American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)

ISSN (Print) 2313-4410, ISSN (Online) 2313-4402

© Global Society of Scientific Research and Researchers

http://asrjetsjournal.org/

A Comprehensive Review of Deep Learning Architectures for Computer Vision Applications

Arman Sarraf [a]*, Mohammad Azhdari[b], Saman Sarraf[c]

[a]Department of Electrical and Computer Engineering Islamic Azad University North Tehran Branch
[b]Data Processing Company (DPCO)
[c]The Institute of Electrical and Electronics Engineers, Senior Member IEEE
[a]Email: armanhs13@gmail.com

Abstract

The emergence of machine learning in the artificial intelligence field led the world of technology to make great strides. Today's advanced systems with the ability of being designed just like human brain functions has given practitioners the ability to train systems so that they could process, analyze, classify, and predict different data classes. Therefore, the machine learning field has become a hot topic for scientists and researchers to introduce the best network with the highest performance for such mentioned purposes. In this article, computer vision science, image classification implementation, and deep neural networks are presented. This article discusses how models have been designed based on the concept of the human brain. The development of a Convolutional Neural Network (CNN) and its various architectures, which have shown great efficiency and evaluation in object detection, face recognition, image classification, and localization, are also introduced. Furthermore, the utilization and application of CNNs, including voice recognition, image processing, video processing, and text recognition, are examined closely. A literature review is conducted to illustrate the significance and the details of Convolutional Neural Networks in various applications.

Keywords: Computer Vision; Machine Learning; Image Classification; Semantic Segmentation; Deep Learning; CNN.

----

* Corresponding author.

1. Introduction

Machine learning is composed of many different algorithms that can learn basic relationships and features from input data and make decisions independently without needing clear instructions or direct human intervention, i.e., unsupervised methods. The primary basis for designing such algorithms was inspired by human behavior. The task was to build a system that can "see" like humans and make decisions after processing what has been seen. Most of the algorithms introduced in the 1990s failed to demonstrate the efficiency and accuracy of recognition tasks compared to human beings. Consequently, several models like artificial neural networks have been implemented to solve image classification, decision-making, and prediction problems [1]. The computer vision field is a branch of artificial intelligence. The main focus is to create an artificial intelligible system that functions like the human brain. This field focuses more on how the computer can automatically achieve a high-level understanding from a system input in texts, photos, and videos, being utilized for various tasks such as image classification, motion detection, 3D modeling, and localization [2]. One of the most important tasks for machine learning methods is to classify operations. Most classifications are for images since most systems' input data are entered in this form. Image classification is a machine learning procedure by which images can be classified into different classes according to the features and visual characteristics. To achieve better accuracy and solve classification problems, variants of methods and algorithms have been introduced, which eventually led to the emergence and introduction of deep learning methods and artificial neural networks [3]. Deep learning is a method introduced to solve machine learning problems in classifying images and learning ability. Deep learning, also known as a deep neural network, consists of many hidden layers that can learn and extract features automatically, even from unlabeled data. It has shown high performance and convincing efficiency in various applications such as image classification, segmentation, and object detection. Various models using different architectures have been proposed to enhance deep learning performance. The most famous architecture is Convolutional Neural Network, the so-called CNN [4]. Convolutional neural networks can be considered as a class of Deep Neural Network. A convolutional neural network has several layers, each with a specific function in processing. This network has shown so much efficiency that large companies such as Google, Facebook, and AT&T hired uncountable scientists to design and introduce the best architectures that can be implemented in CNNs. In this section, components and the best architectures of CNNs are reviewed [5]. After using different applications, the research resulted in satisfactory performance for solving problems, overcoming human error.

