in this data set. These images are 32*32 and are divided into 10 categories, with 6,000 images in each category. There are 50,000 pictures for training, which constitute five batches of 10,000 pictures for each batch. Another 10,000 are used for testing, forming a single batch. The data of the test batches were taken from each of the 10 categories, and 1000 pieces were randomly selected for each category. The rest were randomly arranged into training batches. As shown in Fig.4.



(Fig.4) CIFA-10 Data Set

Firstly, We trained the original PeleeNet model and the improved PeleeNet+FPN model 120 times under the conditions of 0.001 learning rate and Adam optimization function respectively. After CIFA-10 data set test, the recognition rate of the observed pictures is 87.35% and 89.01% respectively. It can be seen that the recognition rate of the improved PeleeNet+FPN model is 1.9% higher than that of the previous PeleeNet model. The results are shown in Table 3 below.

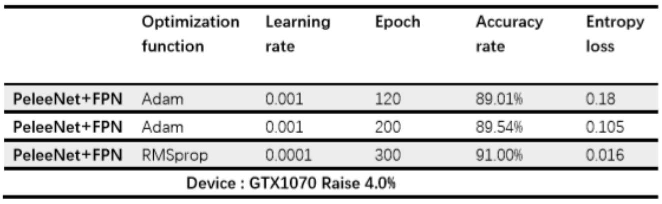
<Table 3> Test Result 1



	Optimization function	Learning rate	Epoch	Accuracy rate
PeleeNet	Adam	0.001	120	87.35%
PeleeNet+FPN	Adam	0.001	120	89.01%
Device : GTX1070 Raise 1.7%				

Then I fine-tuned the PeleeNet + FPN model twice. Entropy loss of PeleeNet+FPN model decreased from 0.18 to 0.016, a decrease of 91%. The results are shown in Table 4 below.

<Table 4> Test Result 2



	Optimization function	Learning rate	Epoch	Accuracy rate	Entropy loss
PeleeNet+FPN	Adam	0.001	120	89.01%	0.18
PeleeNet+FPN	Adam	0.001	200	89.54%	0.105
PeleeNet+FPN	RMSprop	0.0001	300	91.00%	0.016
Device : GTX1070 Raise 4.0%					

5. Conclusion

The PeleeNet+FPN model proposed in this paper has a significant improvement in the performance of target recognition compared with the original PeleeNet model. PeleeNet + FPN model can effectively combine feature maps of different scales, and use the high-resolution information of low-level features and high-level features to achieve the prediction effect by fusing these features of different layers. Good recognition results are obtained on CIFA-10 data set.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, 2009.

[2] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe. Convolutional architecture for fast feature embedding. In ACM, 2014.

[3] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In European conference on computer vision, pp. 21–37. Springer, 2016.

[4] Geoff Pleiss, Danlu Chen, Gao Huang, Tongcheng Li, Laurens van der Maaten, and Kilian Q Weinberger. Memory-efficient implementation of densenets. arXiv preprint arXiv:1707.06990, 2017.

[5] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In CVPR, 2015.

[6] Robert J.Wang, Xiang Li and Charles X.Ling. Pelee: A Real-Time Object Detection System on Mobile Devices. In NIPS, 2018.

[7] Tsung-Yi Lin, Piotr Doll´ ar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection. In CVPR, 2017.

[8] S. Ren, K. He, R. Girshick, X. Zhang, and J. Sun. Object detection networks on convolutional feature maps. PAMI, 2016.

[9] K.He, X.Zhang, S.Ren and J.Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In ECCV. 2014.

[10] R. Girshick. Fast R-CNN. In ICCV, 2015.