**Image Classification with Convolutional Neural Network (CNN)**

Rakibul Newaz Sourav (18-37316-1), Abdur Rakib Howlader (18-37061-1), Chowdhury, Abdur Rahim ( 18-37629-1), Antu, Asif Ikbal (17-34554-2), Islam, Dewan Saiful(18-37623-1)

*Department of Computer Sciences, American International University-Bangladesh*

# Abstract

This report presents an analysis of Convolutional Neural Networks (CNN) architectures. Deep learning is emerging as a powerful tool and has become a leading machine learning tool in computer vision and image analysis. The history and the working principle of the most popular CNN models like AlexNet, GoogleNet, VGG, RestNet, MobileNet has been described. The advantages and disadvantages are also mentioned in this report.

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

# 

# 1. Introduction

Nowadays internet is filled with anabundance of images and videos, which is encouraging thedevelopment of search

applications and algorithms that can examine the semantic analysis [1] of image and videos for presenting the user with

better sear ch content and their summarization. There have been major br eakthroughs in image labeling, object detection,

scene classification [2] [3], areas reported by different researchers across the world. This leads to making it possible to

formulate approaches concerning object detection and scene classification problems. Since artificial neural networks have

shown a performance breakthrough in the area of object detection and scene classification, specially convolutional neural

networks ( CNN)[4] [5] [6], this work focuses on identifying the best network for this purpose. Feature extraction is a key

step of such algorithms. Feature extraction from images involves extracting a minimal set of features containing ahigh

amount of object or scene information fr om low -level image pixel values, therefore, capturing the difference among the

object categories involved. Some of the traditional feature extraction techniques used on images are Scale-invariant feature

transform (SIFT) [7], histogram of oriented gradients (HOG) [8], Local binary patterns (LBP) [10], Content-Based Image

Retrieval (CBIR) [11], etc. Once features are extracted their classification is done based on objects present in an image. A

few examples of classifiers are Support vector machine (SVM), Logistic Regression, Random Forest, decision trees etc

Nowadays internet is filled with anabundance of images and videos, which is encouraging thedevelopment of search

applications and algorithms that can examine the semantic analysis [1] of image and videos for presenting the user with

better sear ch content and their summarization. There have been major br eakthroughs in image labeling, object detection,

scene classification [2] [3], areas reported by different researchers across the world. This leads to making it possible to

formulate approaches concerning object detection and scene classification problems. Since artificial neural networks have

shown a performance breakthrough in the area of object detection and scene classification, specially convolutional neural

networks ( CNN)[4] [5] [6], this work focuses on identifying the best network for this purpose. Feature extraction is a key

step of such algorithms. Feature extraction from images involves extracting a minimal set of features containing ahigh

amount of object or scene information fr om low -level image pixel values, therefore, capturing the difference among the

object categories involved. Some of the traditional feature extraction techniques used on images are Scale-invariant feature

transform (SIFT) [7], histogram of oriented gradients (HOG) [8], Local binary patterns (LBP) [10], Content-Based Image

Retrieval (CBIR) [11], etc. Once features are extracted their classification is done based on objects present in an image. A

few examples of classifiers are Support vector machine (SVM), Logistic Regression, Random Forest, decision trees etc

Nowadays internet is filled with anabundance of images and videos, which is encouraging thedevelopment of search

applications and algorithms that can examine the semantic analysis [1] of image and videos for presenting the user with

better sear ch content and their summarization. There have been major br eakthroughs in image labeling, object detection,

scene classification [2] [3], areas reported by different researchers across the world. This leads to making it possible to

formulate approaches concerning object detection and scene classification problems. Since artificial neural networks have

shown a performance breakthrough in the area of object detection and scene classification, specially convolutional neural

networks ( CNN)[4] [5] [6], this work focuses on identifying the best network for this purpose. Feature extraction is a key

step of such algorithms. Feature extraction from images involves extracting a minimal set of features containing ahigh

amount of object or scene information fr om low -level image pixel values, therefore, capturing the difference among the

object categories involved. Some of the traditional feature extraction techniques used on images are Scale-invariant feature

transform (SIFT) [7], histogram of oriented gradients (HOG) [8], Local binary patterns (LBP) [10], Content-Based Image

Retrieval (CBIR) [11], etc. Once features are extracted their classification is done based on objects present in an image. A

few examples of classifiers are Support vector machine (SVM), Logistic Regression, Random Forest, decision trees etc

