Research and Development of Lightweight Neural Network Mobilenet

YiJie Zhang

Abstract— **The development of deep learning has made huge breakthroughs in the fields of target detection and classification, and achieved accuracy that cannot be achieved by the original traditional methods. Existing neural networks are all developing in a deeper and wider direction, which directly leads to an increasing amount of neural network models and calculations, leading to many obstacles to deployment on mobile devices. For this reason, mobilenet is proposed to allow the model to greatly reduce the amount of calculation and parameters without losing progress**

Index Terms— **Mobilenet, Resnet, Depthwise Separable Convolution, AutoAI**

I. INTRODUCTION

With the development of alexnet[1] and the improvement of Resnet[2], the development of neural networks began to become deeper and deeper, reaching a depth of nearly two hundred layers. And Iceptionnet[3] proposed that a wider neural network can also have good results. Since then, a deeper and wider neural network has become the primary direction of network design. However, this has led to higher and higher requirements for hardware in neural networks, and more and more difficult deployment of mobile terminals. Ordinary mobile terminal equipment cannot withstand the high amount of computing and huge model parameters. And mobilenetv1[4] made the deployment of neural networks on the mobile terminal easier. The depth separable convolution proposed by Mobilenetv1 enables the network to greatly reduce its own parameters and calculations while ensuring accuracy. Mobilenetv2[5] optimizes the fusion of the deep separable convolution structure and the residual structure of Resnet, so that the network can learn better features and obtain higher accuracy while being lighter. Mobilenetv3[6] automatically finds the best neural network architecture through AutoML[7], introduces the SE[8] structure and optimizes the activation function to achieve a faster and lighter lightweight network.

II. DEPTH SEPARABLE CONVOLUTION

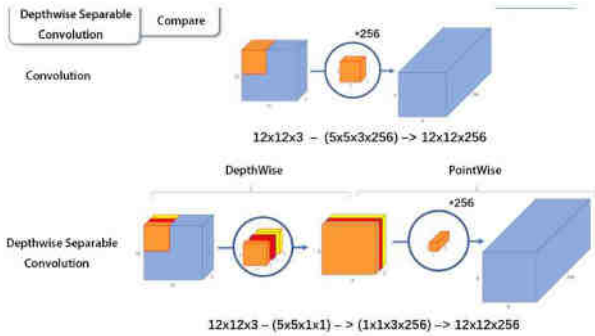

Figure 1

Manuscript received September 22, 2020

A. Process

Ordinary convolution considers channel information and spatial information together, that is, after convolving the input feature map through multiple channels in a convolution kernel, the feature information of the feature map and the channel information will be fused together. Mobilenet proposes that channel information and feature information can be learned separately, first learn the feature information of each channel and then perform fusion learning between channels, as shown in Figure 1. Deep separable convolution first performs DW convolution to learn the feature information in each channel, and then performs PW() convolution to learn the fusion information of all channels.

The DW convolution is shown in Figure 1. When the input feature map is a 12×12×3 3-channel picture, DW provides 3 single-channel convolution checks to convolve each channel of the input picture, and then get a 3 A three-channel feature map composed of convolved single-channel feature maps. Normal convolution at this time is to convolve the input image through one or more 3-channel convolution kernels, and then obtain feature maps of multiple channels. A PW immediately follows the DW, which provides 256 3-channel 1×1 convolution kernels to perform PW convolution on the feature map obtained by the DW. Finally, the 256-channel feature map output result is obtained. Because the convolution method of each convolution kernel in DW is single-channel convolution, the weight is updated in a single channel when it is updated, and the gradient of each channel is different. This can better learn the unique features of each channel, and then use PW to perform 1×1 convolution on each channel feature learned by DW above, which can provide nonlinearity for the network while increasing the dimensionality , And can obtain fusion features through joint learning of the features of each channel through 3 channels of 1×1 convolution. The final feature map output by the depth separable convolution composed of DW and PW convolution has greatly reduced the amount of calculation and parameter, and the computational time and space costs have been further reduced.

B. Calculation

The specific calculation formula of the depth separable convolution is just like the one written in Figure 1. In the DW process, because only three single-channel convolution kernel convolutions have been learned, the uniqueness of each channel of the input feature map has been learned. There are features, so the 1×1 convolution kernel is used when jointly learning the features between channels, so even if the PW convolution process contains 256 convolution kernels, it still greatly reduces the amount of calculation and storage Overhead. Ordinary convolution is to fuse and calculate the channel characteristics and the connection between the channels. As shown in Figure 1, there are 256 3×3

convolution operations, so the calculation amount and parameter amount of normal convolution are almost depth separable convolutions 9 times the product. This improvement has allowed the model to be deployed on mobile devices. And it can provide real-time backbone for target detection, target classification, target segmentation and other networks. However, because the deep separable network disguised one layer of convolution into two layers, the acceleration of parallel devices will not appear to be very fast, because the number of layers becomes more and the single layer utilization becomes lower, which is very suitable for edge devices. CPU operation, but for Nvidia graphics cards, fpga, etc., the acceleration is not as obvious as the CPU, but it can still provide speeds that exceed ordinary convolution.

