Image Classification: A Brief History and Model Architectures

Shaily Sarker (ID: 17-35951-3)[a], Kowshik Chakraborty (ID: 18-36200-1)[a], Saiful Islam (ID:18-37404-1)[a]

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

Abstract

Image Processing includes changing the nature of an image in order to improve its pictorial information for human interpretation, for autonomous machine perception. Machine Learning (ML) generally means that you're training the machine to do something (here, image processing) by providing set of training data's. In this report we will know about the image classification using deep learning, Convolutional neural networks (CNN) have been widely used in automatic image classification systems. In most cases, features from the top layer of the CNN are utilized for classification; however, those features may not contain enough useful information to predict an image correctly. Also with deep learning techniques, a revolution has taken place in the field of image processing and computer vision. The survey paper emphasizes the importance of representation learning methods for machine learning tasks. Deep learning, the modern machine learning is commonly used in the vision tasks—semantic segmentation, image captioning, object detection, recognition, and image classification.

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

1. Introduction

In age of modern science, artificial intelligence (AI) introduces us a new world. The fields of artificial intelligence are taken us in another label of human imaginations. Machine learning is one of the most important and wonderful field of AI. By the help of ML, machines can read the human thoughts. Deep learning is a branch of machine learning and also the subset of artificial intelligence. [1]That methods have the potential in various tasks that involve handling large amounts of digital data, including image, voice and text data. Their image processing and computer vision applications are employing deep learning have achieved remarkable success in applications including the denoising, recognition, detection, and segmentation of objects.

Image classification is a vital part for image processing in deep learning. It is the fundamental task which attempts to comprehend an entire image as a whole. Its goal is to classify the image which are assigning it to find out a specific label for that image. Actually, image classification is referring to images in which only one object appears and is analyzed. In contrast, object detection involves both classification and localization tasks, and is employed to research more realistic cases in which multiple objects may exist in an image. On the other hand, computer vision is that part which trains the computer for interpreting and understanding the virtual world.

Actually, deep learning involves the utilization of computer systems which is understood as neural networks. In neural networks, the input is filtered by the hidden layers of nodes. These nodes each process the input and communicate their results to the subsequent layer of nodes. This repeats until it reaches an output layer, and also the machine provides its answer. There are different kinds of neural networks supported on how the hidden layers work. Image classification with deep learning most frequently involves Convolutional neural networks (CNNs). In CNNs, the nodes within the hidden layers don't always share their output with every node within the next layer (known as convolutional layers). Image classification has the power for allowing the machines to identify and extract features to spot from images. That's why it can learn the features to look for in images by analyzing plenty of images.

Image classification is a very challenging task to find out any image label or category. There are various type of image classifiers which are available to use for image classification purpose. In this report, we will focus on different types of image classifiers-know about them, learn them and find out the best image classifier by the help of their error rate.

For categorizing the images in deep learning, image classification is wonderful solution. As [2] image classification based on visual content is a very risk-full task and largely because there is usually large amount of intra-class variability, arising from different lightings, misalignment, non-rigid deformations, occlusion and corruptions. Actually, [3] image classification is the witnessed the evolution in computer vision algorithm.

2. Literature Review

In the purpose of image classification is to involve with the convolutional neural networks (CNNs). Here, we review the most popular [4] CNN architectures, beginning from the AlexNet model in 2012 and ending at the CapsuleNet model in 2018. By studying these architectures features is the key to help researchers to choose the suitable architecture for their target task.

AlexNet:

In deep CNN structure, AlexNet is introducing us withinside the year 2012. It is relatively reputable as it finished progressive consequences withinside the fields of picture reputation and classification. Krizhevesky first proposed AlexNet and on the alternative facet attempted to enhance the CNN gaining knowledge of potential with the aid of using growing its depth and enforcing numerous parameter optimization strategies. Only for hardware restrictions, deep CNN turned into restrained at now. To triumph over those hardware obstacles problem, GPUs (NVIDIA GTX 580) had been utilized in parallel to teaching AlexNet. Moreover, to make more potent applicability of the CNN to distinctive kinds of picture categories, the variety of feature extraction degrees grew to become prolonged from 5 in LeNet to seven in AlexNet. Regardless of the truth that intensity complements generalization for several picture resolutions, it's been overturning into that represented the most disadvantage related to the intensity. Moreover, with the aid of using decreasing the vanishing gradient problem, ReLU might be applied as a non-saturating activation feature to decorate the charge of convergence. Local reaction normalization and overlapping subsampling had been additionally done to decorate the generalization with the aid of using reducing the overfitting. To enhance the overall performance of preceding networks, different changes had been made with the aid of using the use of large-length filters (five × five and 11 × 11) inside the sooner layers. AlexNet considers importance withinside the latest CNN generations, in addition to starting a progressive studies generation in CNN applications. The architecture of AlexNet is given below:

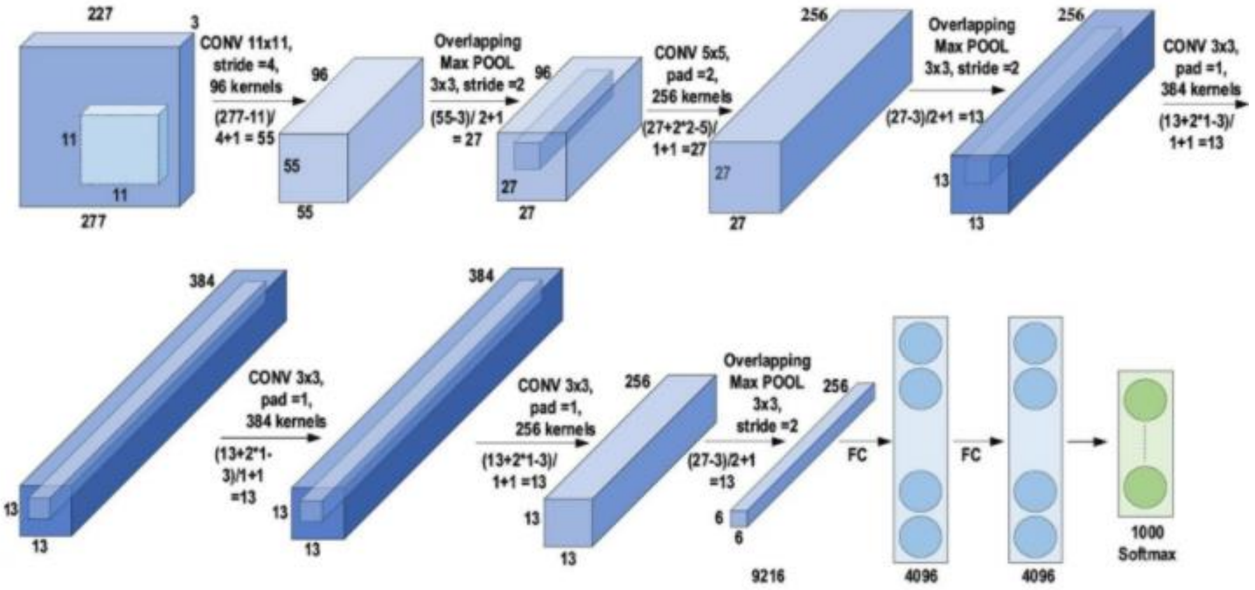

Figure 01: Architecture of AlexNet

Virtual geometry group (VGG):

After finding that CNN has been effective for photograph reputation then Simonyan and Zisserman were built in a clean and inexperienced format principle for CNN. Actually, this grasp piece layout becomes called Visual Geometry Group (VGG). It becomes introduces in year 2014. It is a multilayer version that has nineteen greater layers than AlexNet. That facilitates to simulate the family members of the community representational capability in depth. By the references of ZefNet, VGG began out to insert small sizes for boosting the CNN overall performance wherein inserted layer of the heap of 3 × 3 filters in place of the 5 × 5 and 11×11 filters in ZefNet. Here, it becomes discovered that the parallel undertaking of those small-length filters may want to produce the identical effect on because the large-length filters. By lowering the range of parameters in a filter, there had a further benefit of reducing computational hardship which becomes accomplished through the use of small-length filters. These consequences established a unique research style for walking with small-period filters in CNN. In addition, by putting 1 × 1 convolutions withinside the center of the convolutional layers, VGG regulates the community complexity. In CNN, VGG16 architectures become proposed for item reputation work. But after enhancing the layout structure of VGG16, subsequently got here VGG19 overcame the drawbacks of AlexNet and capable of will increase gadget frequency. Actually, VGG received tremendous consequences for localization issues and photograph classification. It did now no longer get the primary vicinity withinside the 2014-ILSVRC competition, as it wanted popularity because of its enlarged depth, homogenous topology, and simplicity. By the side, VGG's computational value become immoderate because of its usage of around a hundred and forty million parameters, which represented its principal shortcoming. The architecture of VGG is given below:

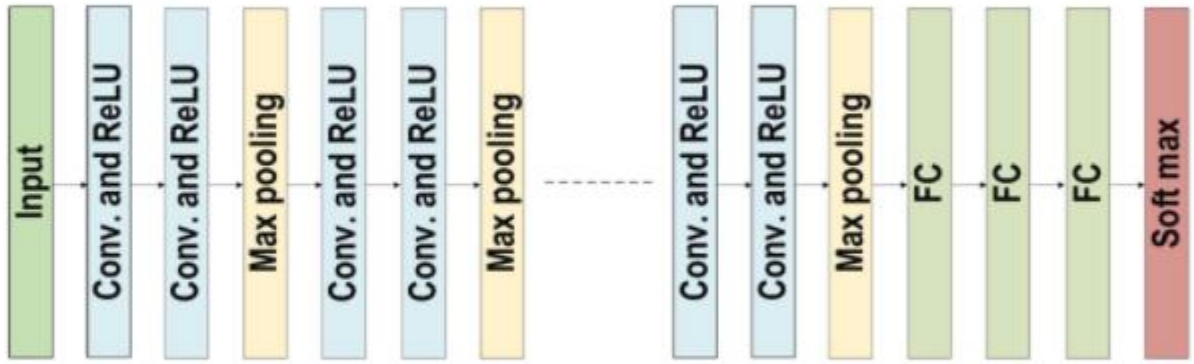

Figure 02: Architecture of VGG

ResNet:

Residual Network (ResNet) is a particular sort of CNN which turned into delivered via way of means of Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun in year 2015. The primary final results turned into [4] accumulating an ultra-deep community as a way to be freed from the vanishing gradient difficulty evaluating the preceding networks of CNN. There are diverse kinds of ResNet that have been developed. These have been primarily based totally on the variety of layers. The particular concept for ResNet is its use of the pass pathway concept. This includes the critical ResNet block diagram. This is a conventional feed in advance network with a residual connection. The residual layer output can be recognized as the $(l - 1)$ th outputs, which is probably delivered from the preceding layer $(xl - 1)$. After executing special operations, the output is $F(xl - 1)$. The completing residual output is xl, which can be mathematically represented as in equation 01-

$$xl = F\ (xl - 1) + x1 - 1 \ldots\ldots\ldots\ldots\ldots(01)$$

There are numerous number one residual blocks included inside aspect the residual network. Based on the shape of the residual network architecture, operations inside aspect of the residual block are moreover changed In the evaluation of the motorway network, ResNet supplied shortcut connections indoors layers to permit cross-layer connectivity, which is probably parameter-loose and facts independent. ResNet turned into wined the first region in ILSVRC and COCO 2015 opposition in ImageNet Detection, ImageNet localization, Coco detection, and Coco segmentation. On the alternative hand, this additionally was given the first region withinside the ILSVRC 2015 category opposition with a top-five mistakes fee of 3.57% (An ensemble model). ResNet has become the winner of the 2015-ILSVRC championship with 152 layers of intensity. It represents eight instances than the intensity of VGG and 20 instances then the intensity of AlexNet. The architecture of ResNet is given below:

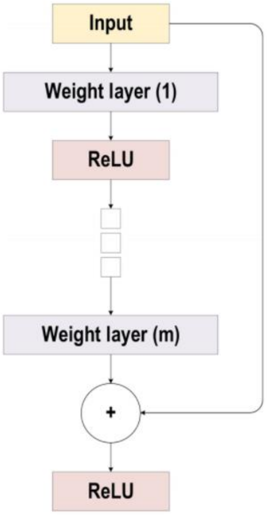

Figure 03: Architecture of ResNet

DenseNet:

DenseNet is used to solve the problem of vanishing gradient in the same way ResNet and Highway model does. A weakness of ResNet is that it preserves information by means of preservative-individuality transformation, since different layers gives very small or no information. ResNet has big number of weights because every layer has a separated group of weights. To solve this issue, cross-layer connectivity is introduced by DenseNet. It links every single layer to all layers in the network by using feed-forward method. The feature maps of each previous layer are employed to input into all of the following layers. In DenseNet, there are l(l + 1) / 2 direct connections whereas in traditional CNN, there are l connections between the previous and current layer. DenseNet illustrates the impact of cross-layer depth wise-convolutions. Because of this, the network achieves the capacity to discriminate between the added and the preserved information, as DenseNet links the features of the preceding layers rather than adding them. DenseNet is parametrically complicated because of its limited layer structure. It is also complicated in case of increasing feature maps. The direct admission of all layers to the gradients increases the information flow all across the network through loss function. Moreover, this includes a regularizing impact, which lowers overfitting on tasks in company with small training sets. Architecture of DenseNet is given below:

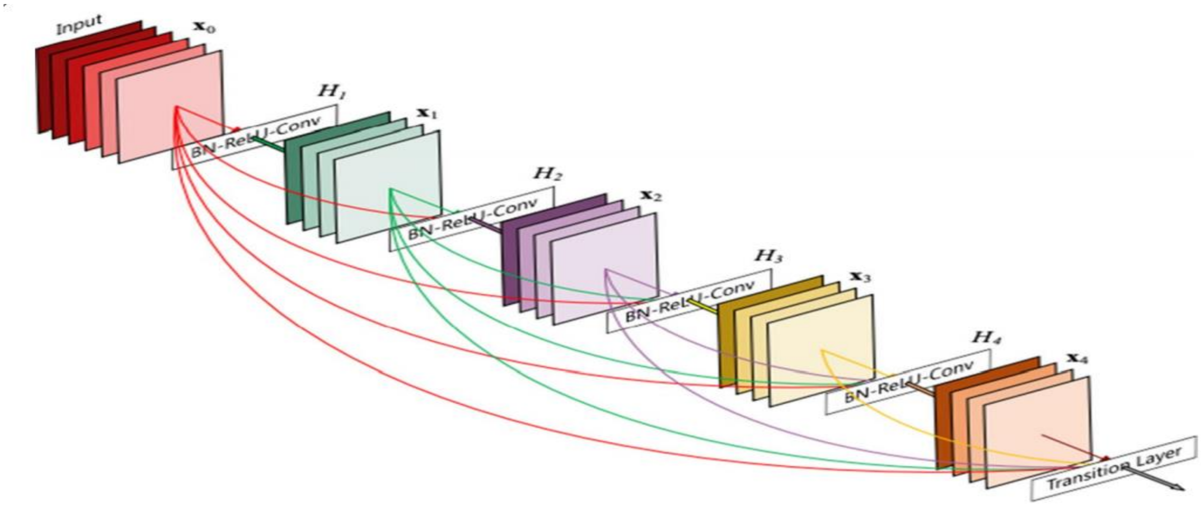

Figure 04: Architecture of DenseNet

Xception:

It is known as extreme inception architecture. It works as depth wise separable convolution. In this model, original inception block is made wider and to reduce computational complexity, exchanging is done with 3x3 dimension followed by 1x1 dimension. This architecture becomes more efficient when decoupling and spatial correspondence is used. Besides, it performs mapping of the convolved output to the embedding small dimension applying 1 × 1 convolutions. Secondly, it works on k-spatial transformations. Here, k means width-defining cardinality and it is acquired through the transformations number in Xception. Convolving each channel around the spatial axes made the computations much easier in Xception.

