2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON) | 978-1-6654-0017-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/CENTCON52345.2021.9687944

A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification

Sheldon Mascarenhas
Lufthansa Technik Services IndiaPvt Ltd.
Bangalore, India.
sheldonoswin@gmail.com

Mukul Agarwal
Deloitte (India)
Mumbai, India
mukul.agrawal06@gmail.com

***Abstract*—Artificial Intelligence advancements have come a long way over the past twenty years. Rapid developments in AI have given birth to a trending topic called machine learning. Machine learning enables us to use algorithms and programming techniques to extract, understand and train data. Machine learning led to the creation of a concept called deep learning which uses algorithms to create an artificial neural network and use it to develop and learn, based on which it makes intuitive decisions by itself. Image classification is a task where we classify the images into sets of different categories, which when performed using deep learning increases business productivity by saving time and manpower. In this paper, we intend to determine which model of the architecture of the Convoluted Neural Network (CNN) can be used to solve a real-life problem of product classification to help optimize pricing comparison. We have compared the VGG16, VGG19, and ResNet50 architectures based on their accuracy while all three of these models solve the same image classification problem. We have concluded that the ResNet50 is the best architecture based on the comparison. These models have provided accuracies of 0.9667, 0.9707, and 0.9733 for VGG16, VGG19, and ResNet50 at epoch 20. The data provided is a real-life data set, sourced from a regional retailer.**

Keywords—Computer vision, Artificial Intelligence, Deep Learning, Convolution Neural Network, CNN architecture, VGG16, VGG19, ResNet50.

I. INTRODUCTION

Artificial intelligence (AI) is a field of technology that is widely being explored to work on tasks and aspects that could never have been imagined. By verbatim, AI implies: mimicking human intelligence. Artificial intelligence on its merit can take decisions and execute them, based on given historic data or by adapting to dynamic real-time data. Machine learning as a subset of Artificial intelligence, using algorithms and programming enables the given intelligent system can make its own decisions based on its input data and conditional environment. To make these AI systems independent, Machine learning which requires human oversight to achieve the most effective outcome uses Deep learning.

Deep learning algorithms are equipped with Artificial Neural Networks which take the human component out of the AI system equation. Deep learning is a key Machine learning subset used especially in computer vision, helping digital technologies to extract information from digital media such as images, videos, gifs, etc. Deep learning also helps in recognizing speech, detecting written and spoken language patterns, and helps in automated vehicular driving with image detection. At the heart of all this lie neural networks, which as the name suggests mimic biological neural networks. Neural networks primarily consist of the following important constituent parts: Input, weight, threshold, and output. These neural networks, when arranged in more than three layers including both inputs and outputs, constitute deep learning. The whole purpose of deep learning is to create automated and trained models that reduce human intervention by great margins. Intelligent learning is enhanced with training and perfecting the algorithm and arranging those layers to achieve maximum accuracy. These models can take up clusters of data and can also employ more data points to increase accuracy. Artificial Neural networks are the most common computational models that have been in use for quite some time now. They are capable of making changes as and when they receive new input, using advanced learning algorithms. More importantly, it is a feed-forward neural network, sending only information forward from the neurons to deeper layers. With the help of rapid serial transformations, the information is easily adapted and transmitted from one hidden layer [11] to another. ANNs output can be calibrated using the information of an output error relayed during the training phase. With time ANN learns how to reduce the chances of error thus increasing efficiency. ANN requires explicitly devised set data points. Moreover, conversion of 2 D images to 1D vectors requires a lot of parameters that in turn require a lot of storage and processing power. Thus ANN serves as a poor neural network for image classification.

Thus, we use another neural network called the Convolution Neural Network. Unlike the ANN instead of deploying neurons or weights, the network uses filters to feature map the image submitted into the network. As a result, a vector output is developed. The CNN by identifying and comparing patterns detects the images. Instead of pushing the data forward, the same data is run multiple times with filters to create a feature map. Since CNN [14] extracts features from the images by itself, it becomes a suitable model when thousands of features have to be extracted and gathered. All of these properties make a CNN model more desirable for image classification [10]. The CNN models exist with multiple architectures adapting and changing with time, each having its own sets of features and efficiencies. To evaluate which model would be the best to solve the problem a detailed comparison of features and benchmarks is required [1].