Nowadays internet is filled with anabundance of images and videos, which is encouraging thedevelopment of search

applications and algorithms that can examine the semantic analysis [1] of image and videos for presenting the user with

better sear ch content and their summarization. There have been major br eakthroughs in image labeling, object detection,

scene classification [2] [3], areas reported by different researchers across the world. This leads to making it possible to

formulate approaches concerning object detection and scene classification problems. Since artificial neural networks have

shown a performance breakthrough in the area of object detection and scene classification, specially convolutional neural

networks ( CNN)[4] [5] [6], this work focuses on identifying the best network for this purpose. Feature extraction is a key

step of such algorithms. Feature extraction from images involves extracting a minimal set of features containing ahigh

amount of object or scene information fr om low -level image pixel values, therefore, capturing the difference among the

object categories involved. Some of the traditional feature extraction techniques used on images are Scale-invariant feature

transform (SIFT) [7], histogram of oriented gradients (HOG) [8], Local binary patterns (LBP) [10], Content-Based Image

Retrieval (CBIR) [11], etc. Once features are extracted their classification is done based on objects present in an image. A

few examples of classifiers are Support vector machine (SVM), Logistic Regression, Random Forest, decision trees etc

Nowadays internet is filled with anabundance of images and videos, which is encouraging thedevelopment of search

applications and algorithms that can examine the semantic analysis [1] of image and videos for presenting the user with

better sear ch content and their summarization. There have been major br eakthroughs in image labeling, object detection,

scene classification [2] [3], areas reported by different researchers across the world. This leads to making it possible to

formulate approaches concerning object detection and scene classification problems. Since artificial neural networks have

shown a performance breakthrough in the area of object detection and scene classification, specially convolutional neural

networks ( CNN)[4] [5] [6], this work focuses on identifying the best network for this purpose. Feature extraction is a key

step of such algorithms. Feature extraction from images involves extracting a minimal set of features containing ahigh

amount of object or scene information fr om low -level image pixel values, therefore, capturing the difference among the

object categories involved. Some of the traditional feature extraction techniques used on images are Scale-invariant feature

transform (SIFT) [7], histogram of oriented gradients (HOG) [8], Local binary patterns (LBP) [10], Content-Based Image

Retrieval (CBIR) [11], etc. Once features are extracted their classification is done based on objects present in an image. A

few examples of classifiers are Support vector machine (SVM), Logistic Regression, Random Forest, decision trees etc

Nowadays internet is filled with anabundance of images and videos, which is encouraging thedevelopment of search

applications and algorithms that can examine the semantic analysis [1] of image and videos for presenting the user with

better sear ch content and their summarization. There have been major br eakthroughs in image labeling, object detection,

scene classification [2] [3], areas reported by different researchers across the world. This leads to making it possible to

formulate approaches concerning object detection and scene classification problems. Since artificial neural networks have

shown a performance breakthrough in the area of object detection and scene classification, specially convolutional neural

networks ( CNN)[4] [5] [6], this work focuses on identifying the best network for this purpose. Feature extraction is a key

step of such algorithms. Feature extraction from images involves extracting a minimal set of features containing ahigh

amount of object or scene information fr om low -level image pixel values, therefore, capturing the difference among the

object categories involved. Some of the traditional feature extraction techniques used on images are Scale-invariant feature

transform (SIFT) [7], histogram of oriented gradients (HOG) [8], Local binary patterns (LBP) [10], Content-Based Image

Retrieval (CBIR) [11], etc. Once features are extracted their classification is done based on objects present in an image. A

few examples of classifiers are Support vector machine (SVM), Logistic Regression, Random Forest, decision trees etc

Nowadays internet is filled with anabundance of images and videos, which is encouraging thedevelopment of search

applications and algorithms that can examine the semantic analysis [1] of image and videos for presenting the user with

better sear ch content and their summarization. There have been major br eakthroughs in image labeling, object detection,

scene classification [2] [3], areas reported by different researchers across the world. This leads to making it possible to

formulate approaches concerning object detection and scene classification problems. Since artificial neural networks have

shown a performance breakthrough in the area of object detection and scene classification, specially convolutional neural

networks ( CNN)[4] [5] [6], this work focuses on identifying the best network for this purpose. Feature extraction is a key

step of such algorithms. Feature extraction from images involves extracting a minimal set of features containing ahigh

amount of object or scene information fr om low -level image pixel values, therefore, capturing the difference among the

object categories involved. Some of the traditional feature extraction techniques used on images are Scale-invariant feature

transform (SIFT) [7], histogram of oriented gradients (HOG) [8], Local binary patterns (LBP) [10], Content-Based Image

Retrieval (CBIR) [11], etc. Once features are extracted their classification is done based on objects present in an image. A

few examples of classifiers are Support vector machine (SVM), Logistic Regression, Random Forest, decision trees etc

Computer vision and neural networks are the hot new IT of machine learning techniques. With advances of neural networks and an ability to read images as pixel density numbers, numerous companies are relying on this technique for more data. For example, speed camera uses computer vision to take pictures of license plate of cars who are going above the speeding limit and match the license plate number with their known database to send the ticket to. Although this is more related to Object Character Recognition than Image Classification, both uses computer vision and neural networks as a base to work.