2. Computer vision

In its most comprehensible and simplest definition, computer vision generally means science that gives computers or machines the ability to recognize. Vision for computers is considered more than just recording light. The main purpose is storing and extracting information and features from what has been seen, and components such as memory, recognition, evaluation, and estimation are the practical and functional elements of the process. The computer vision goal is enabling machines to understand the world - often called image perception - through the processing of digital signals. Such a grasp for machines is getting done by extracting important information and features from digital signals and performing complex arguments [6]. In the last two decades, it is observed that computer vision has made great improvement in recording and recognizing movements based on assessing each part of the image, such as human movements. This process, implemented with sensors entailing various features, is generally divided into four sub-processes: Initialization, Detection, Estimate, and Recognition. Each of them is divided into a large number of different processes and categories. It should be noted that changing system settings and using different data could impact performance and efficiency [7]. One of the most useful computer vision applications, which is very advanced today, is image segmentation. Image segmentation's various methods and algorithms are divided into two categories: Supervised method, Unsupervised method. Supervised method classification can be used for specific class images or image sets with different classes. To implement this method, it is necessary to conduct a mental assessment to be aware of the efficiency and the fact that the segmentation method has done its job properly. In such a way, a human has to be sure of the procedure correctness by interfering and comparing image results of segmentation with predefined categories. Therefore, the supervised method seems to be a problem, causing limitations such as classification on big data or a huge number of classes; this method is not applicable in many computer vision applications. Therefore, an unsupervised method was introduced to solve such issues, which has its drawbacks. An unsupervised method is very useful and practical for real-time segmentation. This method can also adjust required parameters for the next algorithms based on the results they received after evaluation. To implement these segmentation methods that can automatically perform the desired task with high accuracy and precision, heavy algorithms and complex mathematical functions have been designed which are changeable according to the different input forms such as Image, Video, Text [8]. Over the past two decades, the world of computer vision has taken great strides in solving localization problems by using cameras and tracking algorithms. In this way, many algorithms are used to track and identify features that, by increasing the speed of processing and calculations and upgrading the amount of memory is available today, they no longer have the problem of traditional computer vision algorithms. In other words, computer vision algorithms can locate the desired object in real-time by using advanced algorithms on a range of cameras that have microprocessors in inside, and their abilities are not limited to make connections between two points anymore [9]. Also, in recent years, due to a special goal in computer vision, which is to strengthen and enhance the process of analyzing images and extracting features from them, there has been a lot of interest and effort from researchers in the field of context modeling. The main role of context modeling is to simplify and introduce the structure, understanding how to maintain data. Today, there are various applications for those who are willing to use context modeling in their works, in which information such as specific techniques, applications, and approaches are presented and explained. Researches have also shown that recently a new technique for training context modeling has been introduced that allows users to detect objects in computer vision by defining a spatial relationship between objects and models, which reduces the repetitious searching process and increases the procedure's speed [10]. After many difficulties and challenges, computer vision has made significant advances in various applications such as face recognition, object detection, image classification, each of which is a separate set of basic principles and calculations, owing to unique structure and implementation [11].

Researches have shown that systems with the ability to analyze images and extract features such as gender, different body parts have been created combining computer vision and graphics. Today, machines can estimate population density, detect and predict the upcoming event, tracking people thanks to the significantly advanced improvement of such processes reaching modern stages [12].

3. Image Classification

Image classification is a method that contains lots of complex computer vision algorithms that can be designed to classify images according to the content, features, and variants of factors that may impact this process. The image classification aims to identify all pixels in a digital image, classifying and assigning them to several predefined classes. Colorful data are converted to grayscale or a specific color level, being detected and extracted its features. after the data is classified, it can then be used to generate features-maps in an image. Commonly, multi-spectrum data are used for classification, which is the spectral pattern in the data used for each pixel as a numerical basis for classification. The image classification process requires using an algorithm designed to serve a specific purpose from which the algorithm differs. The process is defined and implemented through five main stages: Preprocessing, Feature selection and extraction, selection of training samples, Classification processing, Accuracy assessment [13].

3.1 Preprocessing

The first stage of classification is to make some changes to the image before processing. The RGB image is first converted to a grayscale image, converted and stored as binary data. In principle, after applying the changes, the original image is converted to new pixels based on binary equations and divisions. The main reason for such a transform is to make the important features clearer in the image for further processing [14].

3.2 Feature Selection and Extraction

To perform the classification or detection process, information must be identified from image features. Also, most irrelevant and insignificant information should be discarded. Besides, calculations should be easy in order to extract features and processing be fast-paced. It should be noted that Some important information is automatically lost while capturing images by cameras before entering the process. This data loss has been one of the most troubling issues that harm feature extracting, and as a result, systems cannot recognize the real objects in images. Hence, for various reasons, mentioned above and listed below, eventually led to the introduction of Deep Neural Networks [15].

- Light changes and intensity caused some changes in pixels and the appearance of objects in the image. As a result, diagnosis and classification are disrupted.
- The size and scale of images can be near or far from the camera. This difference could create new classifications that are confusing for computers to predict the same class.
- The cluttered background consisted of colorful and complex pixels that could mislead the classification mechanism and main features recognition.
- The old classification algorithms are not advanced enough to predict that some images are taken from different angles are assigned to a class, having the same features.
- There could be various objects with the same feature and class that feature extraction in image classification could not classify them into one class.