978-1-6654-0017-6/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY SURATHKAL. Downloaded on April 13,2025 at 05:59:04 UTC from IEEE Xplore. Restrictions apply.

II. Convolution Neural Network (CNN)

One of the most popular applications of advanced computation is image recognition and computer vision. To read and understand the image it uses RGB (Red, Green, Blue) channels. An image size represented as A × B × 3, it means that the given image has A columns, B rows and 3 color channels. Since we have a huge pixel density in most images, a fully connected neural network would require a high number of weights in its first hidden layer itself [2]. This is where a Convolution neural network gets to shine, by having one of its neurons connected only to a specific region of layer next to it . Computers detect and extract data from these images by comparing them with a reference image pixel by pixel, thus helping them to relate the reference images to deformed images. To compare these images, the computer detects the image section by section, and these sections are called features. This sectional comparison makes image detection and identification much easier than matching entire images.

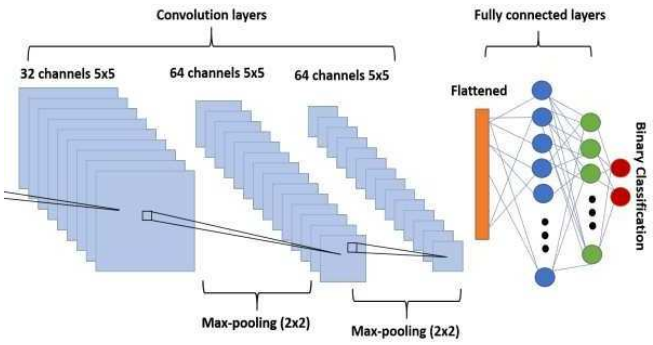

Fig. 1. A pictorial representation of CNN, its components, and mathematical operators [2].

The first layer of the CNN is the Convolution layer is used toextract the features of the image by moving a filter of a specific size over the image and creating a feature map. The feature map created is a matrix of pixel values positive and negative as created by the mathematical operations that are run upon moving the filter. The next layer, called the Rectified linear activation function or ReLU, receives the output from the previous layer as input and produces the input itself as an output if positive. However, if the input pixel value is found to be negative then the output of the ReLU becomes zero. This is followed by the Pooling layer trying to reduce the size of the feature map. This also reduces the time of computation and thus makes the process faster. Pooling features can be executed with different pooling operations such as Max Pooling and Average Pooling. When the output from the Pooling layer passes through a combination of Convolution layer, ReLU layer, and Pooling layer the input shrinks in size further. The last layer of the CNN is the fully connected layer which forms the final layer of any neural network. As a result, it is not specifically a feature unique to CNN. The input vector to the fully connected layer is subjected to an activation function (logistic or Softmax), to calculate the probabilities. A higher probability would indicate the presence of the required feature of the image, thus indicating a successful image detection.

III. Transfer Learning

When training large datasets Deep Neural networks would require a period of several days to weeks. To avoid such a situation, we use pre-trained models [9]. These pre-trained models reduce errors and the time required for training. The pre-trained models use weights that are re-used in layers to adapt to the new problem at hand. A part of the pre-trained model or the entire model can be conjoined into the new neural network model [3]. The weights associated with the pre-trained model can be frozen to make sure that they remain unchanged. The models we intend to use are the VGG network models and a ResNet model, which are pre- trained models. To perform image recognition tasks, these pre-trained models can be accessed from Keras API.

IV. Dataset

Fig. 2. The dataset had 6000 images and some of those images have been provided within this image. The key task of resizing and reducing the size of the image to make it compatible with the CNN architectures was a principal operation in the methodology.

To compare any given architecture of the Convolution Neural network, we require a commonly applicable data set. The application of these models will help us solve a real-life problem that will optimize pricing comparison for a regional retail network. To compare the prices the objects and products concerned need to be classified and segregated into the correct categories [13]. The data set includes both textual data in the form of title and description, along with product images. To classify these products with the help of their image we have developed deep learning models employing different CNN architectures. The dataset had 6000 images to be classified. The data set has also been used to create separate train and test data sets, upon which the accuracy of the CNN model is checked. The data set is also used to determine which CNN architecture is the most effective in terms of image classification. The data provided is a real-life data set, which was used by a regional retailer.