Nowadays internet is filled with an uncountable of images and videos, which is encouraging the development of search applications and algorithms that can examine the semantic analysis of image and videos for presenting the user with better search content and their summarization. There have been major breakthroughs in image labeling, object detection, scene classification, areas reported by different researchers across the world. This leads to making it possible to formulate approaches concerning object detection and scene classification problems. Since artificial neural networks have shown a performance breakthrough in the area of object detection and scene classification, especially convolutional neural networks (CNN), this work focuses on identifying the best network for this purpose. Feature extraction is a key step of such algorithms. Feature extraction from images involves extracting a minimal set of features containing ahigh amount of object or scene information from low-level image pixel values, therefore, capturing the difference among the object categories involved. Some of the models are AlexNet, GoogleNet, VGG-16/VGG-19, MobileNet and so on.

Nowadays internet is filled with anabundance of images and videos, which is encouraging thedevelopment of search

applications and algorithms that can examine the semantic analysis [1] of image and videos for presenting the user with

better sear ch content and their summarization. There have been major br eakthroughs in image labeling, object detection,

scene classification [2] [3], areas reported by different researchers across the world. This leads to making it possible to

formulate approaches concerning object detection and scene classification problems. Since artificial neural networks have

shown a performance breakthrough in the area of object detection and scene classification, specially convolutional neural

networks ( CNN)[4] [5] [6], this work focuses on identifying the best network for this purpose. Feature extraction is a key

step of such algorithms. Feature extraction from images involves extracting a minimal set of features containing ahigh

amount of object or scene information fr om low -level image pixel values, therefore, capturing the difference among the

object categories involved. Some of the traditional feature extraction techniques used on images are Scale-invariant feature

transform (SIFT) [7], histogram of oriented gradients (HOG) [8], Local binary patterns (LBP) [10], Content-Based Image

Retrieval (CBIR) [11], etc. Once features are extracted their classification is done based on objects present in an image. A

few examples of classifiers are Support vector machine (SVM), Logistic Regression, Random Forest, decision trees etc

## **2. Literature Review**

**AlexNet**

The Alexnet has eight layers with learnable parameters. The model has 3 fully connected layers with a combination of max pooling and in each of these layers use Relu activation except the output layer.

It was found out that using relu as an activation function accelerated the training process by 6times.Dropout layers was also used to prevent their model from overfitting.

**GoogLeNet**

In CNN field image classification is the process that takes an input (ex: a picture) and gives the output a class name (ex: cat). GoogLeNet is one of the image classification models, which was introduced in 2014. In addition, this image classification model is 22 layers deep and it is a deeper model since 2014 in CNN history. So, we can say this model is a deeper image classification model but not the deepest. Nowadays GoogLeNet is used for other computer vision problems such as face detection and recognition etc.

The GoogLeNet Network architecture model is 22 layers deep (27 layers including pooling layers) and there are nine inception modules in total [1]. Nowadays, the network model not only trained on ImageNet classifies images into 1000 object categories, such as keyboard, mouse, pencil, and many animals but also classifies images into 365 different place categories, such as field, park, runway, and lobby [2].

**VGGNet**

VGGNet comes after AlexNet. In 2014, Simonyan and Zisserman introduced VGGNet in the paper named “Very Deep Convolutional Networks for Large-Scale Image Recognition “. It has only 3x3 convolutions as AlexNet, but it has lots of filters. It has 138 million parameters.

**VGG-16**

This model has 92.7% top-5 test accuracy in ImageNet, which was a dataset of over 14 million images belonging to 1000 classes.

VGG-16 obtains an 8.8% error rate which means the deep learning network is still improving by adding a number of layers. The 16 in VGG16 refers to it has a total of 16 layers that have weights.

**VGG-19**

VGG-19 is a convolutional neural network that is 19 layers deep and can classify images into 1000 object categories such as a keyboard, mouse, and many animals. The model trained on more than a million images from the ImageNet database with an accuracy of 92%.