Given the mentioned problems, computer scientists have come up with several ideas leading to image classification algorithms' improvement over time. Such approaches are:

- Color Features: The color histogram method, which is the most common method for extracting color characteristics of images, could be applied in recognizing color features and extracting them, regardless of size, zoom, and rotation.
- Texture characteristics: It is a very powerful technique, especially for large images that have duplicate areas. The texture is derived from a set of pixels with certain properties, including two categories: extraction of spectral properties and spatial properties extraction.
- The characteristics of shape: This method, which is commonly used to identify objects and describe shapes, could be applied to classification based on the extraction of environmental and border features of the image or classification based on surface features.

3.3 Selection of training samples

The algorithm preparation for the classification task uses samples collected based on the objective. Therefore, it can be concluded that utilizing a sufficient number of samples is very important and could lead to high efficiency. Managing training samples includes deleting, adding, classifying. Bearing in mind that a selected training sample could activate a new roadmap in the feature extraction step [16].

3.4 Classification processing

Training the model to classify features is done by samples in the previous step. In traditional image classification techniques, electronic devices generally caused noisy samples that could affect the pixels that make up the classification maps. To reduce noises, a process is introduced based on remote spectrometry responses and is performed with the help of a large number of filters. Noise, which means random variation in color information or lightening pixels, is usually caused by sensors or digital cameras during storing images [17].

3.5 Accuracy assessment

Each image classification task is done by two categories of datasets: training dataset and testing dataset. The training dataset is used to train the classification's algorithm, and the testing dataset is for evaluating and determining the classification accuracy. Accuracy assessment is one of the important parts of any classification. It compares the categorized image with a testing dataset that is considered as correct labeled data. This standard procedure makes a square error matrix that rows and columns represent the visualization of the algorithm's performance [18]. The image classification is usually done using several techniques and divided into supervised and unsupervised methods [19] [20]. A supervised method is a classification technique with technical supervision, commonly used to analyze remote sensing data. In a supervised method, classification is done by training an algorithm with given information such as labeled input data, roadmaps, features patterns. An unsupervised method is a technique to train an algorithm using information that is neither classified nor labeled, and the algorithm, consequently, operates on information without guidance. Unsupervised learning algorithms can perform more complex processing tasks than supervised learning systems. As mentioned, several techniques involving complex algorithms are presented for image classification purposes. The techniques are divided into three main categories [21]:

- Support Vector Machine: SVM is a supervised machine learning method that uses classification algorithms to distinguish the image class between two classes. The technique is done by defining the classes and training the model by labeled images. After the training step, the model can classify the non-labeled samples, and the efficiency of the Supper Vector Machine, so-called SVM, varies with the change of hyperplane parameters [22].
- Artificial Neural Network (ANN): Artificial neural networks or connectionism systems are computational systems that mimic functions similar to the human brain. Every artificial neural network, the so-called ANN, is composed of several layers. Each one is made up of many neurons connected to all neurons in the previous layer. This system gets trained by inputs that determine the system function. This network's efficiency can increase or decrease depending on the architecture, parameters, and inputs [23].
- Decision Tree: This algorithm, which belongs to the supervised algorithms' family, tries to classify and solve problems by downsizing the data into smaller subsets, which draws the decisions required to be made graphically. In table 1, the advantages and disadvantages of the three methods are discussed.

Table 1: Comparison of strengths and weaknesses of SVM, ANN, and DT

Methods	Strengths	Weaknesses
Support Vector Machine	• Elude from Overfitting • Provide Unique Explanation • More Efficiency	• Slow Implementation • Very Intricate Algorithm
Artificial Neural Network	• Applicable for Big Data • Practicable On Noisy Samples	• Learn Slowly • Requires Powerful Systems
Decision Tree	• Simple Understanding • Not Much Knowledge Needed	• High Loss Value • Confusing Splits

Although these accurately practical methods were provided for classification, there were still some drawbacks. At the forefront of these problems, there was this controversial question: Does the 'X,' identified object to Class 'A,' really belong to Class 'A'? This question was not answered due to problems such as interfering noisy samples, lack of ability to fail training systems in the right way completely and disrupting the inputs used for training purposes [24]. To solve such issues, scientists have developed a deep-learning neural network, which led to significant artificial intelligence advances [25].

