Image Classification Architecture

Gupta, Rajan Das (id: 18-36304-1)[a], Jui Saha Pritha 2 (id: 17-35507-3)[a], Paul Bishal (id: 17-35836-3)[a], Zubaer Hossain Asif 4 (id: 18-36248-1)[a]

[a] *Department of Computer Sciences, American International University-Bangladesh*

Abstract

Image classification is the process of categorizing and labeling groups of pixels or vectors within an image based on specific rules.Convolutional networks (ConvNets) have recently enjoyed a great success in large-scale image and video recognition. In this report, we propose an efficient and noncomplex image classification architecture using deep learning based on the most popular algorithms: Convolutional Neural Network (CNN)) for Feature extraction, Stacked Auto Encoder (SAE) for reducing the dimensionality, and Recurrent Neural Network (RNN) for increasing the accuracy. Convolutional networks (ConvNets) have recently enjoyed a great success in large-scale image and video recognition. To improve their performance, we can collect larger datasets, learn more powerful models, and use better techniques for preventing overfitting.

Keywords: Deep learning, Computer Vision, Object detection, NN, CNN

1. Introduction

Computer image classification is to analyze and classify images into certain categories to replace human visual interpretation. It is one of the hotspots in the field of computer vision. Because the features are very important to classification, most of the researches on image classification focus on image feature extraction and classification algorithms.

Convolutional neural networks have the ability of self-learning, self-adapting, and self-organizing; so, it can automatically extract features by using the prior knowledge of the known categories, and avoid the complicated process of feature extraction in traditional image classification methods. At the same time, the extracted features are highly expressive and efficient. Deep convolutional neural network (CNN) has achieved significant success in the field of computer vision, such as image classification, target tracking, target detection, and semantic image segmentation . For example, in the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012), Krizhevsky et al. won the championship with an AlexNet model of about 60 million parameters and eight layers. In addition, VGG with 16-layer, GoogleNet with Inception as the basic structure, and ResNet with residual blocks that can alleviate the problem of gradient disappearance have also achieved great success. However, the deep convolutional neural network itself is a dense computational model. The huge number of parameters, heavy computing load, and large number of memory access lead to huge power consumption, which makes it difficult to apply the model to portable mobile devices with limited hardware resources.

Compared with VGG-16 network, MobileNet is a lightweight network, which uses depthwise separable convolution to deepen the network, and reduce parameters and computation. At the same time, the classification accuracy of MobileNet on ImageNet data set only reduces by 1 Percent. However, in order to be better applied to mobile devices with limited memory, the parameters and computational complexity of the MobileNet model need to be further reduced. Therefore, we use dense blocks as the basic unit in the network layer of MobileNet. By setting a small growth rate, the model has fewer parameters and lower computational cost. The new models, namely Dense-MobileNets, can also achieve high classification accuracy.

1. ***Capture reader's interest*** - The memory intensive and highly computational intensive features of in deep learning restrict its application in portable devices. Compression and acceleration of network models will reduce the classification accuracy.
2. ***General aims*** – Image classification is the primary domain, in which deep neural networks play the most important role of medical image analysis. The image classification accepts the given input images and produces output classification for identifying whether the disease is present or not.
3. ***Specific objectives*** – 1.Input: Input is a collection of N images; each image label is one of the K classification tags. This set is called the training set.
   2.Learning: The task of this step is to use the training set to learn exactly what each class looks like. This step is generally called a training classifier or learning a model.

3.Evaluation: The classifier is used to predict the classification labels of images it has not seen and to evaluate the quality of the classifiers. We compare the labels predicted by the classifier with the real labels of the image. There is no doubt that the classification labels predicted by the classifier are consistent with the true classification labels of the image, which is a good thing, and the more such cases, the better.

4. ***List your research questions*** -

   - Q1: What is image classification technology and how does it work?
   - Q2: What are the research approaches followed in this study?
   - Q3: Does image classification make sure safety of a transaction?
   - Q4: Where are the technology is being used?

5. ***Provide an overview of the forthcoming chapters*** - Image Image Classification, we now have more user and trend data than ever. Varying in form, data could be text, image, speech, or a mix of these. Images now constitute a part of user data more prominently than ever.However, the image data that we have, is unstructured and requires advanced methods like deep learning models to analyze it. Arguably the most crucial part of digital image analysis, image classification today, uses AI systems based on deep learning models to achieve better and more accurate resultsWhile a person can naturally classify images, one might wonder how a computer learns to do that. The answer is, using Convolutional Neural Networks (CNN). A CNN is a framework built using concepts of machine learning.