VGG-19 has a 9.0% error rate which means the deep learning network is NOT improving by adding a number of layers. Thus, the authors stop adding layers.

**ResNet**

Deeper neural networks are more difficult to train. We present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. We explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. We provide comprehensive empirical evidence showing that these residual networks are easier to optimize, and can gain accuracy from considerably increased depth. On the ImageNet dataset we evaluate residual nets with a depth of up to 152 layers---8x deeper than VGG nets but still having lower complexity. An ensemble of these residual nets achieves 3.57% error on the ImageNet test set. This result won the 1st place on the ILSVRC 2015 classification task. We also present analysis on CIFAR-10 with 100 and 1000 layers.[1]

The depth of representations is of central importance for many visual recognition tasks. Solely due to our extremely deep representations, we obtain a 28% relative improvement on the COCO object detection dataset. Deep residual nets are foundations of our submissions to ILSVRC & COCO 2015 competitions, where we also won the 1st places on the tasks of ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation.

**MobileNet**

For portable and embedded vision applications, we offer MobileNets, a family of efficient models. MobileNets are built on a simplified design that builds light weight deep neural networks using depth-wise separable convolutions. Two simple global hyper-parameters are introduced that efficiently trade off latency and accuracy. These hyper-parameters help the model builder to select the appropriate model size for their application depending on the problem constraints. We report comprehensive studies on resource and accuracy tradeoffs, as well as good results on ImageNet classification when compared to other popular models. The efficacy of MobileNets is then demonstrated across a variety of applications and use cases, including item identification, fine-grain classification, face characteristics, and large-scale geo-localization.

## **3**. **Proposed Method**

**AlexNet**

The first convolution layer with 96 filters of size 11X11 with stride 4. The activation function used in this layer is relu. The output feature map is 55X55X96.

output size of a convolution layer = ((Input-filter size)/ stride)+1

Alexnet used Rectified Linear Units (ReLU) instead of the tanh function, which was standard at the time. ReLU’s advantage is in training time; a CNN using ReLU was able to reach a 25% error on the CIFAR-10 dataset six times faster than a CNN using tanh.

Back in the days Gpu’s were around 3 gigabytes.And Alexnet allowed multi Gpu training by putting half of the model’s neurons on one GPU and the other half on another GPU.