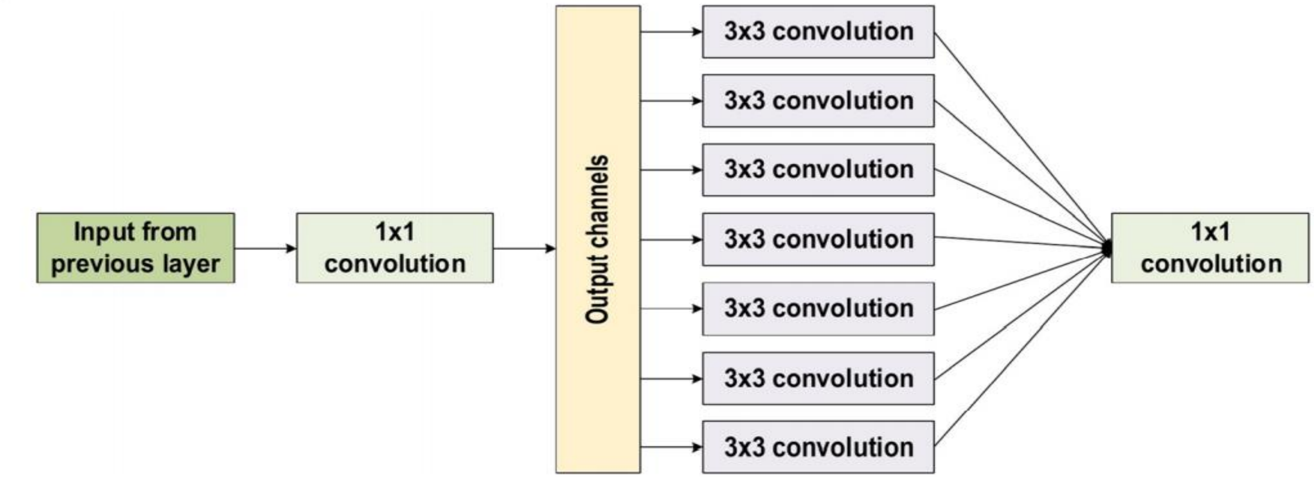

Figure 05: Xception Block Architecture.

For performing cross-channel correspondence, these axes are used afterwards as the 1 × 1 convolutions (pointwise convolution). Depth of the channel is regularized using the 1 × 1 convolution. In Xception, traditional convolutional operation utilizes a number of transformation segments correspondent to the number of channels; Inception, moreover, utilizes three transformation segments and on the other hand, traditional CNN architecture utilizes only one transformation segment. Xception model gains additional learning efficiency and good performance but it has a little problem and that is it cannot minimize the number of parameters.

CapsuleNet:

CapsuleNet can work on feature properties such as size, orientation, perspective, etc. It has the ability to effectively detect the face with several types of information. Numerous layers of capsule nodes are used to build the capsule network. Besides, encoding unit containing three layers of capsule nodes forms the CapsuleNet or CapsNet (initial version of the capsule networks).
To understand the functionality of CapsuleNet, MNIST dataset can be taken as example. CapsuleNet comprises 28 × 28 images, applying 256 filters of size 9 × 9 and with stride 1 on this dataset. The output is 28 - 9 + 1 = 20 and 256 feature maps. Now, these outputs are used as inputs for the first capsule layer during creating an 8D vector rather than scalar; this is a modified convolution layer. Stride 2 with 9 × 9 filters is employed in the first convolution layer. After that, the dimension of the output becomes (20 − 9)/2 + 1 = 6. The first capsules employ 8 × 32 filters, reason of generating 32×8×6×6 (32 for groups, 8 for neurons and 6×6 is neuron size).
In the following figure, encoding and decoding process of CapsuleNet is described;