4. Deep Learning

Deep learning is an artificial intelligence function that uses multiple layers to extract higher-level features from raw input stages. For instance, while inserting an image input into a deep neural network, the edges are identified in the first layers, and the most obvious features are extracted, then the higher layers identify deeper features [26]. Deep learning, which is currently one of the most important and practical machine learning techniques, has been very successful in many applications such as image analysis, speech recognition, and text recognition [27]. This technique teaches how to detect and classify objects by extracting features from input images through two strategies: supervised and unsupervised with a unique architectural characteristic. Deep Learning's history can be traced back to 1943 when Walter Pitz and Warren McCulloch invented a computer model based on human brain neural networks. They used a combination of algorithms and mathematics called "threshold logic" to mimic the human thought process, however for various reasons, including the lack of advanced hardware and software, this model was not considered useful until recent years. Deep learning returned to the field in 2006, becoming a hot topic for researchers ever since [28]. DL has shown a high-level efficiency in various fields such as Image Classification, Object Detection, Video Processing, Natural Language Processing, Speech Recognition. Given that each application has its algorithm and processing method, various models and algorithms of deep learning have been introduced over the past few years. The most important introduced models that have improved performance and reduced the problems of features extracting are [29]:

4.1 Stacked Autoencoder (SAE)

A stacked autoencoder is the simplest deep learning model, being a subset of the unsupervised method. SAEs are usually designed from multiple scattered layers, each composed of several auto-encoders. In this architecture, each layer input is the output of the previous layer. SAEs solve classification problems by placing several automatic encoders consisting of two main steps: Encryption, Decryption. Auto-encoders typically use backpropagation to change and reduce the size of weightlift inputs, which somehow prepares the input values for better feature extraction [30]. In figure 1, an overview of SAE architecture can be observed.

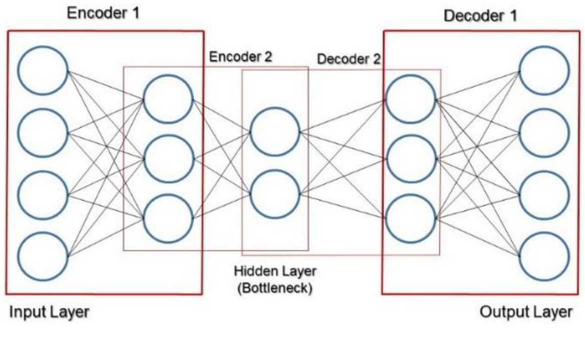

Figure 1: A simple view of stacked auto-encounter architecture

4.2 Deep Belief Network (DBN)

The first deep learning model is the deep belief network, which could get fully trained. DBN and SAE's difference structure is that DBN comprises several restricted Boltzmann machines that include two visible (V) and hidden (H) layers. The restricted Boltzmann machine uses Gibbs sampling to train its parameters. Restricted Boltzmann (RBM) uses conditional probability P (h | v) to calculate the value of each unit in the hidden layer and then the conditional probability p (h | v) to calculate the value of each unit in the visible layer. This process is repeated until the model is fully trained. Figure 2 illustrates a deep belief network architecture [31]. As can be implied, this model has a more complex architecture and calculations that slow down the process [32].

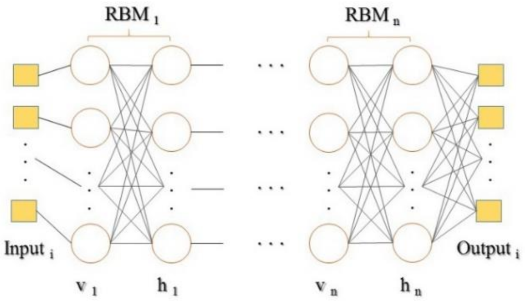

Figure 2: A deep belief network architecture

4.3 Convolutional Neural Network (CNN)

A convolutional neural network (CNN) is a special type of artificial neural network and an important subset of the supervised method. CNN uses machine learning algorithms to analyze data and extract features. CNN architecture is designed to mimic the human brain neuron pattern. Its Layers are divided into a three-dimensional structure in which each set of neurons in layers analyzes a specific area of the image [33].

4.4 Recurrent Neural Network (RNN)

This model has been introduced due to the lack of previous models' ability to extract serial data's features. The problem was they failed to learn and extract features from texture inputs. A recurrent neural network (RNN) learns the features and information from the continuous data stored in the neural network's internal state's predefined memory input. Its purpose is like predicting the continuation of a sentence or extracting its meaning and key features. As it is known, each word is related to other words in a sentence; hence, one or more previous words must be taken to account for the model to extract features. In RNN's architecture, each node is connected from one layer with a direct (one-way) connection to other nodes in the next sequential layer [34]. An RNN neural network architecture is shown in Figure 3.