V. CNN Architectures

The Convolution neural network with an architecture of VGG16 was introduced by NK. Simonyan and A. Zissermanfrom the University of Oxford in the paper “Very Deep Convolutional Networks for Large-Scale Image Recognition” [4]. The model was able to achieve a test data accuracy of 92.77%, with ImageNet, where the dataset has a total of 14 million images, with a total number of classes of 1000. Images of size 224×224×3, enter 2 convolution networks, a max-pooling layer, followed by 2 convolution

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY SURATHKAL. Downloaded on April 13,2025 at 05:59:04 UTC from IEEE Xplore. Restrictions apply.

layers and one more Max Pooling layer. Following this the architecture is provided with 3 Convolution layers, 1 MaxPooling layer, 3 Convolution layers, another Max Pooling layer, 3 Convolution layers, and 1 more Max Pooling layer. This is followed by fully connected and ReLU layers. The number of filters keeps changing with the layers. The convolution layer has a 3×3 filter size with a stride of 1 while the Max Pooling layer has a filter size of 2×2 and a stride of 2. Similarly, the other VGG 19 architecture consists of (3 Fully connected layers, 16 convolution layers, 1 SoftMax layer, and 5 MaxPool layers). The number of filters in the convolution layers includes 64, 128, and 256.

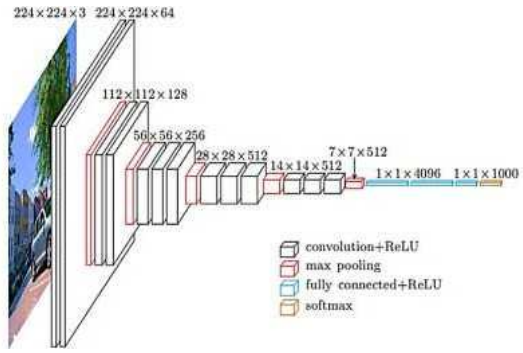

Fig. 3. A complete opened-up view of the VGG16 Architecture. VGG 19 has been provided with additional layers. They all belong to the same category of Pre-trained models, requiring input in the form of 224×224×3 [5].

ResNet50 is another convolutional neural network-based model, consisting of 48 convolution layers. The other two layers are 1 Max Pooling layer and 1 Average Pooling layer. ResNet architecture allowed CNN to operate with multiple layers. Deep neural networks with multiple stacked layers tended to produce greater training error percentages than models with lesser layers. This residual network framework allowed the addition of shortcut connections and usage of residual functions, thus allowing stacked layers of deep neural networks to reduce their training errors. The direct connection set in this architecture helps skip some of the layers of the model 0.

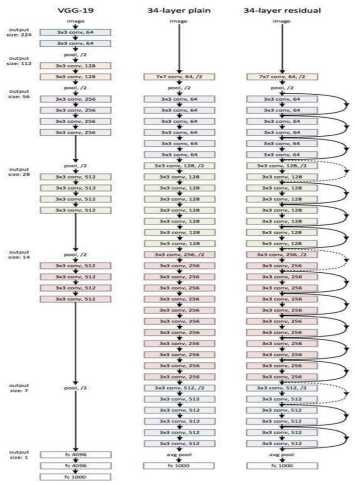

Fig. 4. A Comparative representation of VGG19, and ResNet 34 architecture. The given Image intends to represent how a residual network is organized to produce effective results. We have used a ResNet 50 model for the image. [7].

VI. Methodology

The given data set is classified into two steps. The first step involves feature extraction which extracts the features of the image to identify them. The next step is to classify the images with the help of Exploratory Data Analysis.

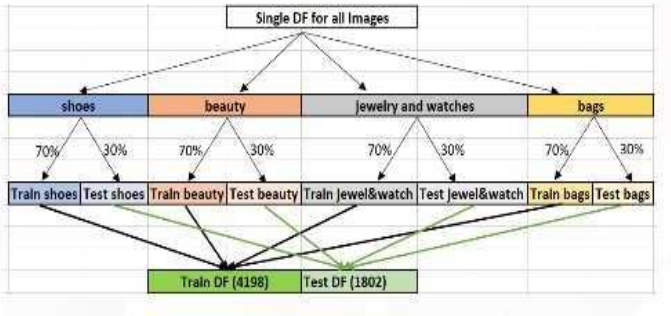

Fig. 5. The dataset needs to be split into 2 sections, train and test data sets where the train data set is used to train the model and the test dataset is used to test the validatory accuracy for the given. Model. The split ratio is respectively 70:30.