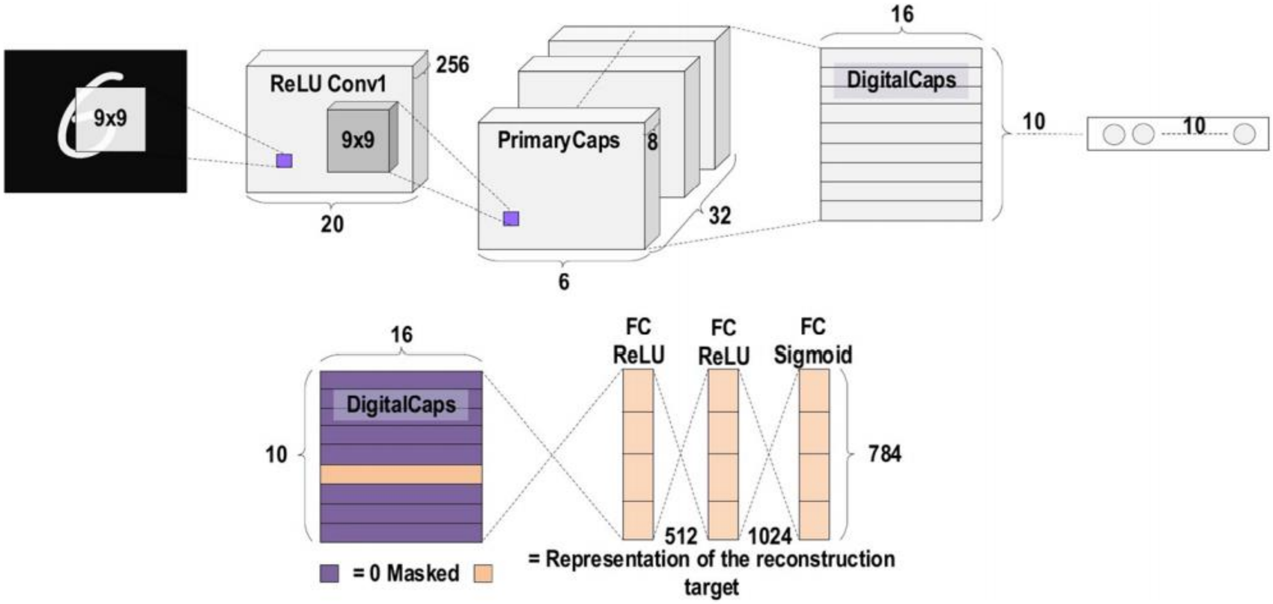

Figure 05: CapsuleNet Architecture (Encoding and Decoding Process).

In CNN, a max-pooling layer is used for handing the changes in translation. It can detect the feature moves in the event that the feature is still within the max-pooling window. Overlapped features can be found by using this method. It is the most notable in detection and segmentation operations because the capsule involves the weighted features addition from the preceding layer. A particular cost function is employed in CNN to estimate the global error that rises toward the back all over the training process. In that situation, the activation of a neuron will not grow anymore if weight between two neurons becomes zero. The signal is conducted based on the features parameters instead of single size being given with the complete cost function in monotonous dynamic routing with the agreement.

3. Discussion

Every year, a competition is held to find out new and best CNN models. Researchers from all over the world develop their own CNN models and then participate in that competition. ImageNet, CIFAR-100, MNIST and more big datasets are used to check the performance of those models. AlexNet, VGG 16 & 19, ResNet and all the models come from that competition. A description table of our selected models and their performance on that competition is given below:

Model	Depth	Dataset	Error Rate	Input Size	Year
AlexNet	8	ImageNet	16.4	227x227x3	2012
VGG	16,19	ImageNet	7.3	224x224x3	2014
ResNet	152	ImageNet	3.57	224x224x3	2016
DenseNet	201	CIFAR-10,CIFAR-100, ImageNet	3.46, 17.18, 5.54	224x224x3	2017
Xception	71	ImageNet	0.055	229x229x3	2017
CapsNet	3	MNIST	0.00855	28x28x1	2018

From the above table, it is clearly seen that Xception and CapsNet gives the better performance. Their error rate is very low (0.055 and 0.00855 respectively) which is very good and by using this two model, any kinds of image classification can be done perfectly. VGG-16 & 19, DenseNet are also good but their error rate is a bit high if compared with Xception and CapsNet. So, overall, Xception and CapsNet can be good model for image processing works in future.

4. Conclusion

The processing of images is faster and more cost-effective. One needs less time for processing, as well as less film and other photographing equipment. It is more ecological to process images. CNN is the best artificial neural network, it is used for modeling image but it is not limited to just modeling of the image but out of many of its applications. No processing or fixing chemicals are needed to take and process digital images.

References

[1] Doohee Lee[1], Jingu Lee[1], Jingyu Ko[1], Jaeyeon Yoon[1], Kanghyun Ryu[2], Yoonho Nam[3], Deep Learning in MR Image Processing, DOI: https://doi.org/10.13104/imri.2019.23.2.81, 30 June 2019.

[2] Tsung-Han Chan, Kui Jia, Shenghua Gao, Jiwen Lu, Zinan Zeng, and Yi Ma, PCANet: A Simple Deep Learning Baseline for Image Classification?, arXiv:1404.3606v2 [cs.CV] 28 Aug 2014.

[3] Manali Shaha[1], Meenakshi Pawar[2], Transfer learning for Image Classification, Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018) IEEE Conference Record # 42487; IEEE Xplore ISBN:978-1-5386-0965-1

[4] Laith Alzubaidi1,5* , Jinglan Zhang1, Amjad J. Humaidi2, Ayad Al-Dujaili3, Ye Duan4, Omran Al-Shamma5, J. Santamaría6, Mohammed A. Fadhel7, Muthana Al-Amidie4 and Laith Farhan8, Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, Alzubaidi et al. J Big Data, https://doi.org/10.1186/s40537-021-00444-8, 8:53 (2021).