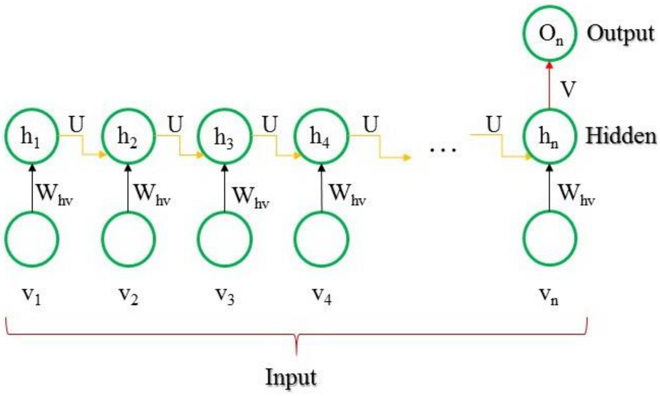

Figure 3: An architecture of a Recurrent Neural Network

So far, most machine learning and digital signal processing techniques were used by low-depth architectures. A common trait between low-depth learning models is a relatively simple architecture consisting of only one layer. It has the task of converting raw input signals or features into a specific roadmap for problem-solving, which may be uncontrollable (Such as SVM). Such models can only ultimately use one layer to separate the shallow linear pattern and extract features when used. Models with low-depth architecture effectively solve simple problems; however, the performance and synchronism are limited. This flaw prevented the system from performing well when faced with more complex tasks in the real world, including natural signals such as human speech, natural language, real-time images, and visual data. Such problems and deficiencies have attracted more complex and practical neural networks such as convolutional neural networks [35]. It should be noted that the process of deep learning implementation is a complex task in order to achieve the highest efficiency. Variants of parameters and hyperparameters can affect the process. Therefore, there is still no definite pattern to show which criteria improve network performance. It is also difficult to answer which algorithm reduces the complexity of optimization and can improve the Processing speed. However, it must be said that considering and paying attention to a combination of network size, detecting the edges and middle limits, and the resolution of images can cause significant changes while training the network and its output [36]. After introducing various algorithms in Deep Learning, networks have finally been developed that advanced, sophisticated, and accurate that they can determine the location and type of objects in the images. For example, these advanced networks can recognize and classify birds in an image and identify the breed and their species. Classification approaches based on fine-grained images can be classified into four groups in the field of deep learning [37].

- Approaches that directly use deep neural networks (mainly CNNs) to classify images with sub-features.
- Approaches use deep neural networks more as feature extractors to locate and align different parts of the fine-grained object.
- Approaches that use multiple deep neural networks to better distinguish between very small visual images.
- Approaches that use the visual attention mechanism can detect different fine-grained images that are difficult to determine.

5. Convolutional Neural Network

Since this article's main purpose is to review convolutional neural networks (CNN), it is better to begin with, its creation history. The idea of neural networks with a structure similar to the human brain was discussed even before computers' invention. In the early stages, neural networks were evaluated by propositional logic. With the emergence of artificial intelligence fields such as convolutional and backward propagation for deep neural networks, neural networks made great strides; however, CNN's improvements have come with a major hurdle. The interruption was due to a lack of systems that could perform a heavy and twisted process. Therefore, neural networks were not commercially viable when they were introduced, and the fluctuation continued until the invention of GPUs. Nowadays, CNN can be used in real-life applications on advanced systems and network architectures [38]. Table 2 briefly summarizes the progress.

Table 2: Creation and progress of convolutional neural networks

Cycle	Improvement	Year	Innovation
1940 - 1979	Beginning of NN	1943	Evaluate neural function with predictive logic
		1949	Proposal of cellular interpretation theory
		1962	Recording Cat's neurons electrical activity to achieve pattern functions
1980 - 1998	Creation of CNN	1980	Inventing a self-learning neural network that could represent basic geometrics
		1989	Utilization backpropagation CNN for the actual application
1999 - 2010	Development of CNN	1999	Proposal of Max-Pooling
		2006	Presentation Max-Pooling for CNN
2011 – 2015	Merging GPUs with CNN	2011	Training a CNN model with GPU for the first time
		2012	Proposal of Dropout technique by Google researchers
		2013	Proposal of Drop-Connect for CNN
		2014	Presentation of many more useful architectures like VGG/RCNN
		2015	Releasing different open-source libraries for CNN by Google
2016 - 2020	Introduction of advanced CNN's Architecture	2016	Improvement CNN for real-time classification by Introducing Yolo/SSD
		2017	Introduction of upgraded models for getting more performance
		2018	Pre-training language models