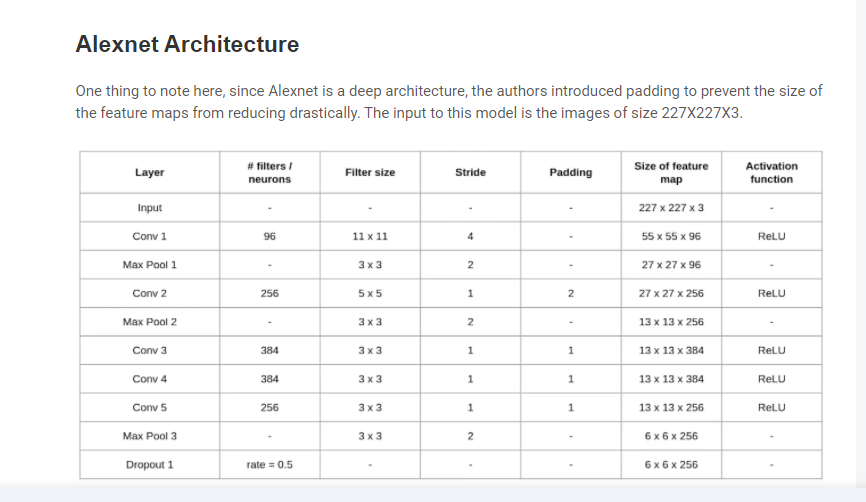


Figure:1 AlexNet Architecture

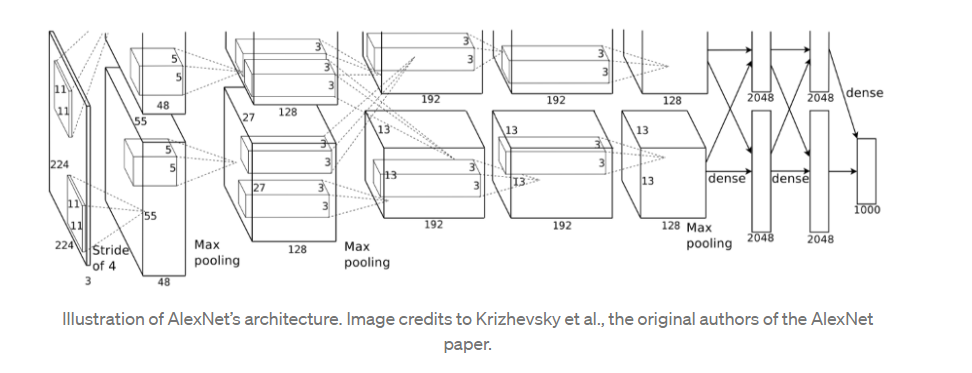
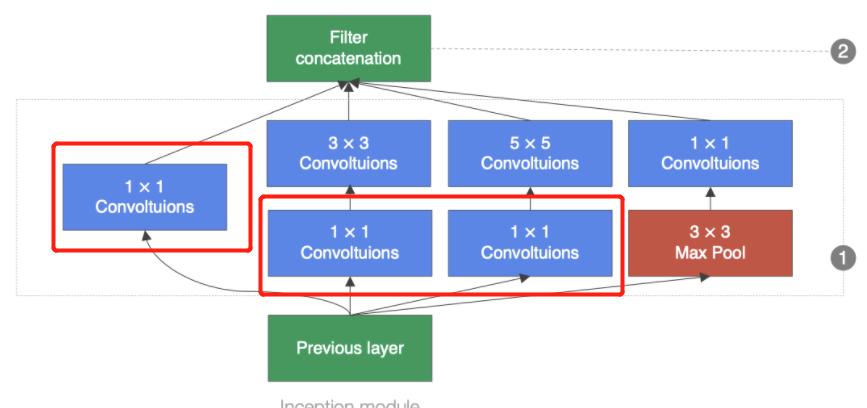


Figure 2: ALexNet Model

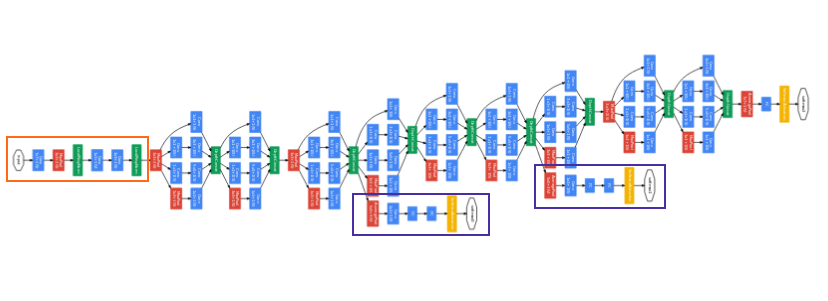
**GoogleNet**

Being an image classification model, it has some unique characteristics and features. This model architecture proposed method consists of 22 layers and parts of these layers are total of 9 inception modules (***Figure 1***). The GoogLeNet architecture input layer takes an image with the dimension 224 ×224.



**Figure 1:** Inception module of GoogLeNet

Also, the patch size refers to the sweeping window have equal height and width across convolution and pooling layers. The resulting output dimensions of current architecture components maps after the input is passed through the layers [2].