2. Literature Review

In this particular section, we will discuss 5 different model or architectures 44 45 of convolutional neural networks(CNN).Here we will know about the model 45 46 design about architecture.

2.1 Alexnet AlexNet Krizhevsky et al. [2012] is one of a convolutional neural network (CNN) architecture.The network achieved a top-5 error of 15.3 parcent, more than 10.8 percentage points lower than that of he runner up. The original paper's primary result was that the depth of the model was essential for its high performance, which was computationally expensive, but made feasible due to the utilization of graphics processing 54 units (GPUs) during training Krizhevsky [2017] the architecture is described in figure 1 in

AlexnNet architecture. In the picture, we can see the net contain eight-layer with weight the first five are convolutional and the rest are fully connected are fed to a 1000- way softmax.in this architecture, it says The input to the network is an image of dimen-sions (227, 227, 3). it has 5 convolution layers in the first convolution layers it takes 227*227 RGB images.

The input to the network is an image with dimensions (227, 227, 3). the first

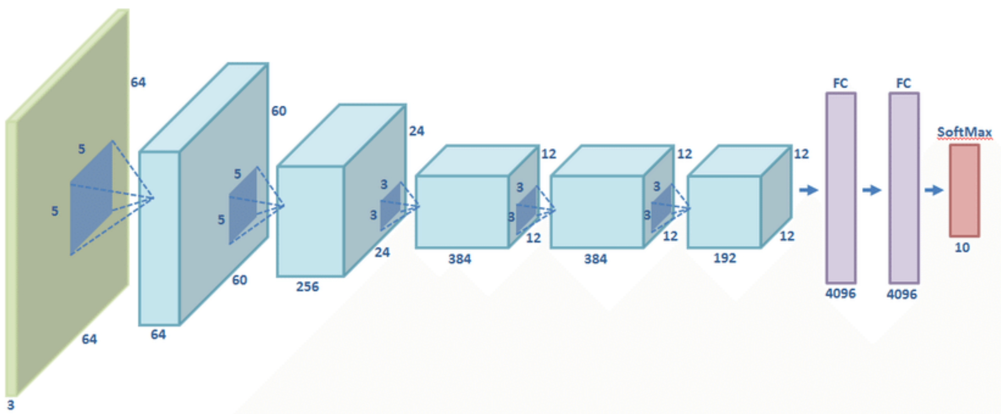

Figure 1: ALexNet Model

layer has 96 Channels of 11*11 filter size with stride 4*4 and the activation layer will be relu followed by a max-pooling pool size 2*2 with stride 2*2 In the next convolution layer, it takes 256 filters with the size of 11*11 and this convolution stride will be 1*1 And same as before the activation layer will be relu then the max pooling with pool size 2*2 and stride 2*2 The next 2 convolution layers have the same kernel size 3*3 with filter 384 and also the same stride 1*1 Followed by same activation function relu and in this layers don't have any max pool . For normalization or preprocessing they firstly take the mean from all the images and then subtract from the original image then fit the image set in the model.

2.2 VGG-Net VGGSimonyan and Zisserman [2014] net is a another model or architecture of convolutional neural net- work (CNN) . the speciality of this networks is increasing depth using an architecture with very small (3×3) convolution filters, which shows that asignificant improvement on the prior-art configurations can be achieved by pushing the depth to 16–19 weight layers. During training, the input to this model is a fixed-size 224×224 RGB image. They do only one preprocessing here . in this paper they subtracting the mean RGB value, computed on the training set, from each pixel. The image is passed through a stack of convolutional (conv.) layers, where we use filters with a very small receptive field: 3×3. In one of the con- figurations we also utilise 1×1 convolution filters, which can be seen as a linear transformation of the input channels .The convolution stride is fixed o 1 pixel; the spatial padding of conv. layer input is such that the spatial resolution is preserved after convolution, i.e. the padding is 1 pixel for 3×3 conv. layers. Spatial pooling is carried out by five max-pooling layers, which follow some of the conv. layers (not all the conv. layers are followed by max-pooling).