5.1 Review

A convolutional neural network consists of many hidden neural neurons in which neurons can recognize and classify particular features. This algorithm reduces the size of the input without affecting the characteristics of the images. Extracting features by neurons can be anything from the image edge to the speech purpose. The CNN structure comprises many convolutional cores to highlight important image features making them understandable to the system. The reason behind such a thing is to avoid wasting time processing on many unnecessary pixels presented in each image with no useful information. a Convolutional neural network comprises several neural layers that could extract more high-level features by increasing the number of layers [39]. CNN was first used in 1989 by a scientist named LeCuN to analyze network-like topological data (images and time series data), showing its good performance that attracted researchers. As a result, CNN was inspired by the same initial structure later [40]. The convolutional neural network's simplest structure usually consists of two convolutional layers: Conv layer and two sub-sampling layers that perform the fully connected layer's sampling process. The process begins by entering input into the first hidden layer, which consists of three filters. After processing, three maps of the features are determined and weighted. Then, a new feature map is obtained through a nonlinear activation function in the sub-sampling layer. Afterward, these maps are transferred to three trained filters from the Conv layer, and as a result, three new maps are obtained through the next sub-sampling layer. The second sub-sampling layer's output is transferred to a fully connected layer, converted to a vectorized coordinate, and then the output of the fully connected layer is entered into the neural network for the training step. The schematic of steps is depicted in Figure 4.

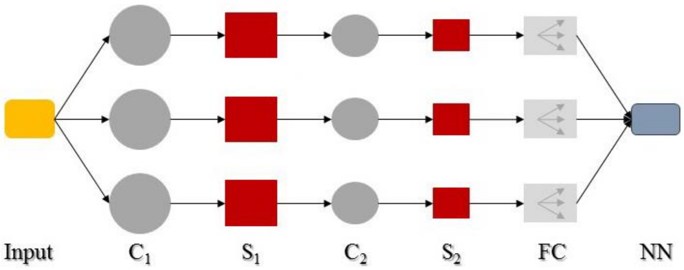

Figure 4: C refers to the convolutional layer, S refers to the sub-sampling layer, & FC is a fully connected layer

The three-dimensional structure makes the model more advantageous. CNN can also be used in large datasets because it makes processing easier by reducing and changing sample sizes. They are mainly used to identify spatial differences, scales and classify two-dimensional graphic images. Furthermore, the network can perform the learning process in parallel due to the same weight of neurons in the map of similar features. The special structure of local weight distribution gives the convolutional neural network unique speech recognition and image processing advantages. The weight distribution, which reduces the network complexity, makes extraction and classification features far easier. The advantages can be listed as following [41]:

- The designed network features maximum compatibility with the input.
- Parallelism makes the simultaneous feature extraction and pattern classification implementable in the training phase.
- Weight distribution reduces the training parameters and makes the convolutional neural network structure easier and more compatible.

5.2 Architecture

Convolutional Neural Network has undergone many changes since its introduction for improving performance, resolving various issues such as extracting feature. Today, after all the changes and improvements, CNN architecture is generally composed of three main neural layers: Convolutional layers, Pooling Layers, Fully Connected Layers doing a large volume of numerical processing calculations [42]. However, regulatory techniques such as dropout or batch normalization are sometimes added to network settings, which improve the performance of the neural network [43]. Moreover, the network training is accomplished by two stages: forward and backward steps. The forward step aims to maintain parameters such as inputs weight and value when entering into each layer. Afterward, data loss is calculated using predicting the outputs. In the backward step, based on the loss calculated in the first step, the gradient of each parameter is determined by chain rules calculations updating parameters for the next forward step. These two steps are repeated until the network is fully trained [44]. A concept of CNN architecture is shown in Figure 5.