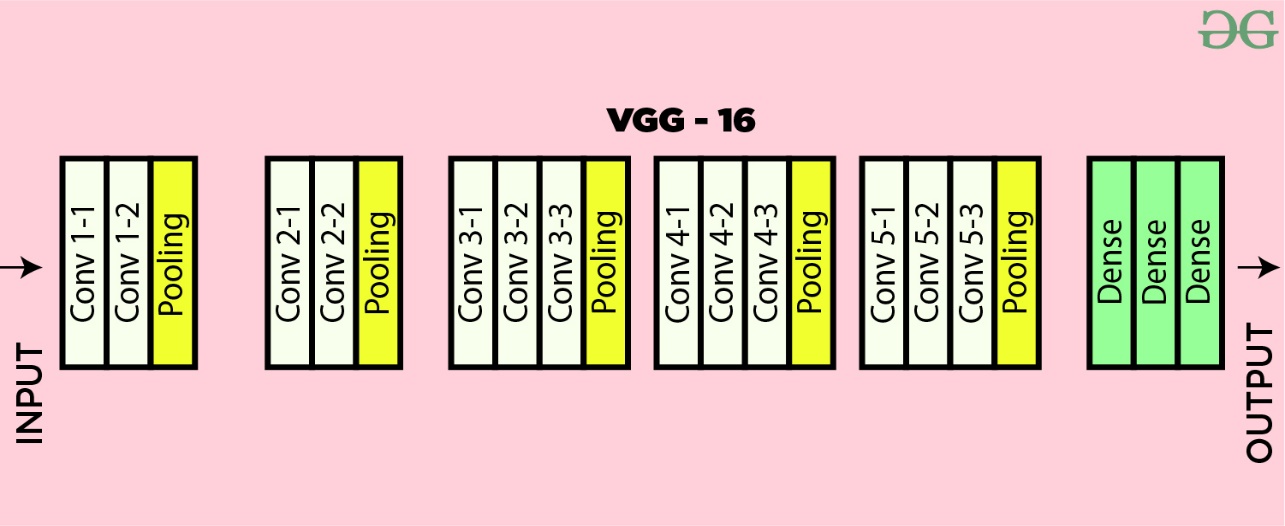


**Figure 2:** Architecture of the GoogLeNet

**VGGNet**

The input to the network is an image of dimensions (224, 224, 3). The first two layers have 64 channels of 3\*3 filter size and the same padding. Then after a max pool layer of stride (2, 2), two layers have convolution layers of 256 filter size and filter size (3, 3). This followed by a max-pooling layer of stride (2, 2) which is the same as the previous layer. Then there are 2 convolution layers of filter size (3, 3) and 256 filters. After that, there are 2 sets of 3 convolution layers and a max pool layer. Each has 512 filters of (3, 3) size with the same padding. This image is then passed to the stack of two convolution layers. In these convolution and max-pooling layers, the filters we use are of the size 3\*3 instead of 11\*11 in AlexNet and 7\*7 in ZF-Net. In some of the layers, it also uses 1\*1 pixel which is used to manipulate the number of input channels. There is a padding of 1-pixel (same padding) done after each convolution layer to prevent the spatial feature of the image.

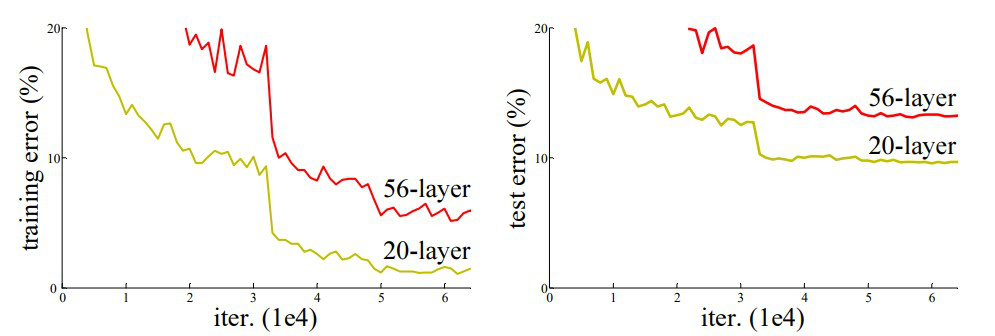
 **Figure:** VGGNet Architecture



**Figure 2:** VGGNet 16 Layers

**RestNet**

After the first CNN-based architecture (AlexNet) that win the ImageNet 2012 competition, Every subsequent winning architecture uses more layers in a deep neural network to reduce the error rate. This works for less number of layers, but when we increase the number of layers, there is a common problem in deep learning associated with that called Vanishing/Exploding gradient.[2] This causes the gradient to become 0 or too large. Thus when we increases number of layers, the training and test error rate also increases.



In the above plot, we can observe that a 56-layer CNN gives more error rate on both training

Figure 1

and testing dataset than a 20-layer CNN architecture, If this was the result of over fitting, then we should have lower training error in 56-layer CNN but then it also has higher training error. After analyzing more on error rate the authors were able to reach conclusion that it is caused by vanishing/exploding gradient. ResNet, which was proposed in 2015 by researchers at Microsoft Research introduced a new architecture called Residual Network.