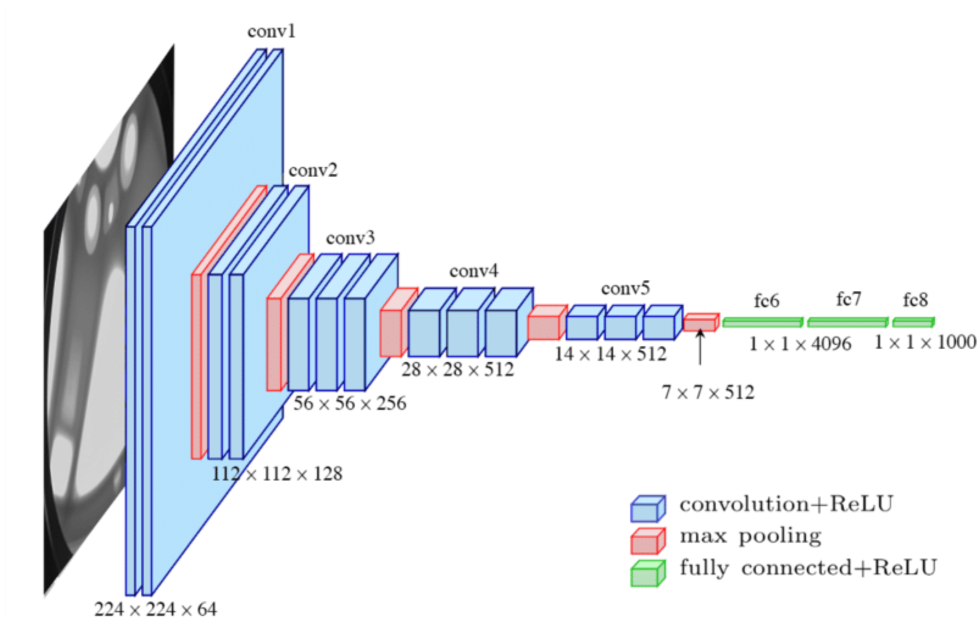

Figure 2: VGG Model

Max-pooling is performed over a 2×2 pixel window, with stride 2. A stack of convolutional layers (which has a different depth in different archi- tectures) is followed by three Fully-Connected (FC) layers: the first two have 4096 channels each, the third performs 1000- way ILSVRC classification and thus contains 1000 channels (one for each class).

2.3 GoogleNet This Szegedy et al. [2015] is another reworded architecture this architecture is a bit different from AlexNet , VGG-Net. In their paper, they said there is a fixed convo- lution size for each layer. In the inception module, 1*1,3*3,5*5 convolution and 3*3 max-pooling perform parallel way of the input and out of these are stacked together to generated final output.

- **Auxiliary classifier :** Inception

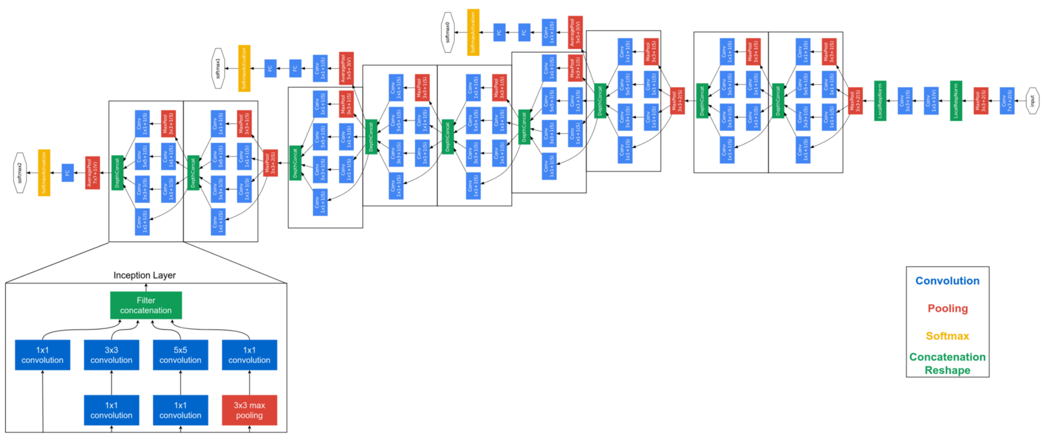

Figure 3: GoogleNet Model

The overall architecture is 22 layers deep architecture in this paper they claimed this architecture was designed to keep computational efficiency in mind .it can ruin individual devices with low computational resources. The architecture also contains two auxiliary classifier layers connected to the output of inception (4a)and inception (4d)layers.

This architecture takes image of size 224 x 224 with RGB color channels. All the convolutions inside this architecture uses Rectified Linear Units (ReLU) as their activation functions.

- ***2.4 ResNet*** TheHe et al. [2016] constitutional layers have 3×3 filters and follow two rules. 1.the layers have the same number of filters for the same number of output map sizes. and (ii) As the time complexity preserve layer halved, the feature map size in filters is doubled.34 layer baseline has 3.6 billion FLOPs that only 18 percent of VGG-19.