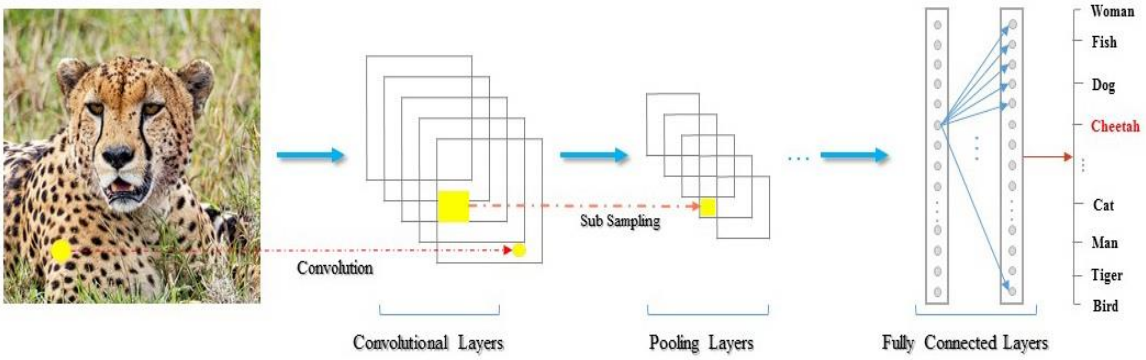

Figure 5: The concept of convolutional neural network architecture

5.3 Convolutional Layer

The convolutional layer (Conv) is the main layer of this neural network architecture that determines the amount of output associated with the input to the receiver. Conv layer consists of filters, each composed of neurons that the neurons inside each filter are considered a kernel. Convolution kernels divide input images into smaller parts, commonly known as receiver fields. The concept behind cutting and dividing the input into smaller sizes emphasizes key information making it easier to extract important features. The outputs are identified by the cores that process the length and width of the data. A 2D activation map of features is created as well. By such a thing, the network quickly learns those filters activated by observing a certain type of feature in the input. Depending on each Conv layer’s weight distribution capability, different sets of features within the image can be extracted from the input without damaging the input weight. Processing operations may be divided into different methods based on the type and size of filters. A simple diagram of the Conv layer process is shown in Figure 6.

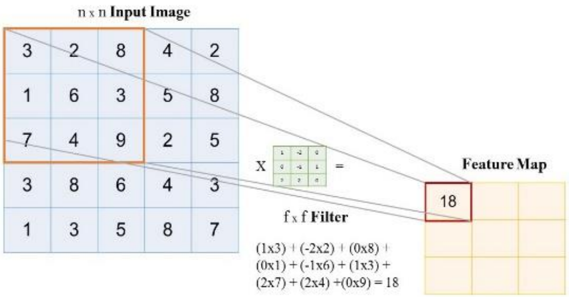

Figure 6: Building feature map matrix from the input information

5.4 Pooling Layer

The pooling layer usually comes between convolution layers. The main task of this layer is to reduce the dimensions of the feature map and the parameters. Using pooling operations helps to extract a combination of features. Reducing the feature map's size to smaller subsets and excessive connections regulates the network's complexity and increases performance. Pooling is the most widely discussed among the three layers. There are three approaches with different goals of pooling operation [45].

- Stochastic Pooling: The disadvantage of max-pooling is its sensitivity to overfitting, making it difficult to generalize the training process's data. To resolve the drawback, stochastic pooling was invented to prevent overfitting. This technique replaces a stochastic operation with max pooling, randomly selecting activation in each pooling layer area. The stochastic process is similar to max-pooling, except many duplicates of the input image undergone small changes are utilized. This random selection of activations is helpful to overcome overfitting problems [46].
- Spatial Pyramid Pooling (SPP): CNNs usually require invariant-size input images. This limitation caused problems such as reduced input recognition features differing in size and scales. To solve this problem, a method has been proposed that helps to manage inputs with various scales. In this method, the last pooling layer is replaced by a spatial pyramid pooling layer (SPP) in CNN architecture. SPPs can extract fixed-length views of images or desired areas, indicating a flexible solution for extracting different scales, sizes, and aspect ratios. SPP technique can enhance model performance in any CNN structure [47].
- Def Pooling: Solving input deformation is one of the major challenges in computer vision, especially in object recognition. Max and average pooling are useful for recognizing different angles and boundaries of images. Nevertheless, they are not able to learn different geometric shapes of images. Therefore, a new layer called deformation constrained pooling was introduced to recognize the geometric and shape. This technique recognizes various shapes by learning the geometric changes of images in visual patterns. It should be noted that this layer can be replaced by the max-pooling layer in any CNN architecture [48].

5.5 Fully Connected Layer (FC)

In convolutional neural network architecture, one or more fully connected layer is defined after the last pooling layer. FC layer converts two-dimensional feature maps created from previous layers into a one-dimensional vector for displaying more features. Fully connected layers, composed of nearly 90% of a CNN model, act as a traditional neural network. It gives a predefined length vector of extracted features from its output, which can be used for classification purposes or considered a feature vector for re-processing [49]. Neurons do processing in each filter that is fully connected to all neurons of the previous layer. The FC layer's disadvantage is its large number of parameters that make it hard to be trained. Therefore, one of the hot topics for scientists is how to reduce the number of connections and the number of filters while maintaining accuracy [50].