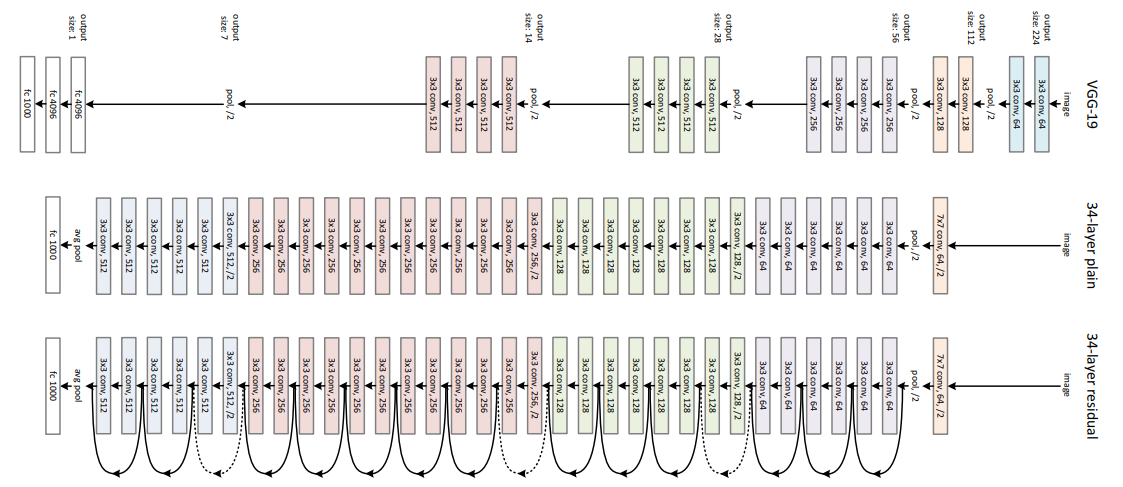
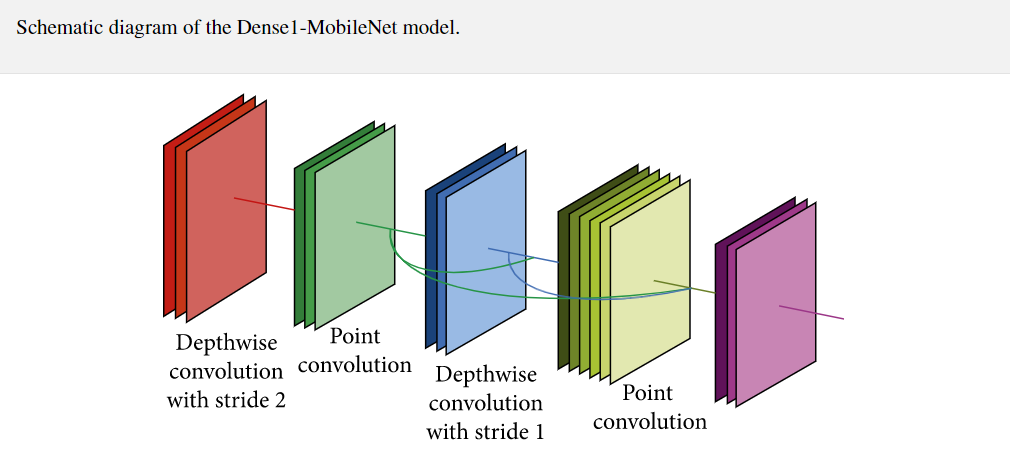


Figure 2: architecture of resnet

**MobileNet**

The MobileNet model is a network model in which the basic unit is depthwise separable convolution. It has two layers of depthwise separable convolution: depthwise convolution and point convolution. The depthwise convolution layer and the point convolution layer are two distinct convolution layers in the Dense1-MobileNet model. As shown in Figure , the input feature maps of each depthwise convolution layer in the dense block are superpositions of the output feature maps of the previous convolution layer, and the input feature maps of each deep convolution layer are also superpositions of the output feature maps of the previous convolution layer. Because depthwise convolution is a single-channel convolution, it has a limited range of applications.



The number of output feature maps in the middle depthwise convolution layer is the same as the number of input feature maps, which is the total of all previous layers' output feature maps. Between two dense blocks in DenseNet, there is a transition layer. The transition layer uses a 1 \* 1 convolution kernel to minimize the number of input feature maps and a 2 \* 2 average pooling layer to halves the number of input feature maps.

## **4. Discussion**

**AlexNet**

**Advantages:**

* First major CNN model that used GPU’s for training. This lead to faster training of models.
* Deeper architecture with 8 layers which means that is better able to extract features when compared to CNN. It also worked well for the time with color images.
* The ReLu activation function used in this network has 2 advantages. It does not limit the output unlike other activation functions.
* It negates the negative output of summation of gradients and not the dataset itself. This means that it will further improve model training speed since not all perceptrons are active.

**Disadvantages:**

* We can see that it takes more time to achieve higher accuracy results compared to future models.

**GoogLeNet**

This image classification model already learned different class representations of a wide range of images. Some advantages of this network model given below

**Advantages:**

* GoogleNet trains faster than VGG.
* The size of a pre-trained GoogleNet is comparatively smaller than VGG. A VGG model can have >500 MBs, whereas GoogleNet has a size of only 96 MB.
* GoogLeNet reduces the error rate to 6.656%.
* GoogLeNet is 22 layers deeper image classification model.

**Disadvantages:**

* GoogLeNet is not the deepest model.