We discovered the existence of 6000 images that needed to be classified into 5 different unique categories of shoes, beauty, jewelry, watches, and bags. Each of these categories was assigned integer vectors to be converted into hot encoded vectors. The data set was split with a ratio of 70:30 into testing and training datasets. The images that were found in the data set were clearly of greater size than the ones with which the convolution layers could be input with and hence they had to be compressed to the size of 224×224×3. The features of the images were extracted separately with the help of VGG16, VGG19, and ResNet50 [12]. The data set was run with an Epoch of 1, 3, 5, 10, 15, 20, and 100. The accuracies of each architecture were then observed and compared.

VII. Results and Discussions

TABLE I. TABLE REPRESENTING ACCURACIES OF DIFFERNET MODELS FOR VARYING EPOCH NUMBERS.

epoch	VGG 16 (Training data Accuracy)	VGG 16 (Test data Accuracy)	VGG19 (Training data Accuracy)	VGG 19 (Test data Accuracy)	ResNet 50 (Training data Accuracy)	ResNet 50 (Test data Accuracy)
1	0.8516	0.9367	0.843	0.924	0.9176	0.9656
3	0.9552	0.9545	0.935	0.9489	0.978	0.9716
4	0.9621	0.9478	0.9574	0.9556	0.9762	0.9472
5	0.9664	0.9578	0.9707	0.9672	0.9814	0.9633
10	0.9852	0.9589	0.9859	0.965	0.9971	0.9722
15	0.9912	0.9639	0.9931	0.9744	0.9938	0.9683
20	**0.9919**	**0.9667**	**0.9936**	**0.9707**	**0.9976**	**0.9733**
100	1	0.9633	1	0.9689	1	0.9533

The process of image classification was run on Core i5-43100U (CPU) on 64-bit Windows 10 OS. All the image

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY SURATHKAL. Downloaded on April 13,2025 at 05:59:04 UTC from IEEE Xplore. Restrictions apply.

classification and image recognition models used have been run with a Keras API. The activation function used in all the above architectures is Softmax and the Loss function used is Categorical cross-entropy used for fine-tuning. Accuracy is based on the confusion matrix [8] developed. The training data accuracy has been determined while fitting the model using the training data, and the test data accuracy is based on the confusion matrix developed. The selected batch for epoch is 32 [1]. Epoch 20 VGG19 records a training data accuracy of 99.36% implying that the model is exhibiting an accuracy of image classification with the training data set. At Epoch 20, VGG 19 records an accuracy of 97.07 %, implying that the model with this accuracy can classify images of the test data set. A greater value of training data set accuracy indicates a better model.

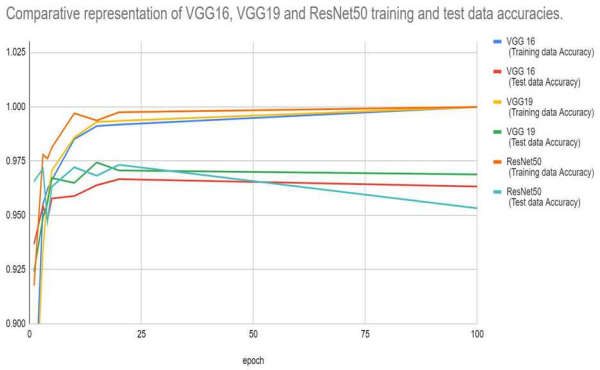

Fig. 6. Comparison of accuracies of the different architectural models. The given comparison includes both training and test accuracy.

All the CNN models exhibit a high accuracy of test data [1], but it is ResNet50 that exhibits the highest accuracy [8] followed by VGG 19 and then the VGG16. It is important to note when epoch values range from 15 to 20, CNN models showcase the highest and absolute fit. As indicated by Figure 6 and TABLE I. we notice that as the Epoch values go above 20, the models start overfitting and when below 15 they show an underfitting. Fig. 6 also shows a similar trend where optimal accuracy is represented in between Epoch 15 to 20. Based on these tests, training accuracy comparison we can conclude that the ResNet 50 model exhibits the best accuracy for the task of image classification with a test data accuracy of 97.33%.