III. RESIDUAL STRUCTURE

In mobilenetv2, the idea of resnet's residual network is absorbed, that is, after a VGG-like convolution block, the output and the input of the convolution block are added. This can effectively prevent the model from gradually disappearing or exploding during the transmission process, making the network deeper. And in the subsequent series of resnet, each convolution must first reduce the dimensionality of the channel through a 1×1 convolution, and then convolve the 3×3 convolution kernel and then pass the 1×1 convolution kernel to upgrade the dimension back. This way can significantly reduce the amount of calculation.

In mobilnetv2, not only the idea of residual error has been absorbed, but also new considerations have been made to the maintenance of resnet. Because the core of the deep separable convolution is to learn the features of each channel separately from the DW, and this part of the calculation is very small, in order to better learn the individual features of each channel, mobilenetv2 adds 1 before the input is sent to the DW The ×1 convolutional layer performs an upscaling operation on the input, adding more channels to the features of the input DW. This method can better learn the distribution of features. At the same time, in order to reduce the amount of calculation, the subsequent PW will dimensional The number of channels before the upgrade is restored. The multiplier of the ascending dimension is 6. In the paper, it is mentioned that the ascending and descending dimensionality multiplier can have a better feature protection effect after 15, but considering the amount of parameters and the amount of calculation, the paper has made a compromise and selected 6 as the descending and descending dimension multiple. And made these structures into a block. At the beginning of the block is a PW convolutional layer for dimension enhancement, followed by a DW feature learning layer, and then a PW convolutional layer to reduce dimensionality, and the final output of the block and the input are added as the input of the next block. This structure makes the data in the model very narrow, but it is very wide in each block, which is also very helpful for reducing the amount of model parameters and calculations. Compared with v1, v2 version of the model is better at the same time, the calculation amount is reduced by two-fifths. It was done quickly and well.

IV. AUTO SEARCH

Neural network automatic search is used in mobilenetv3. The automatic search network can preset a series of different models, and then obtain a roughly accurate model through training, and finally select the model structure with the best accuracy. The automatic search network replaces manual selection, which has a better effect. In addition, in order to further reduce the amount of parameters, v3 also changed the first layer of the model's convolution channel to 16, which increased the accuracy by 3ms while ensuring the accuracy. On this basis, the v3 version adds the SE structure, which is equivalent to the channel attention mechanism, and performs weighting calculations for each channel. However, because the SE structure increases the computational overhead, the dimension is reduced to a quarter of the original when the SE structure is introduced, which improves the accuracy without increasing the time consumption. In addition, the tail structure of V2 is modified, and the average pooling of the tail is mentioned before the 1×1 convolution, which greatly reduces the amount of tail calculation. At the same time, the 3×3 and 1×1 convolutional layers of the tail block are deleted. Through these operations to reduce the amount of calculation, 15 ms is improved while ensuring accuracy. The final h-mish operation is also maintained to bring better convenience for quantification.

CONCLUSION

The Mobilenetv1 to v3 series have advanced the structure of the deep separable convolution to the improvement of each version, so that the amount of calculation has been greatly reduced while the accuracy has been continuously increased, and the final v3 version has optimized the quantization design. Makes the mobilenet series can be better deployed on the mobile terminal, achieving faster and more accurate. However, due to the inherent limitations of the mobilenet structure and the excessive number of layers, the computing chip unit cannot be fully utilized when facing parallel computing devices, resulting in the inability to achieve theoretical acceleration. But on mobile devices with cpu, it can be accelerated close to theory.

REFERENCES

[1] Krizhevsky A , Sutskever I , Hinton G . ImageNet Classification with Deep Convolutional Neural Networks[J]. Advances in neural information processing systems, 2012, 25(2).

[2] He K , Zhang X , Ren S , et al. Deep Residual Learning for Image Recognition[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society, 2016.

[3] Szegedy C , Liu W , Jia Y , et al. Going Deeper with Convolutions[J]. 2014.

[4] Howard A G , Zhu M , Chen B , et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications[J]. 2017.

[5] Rosenfeld A , Tsotsos J K . Incremental Learning Through Deep Adaptation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017.

[6] Howard A , Sandler M , Chu G , et al. Searching for MobileNetV3[J]. 2019.

[7] Guyon I , Bennett K , Cawley G , et al. Design of the 2015 ChaLearn AutoML challenge[C]// International Joint Conference on Neural Networks. IEEE, 2015.

[8] Hu J , Shen L , Albanie S , et al. Squeeze-and-Excitation Networks[C]// IEEE, 2017.