**VGGNet**

**Advantages:**

* VGG brought with it a massive improvement in accuracy and an improvement in speed as well. This was primarily because of improving the depth of the model and also introducing pre-trained models.
* VGG brought with it various architectures built on a similar concept. This gives more options to us as to which architecture could best fit our application.
* .Don't require mastery in Deep Learning to use pre-trained models.

**Disadvantages:**

* More heavier model
* More training time
* Vanishing gradient problem
* It will take a large number of resources(time and computation power) to train big models from scratch

**RestNet**

A similar approach to ResNets is known as “highway networks”. These networks also implement a skip connection, however, similar to an LSTM these skip connections are passed through parametric gates. These gates determine how much information passes through the skip connection. The authors note that when the gates approach being closed, the layers represent non-residual functions whereas the ResNet’s identity functions are never closed. Empirically, the authors note that the authors of the highway networks have not shown accuracy gains with networks as deep as they have shown with ResNets.

**Advantage**

* Networks with large number (even thousands) of layers can be trained easily without increasing the training error percentage.
* ResNets help in tackling the vanishing gradient problem using identity mappin.

**Disadvantage**

* Computationally expensive: Neural networks are computationally expensive than any other traditional algorithms.
* Determination of proper network structure: There is no specific rule for determining the structure of a neural network. The appropriate network structure is achieved through experience and trial and error.
* The duration of the network is unknown: Reducing the network to a certain value of the sampling error implies completing the training.

**MobileNet**

MobileNet model with dense blocks for image classification. The dense blocks are used as the basic structure to improve the structure of MobileNet, and two improved models are proposed. These two models can reduce the parameters and calculation by setting the hyperparameter growth rate. At the same time, experiments show that Dense2-MobileNet can also increase the accuracy of classification. Compared with the MobileNet model, although the classification accuracy of Dense1-MobileNet is reduced, it reduces the number of parameters by at least half and the amount of calculation by nearly half. Generally speaking, the models proposed in this paper can be better applied to mobile devices.

**Advantage :**

* MobileNets are small deep neural networks that are well-suited to mobile and embedded vision applications.
* MobileNets are built using depthwise separable convolutions and a simplified design.
* MobileNet makes use of two basic global hyperparameters to effectively balance accuracy and latency.

## **5. Conclusion**

In this report, we talk about some image classification models and their proposed method since the journey of CNN (Convolutional Neural Network) began. Also, we try to explain how the proposed method work to solve the real-life problem with the desire output result. In addition, we try to discuss all the advantages and disadvantages of each and every image classification model listed in this report. Lastly, we hope that we are able to give a brief overview of the image classification model's history since the journey began and the current development picture of this field.

## **References**

1] <https://www.analyticsvidhya.com/blog/2021/03/introduction-to-the-architecture-of-alexnet/>

[2] <https://towardsdatascience.com/alexnet-the-architecture-that-challenged-cnns-e406d5297951>

|  |  |
| --- | --- |
| [3] | R. Alake, "Deep Learning: GoogLeNet Explained | by Richmond Alake | Towards Data Science," [Online]. Available: https://towardsdatascience.com/deep-learning-googlenet-explained-de8861c82765. |
| [4]  [5] | "GoogLeNet Model | Deeplearning Convolutional Neural Network," [Online]. Available: <https://www.mathworks.com/help/deeplearning/ref/googlenet.html>.  Simonyan, K., Zisserman, A.. Very deep convolutional networks for large- scale image recognition. arXiv preprint arXiv:14091556 2014;. |
| [6] | https://medium.com/analytics-vidhya/cnns-architectures-lenet-alexnet-vgg-googlenet-resnet-and-more-666091488df5 |

[7] Bai, H.; Shi, G. Gas Sensors Based on Conducting Polymers. Sensors **2007**, 7, 267–307. [CrossRef]

[8] Szulczy ´ nski, B.; Armi´ nski, K.; Namie´snik, J.; G˛ebicki, J. Determination of Odour Interactions in Gaseous Mixtures Using Electronic Nose Methods with Artificial Neural Networks. Sensors **2018**, 18, 519. [CrossRef] [PubMed]

|  |  |  |
| --- | --- | --- |
| Name | Id | Model |
| Antu Asif Ikbal | 17-34554-2 | AlexNet |
| Islam, Dewan Saiful | 18-37623-1 | GoogleNet |
| Rakibul Newaz Sourav | 18-37316-1 | VGGNet (16,19) |
| Chowdhury, Abdur Rahim | 18-37629-1 | RestNet |
| ABDUR RAKIB HOWLADER | 18-37629-1 | MobileNet |