VIII. Conclusion and Future Scope

The result of this study has clearly shown us that the ResNet50 model has the best accuracy for image classification. The applications of this study extend from health care to industrial sectors. These models of CNN can be used to detect diseases [5] and better the model, better is the accuracy of its visual detection [3]. Increasing accuracies in health care can be of life saving importance. Industrial and market applications could range from identification and classification of raw materials to wholesale goods in the retail industry. These methods save time and revenue to a great extent. These models discussed can be streamlined into creating object detection platforms that have multiple models working on real time data. Along with object detection and classification if the platform provides a comparative study of accuracies of these models, the user could adopt the model providing them the best results.

References

[1] TTheckedath, D., Sedamkar, R.R. Detecting Affect States Using VGG16, ResNet50 and SE-ResNet50 Networks. *SN COMPUT. SCI.* 1, 79 (2020). https://doi.org/10.1007/s42979-020-0114-9.

[2] S. Khaleghian, H. Ullah, T. Kræmer, N. Hughes, T. Eltoft, and A. Marinoni, “Sea Ice Classification of SAR Imagery Based on Convolution Neural Networks,” *Remote Sensing*, vol. 13, no. 9, p. 1734, Apr. 2021.

[3] Shin, Hoo-chang & Roth, Holger & Gao, Mingchen & Lu, Le & Xu, Ziyue & Nogues, Isabella & Yao, Jianhua & Mollura, Daniel & Summers, Ronald. (2016). Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. IEEE Transactions on Medical Imaging. 35. 10.1109/TMI.2016.2528162.

[4] Simonyan, Karen & Zisserman, Andrew. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 1409.1556.

[5] Pravitasari, A. A., Iriawan, N., Almuhayar, M., Azmi, T., Irhamah, I., Fithriasari, K., Purnami, S. W., & Ferriastuti, W. (2020). UNET-VGG16 with transfer learning for MRI-based brain tumor segmentation. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *18*(3), 1310. https://doi.org/10.12928/telkomnika.v18i3.14753.

[6] Q. A. Al-Haija and A. Adebanjo, "Breast Cancer Diagnosis in Histopathological Images Using ResNet-50 Convolutional Neural Network," 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 2020, pp. 1-7, doi: 10.1109/IEMTRONICS51293.2020.9216455.

[7] Vu, Tuan-Hung. (2018). Learning visual models for person detection and action prediction.

[8] W. Li et al., "Classification of High-Spatial-Resolution Remote Sensing Scenes Method Using Transfer Learning and Deep Convolutional Neural Network," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 1986-1995, 2020, doi: 10.1109/JSTARS.2020.2988477.

[9] R. Pires de Lima and K. Marfurt, “Convolutional Neural Network for Remote-Sensing Scene Classification: Transfer Learning Analysis,” Remote Sensing, vol. 12, no. 1, p. 86, Dec. 2019.

[10] J. Naranjo-Torres, M. Mora, R. Hernández-García, R. J. Barrientos, C. Fredes, and A. Valenzuela, “A Review of Convolutional Neural Network Applied to Fruit Image Processing,” Applied Sciences, vol. 10, no. 10, p. 3443, May 2020.

[11] Altwaijry, N., Al-Turaiki, I. Arabic handwriting recognition system using convolutional neural network. Neural Comput & Applic 33, 2249–2261 (2021),https://doi.org/10.1007/s00521-020-05070-8.

[12] Yang Liu, Zelin Zhang, Xiang Liu, Lei Wang, Xuhui Xia, Deep learning-based image classification for online multi-coal and multi-class sorting, Computers & Geosciences, Volume 157, 2021, 104922, ISSN 0098-3004, https://doi.org/10.1016/j.cageo.2021.104922.

[13] J. R. Rajayogi, G. Manjunath and G. Shobha, "Indian Food Image Classification with Transfer Learning," 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), 2019, pp. 1-4, doi: 10.1109/CSITSS47250.2019.9031051.

[14] B Bansal, M., Kumar, M., Sachdeva, M. et al. Transfer learning for image classification using VGG19: Caltech-101 image data set. J Ambient Intell Human Comput, 2021.https://doi.org/10.1007/s12652-021-03488-z.

.

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY SURATHKAL. Downloaded on April 13,2025 at 05:59:04 UTC from IEEE Xplore. Restrictions apply.