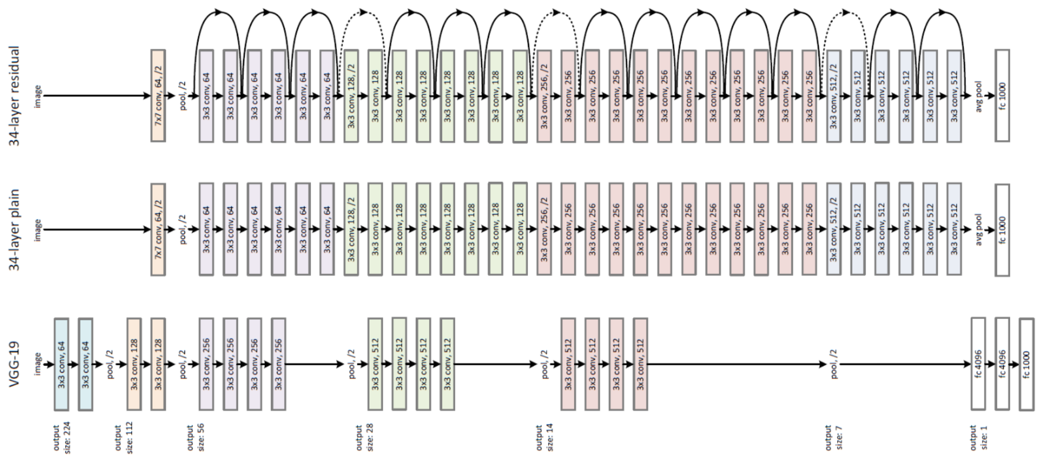

Figure 4: RestNet Model

For VGG-19 the image size 3x3 convo,64 the output size will . Then goes pool, /2, The image has 3x3 Conv, 128 3x3 Conv, the output size 112 for both residual layer and 34-layer plain. Next, go to image process pool, /2,3x3 Conv, 256, The 34-layer plain,34-layer residual 3x3 Conv, 64, the output of size is 56. It gradually decreases. Than go to the next image line pool, /2,3x3 Conv, 512 ,at the same time 34-layer plain and 34-layer residual 3x3 Conv, 256, /2,output size is 14.The VGG-19 image has/pool, the 34-layer plain and 34-layer residual have 3x3 Conv,, the output is 7. The last step pool/2, the last VGG-19 model has FC 4096, FC 4096, FC 1000, the output of the size 1.34-layer plain,34-residual layer goes to avg pool FC 1000 with output size 1.

- ***2.5 MobileNet***

The Mobile Network structureHoward et al. [2017] is built on a separable department-wise first layer which is full convolution. For the classification of the softmax layer, ReLU has non-linearity of the final connected layer which has no non- linearity and feeds into a soft max layer for classification. In this figure, the contrasts layer with involve regular convolutions. ReLU and batch norm lin- earity involves with the factorized layer.Mobile Net has 28 layers, by Counting the point wise and depth wise convoke locutions these are separate layers. It is also important to check its efficiency for implementation. dense matrix is faster than sparse matrix because dense has a very high level of sparsity. Our structured model has all of the computation into dense input size is $112 \times 112 \times 32$. Mobile architecture body type and stride 193 Conv / s1 filter size will be $1 \times 1 \times 64 \times 128$ and input size will be 56 $\times$ 56 194 $\times$ 64.Type/stride Conv DW / s2 filter size $3 \times 3 \times 128$ dw and input size can be $56 \times 56 \times 128$. By analyzing all these data it gradually increases filter size, input size. Type and size of stride Conv / s1 filter shape is $1 \times 1 \times$ and input data is 128×256 $28 \times 28 \times 128$.If the resource layer has fully Connected the multi-add will 0.18 percent and the parameter will 24.33 percent

3. Discussion

Among all the mentioned models the VGG-NET is a keras model with 16 and 19 layer network that has an input size of 224X224. With the achivement of 92.7The preceding models square measure the foremost used and effective means that of classifying pictures dataset for its applications in deep learning. These models not solely facilitate improve the potency and accuracy of our results however conjointly provides us with easier ways that to hold out image classification in our Deep Learning comes.

4. Conclusion

In this paper, we have discussed different image classification for dividing different categories of pictures. The paper also discussed various situations for image classification techniques. Our study also discussed different scenarios for different image classification techniques and the pros and cons of each of them.Therefore, this paper will help us in making the right choice the process of distinguishing between all available strategies.

References

Krizhevsky, A., Sutskever, I., Hinton, G.E.. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 2012;25:1097–1105.

Simonyan, K., Zisserman, A.. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556 2014;.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, p. 1–9.

He, K., Zhang, X., Ren, S., Sun, J.. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, p. 770–778.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:170404861 2017;.