5.6 Activation Function

Activation functions are a set of mathematical equations that determine each layer's output in the neural network. These functions are connected to each neuron in the network. This technique determines whether the functions should be activated or not [51] due to each neuron's task. Activating functions means deciding to stop processing and sending output to the next layer. Various activation functions such as Sigmoid, Tanh, Maxout, Swish, ReLU and its types such as leaky ReLU, ELU, PReLU have introduced different mathematical equations to be used in different neural networks [52]. Choosing the right activation function can speed up the learning process because they understand nonlinear features. One of the recently considered activation functions is MISH, which has outperformed ReLU in deep neural networks [53].

5.7 Batch Normalization

Batch normalization or batch norm is a technique to train very deep neural networks. It is used to speed up the procedure, improve the performance and stability of artificial neural networks. Batch norm standardizes inputs into small layers according to their categories, making the training step more stable. It significantly reduces the number of cycles required for training a deep network [54].

5.8 Dropout

Dropout is a regularization technique introduced by Google Research that reduces excessive connections between neurons in the network by preventing complex coordination in the training data. This technique is a very impressive way to improve efficiency in neural networks, especially in CNNs. In neural networks, the various connections that learn a nonlinear relationship are closely fitted, causing overfitting. Dropout creates several thin network architectures by reducing some connections. A comprehensive network with small weights is selected at the end. The selected network architecture is considered for almost all other networks in the model, continuing the patch [55].

6. CNN Models

As mentioned earlier, Convolutional Neural Networks (CNNs) are the most popular models of neural networks that have made great strides in many areas, including classification, face recognition, and have been able to solve variants of problems. On the other hand, each CNN network consists of many parameters and hyperparameters, including weight, number of layers, processing unit (neuron), filter size, stride, activation function, and learning rate [56]. Since the convolution layer performs processing on the input pixels, different input levels can be examined by changing the size of the Conv filters. Therefore, various standard architectures have been introduced and tested for these purposes [57].

6.1 LeNet

LeNet architecture was introduced in 1998. Due to its historical importance, it is known as the first CNN model. It has also made great function in MNIST handwritten digital ID patterns. LeNet model, which is usually composed of 5 layers, accepts grayscale images with 32 x 32 x 1 as input. The inputs are transferred to the Conv layer and then to sub-sampling. Afterward, there are other Conv layers, followed by a pooling layer, and at the end of the architecture, FC layers including the output at the last layer are defined [58]. This model was the first CNN architecture reducing the number of parameters and capable of automatically learning features from raw pixels. This model was introduced to identify digital manuscript patterns and postal codes in post offices [59]. A LeNet architecture is shown in figure 7.

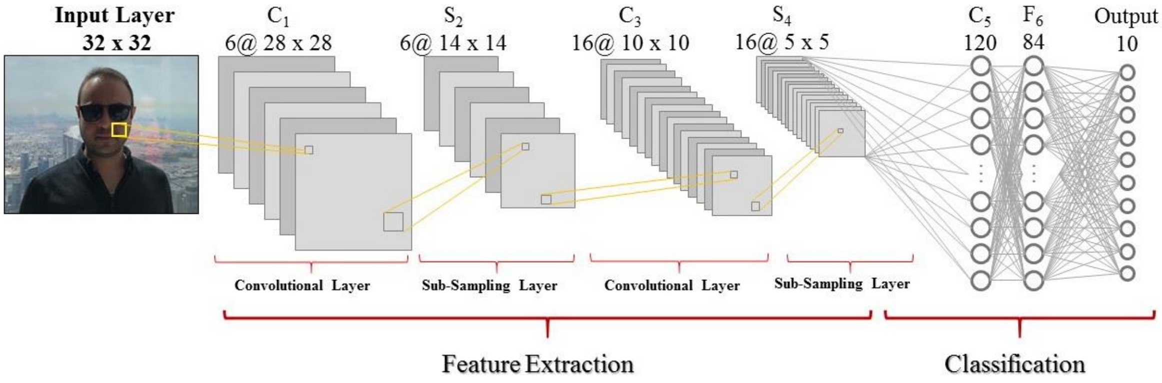

Figure 7: LeNet Model, which is one of the simplest architectures composed of 5 layers considering Conv and pooling layers and then a fully connected layer

6.2 AlexNet

Although CNN’s history began with LeNet, it was limited to recognizing digital manuscript patterns and did not perform well in classifying image classes. AlexNet is considered the first deep CNN architecture used for classification and recognition tasks. AlexNet, which has a deeper architecture than LeNet, consists of five Conv layers, one max-pooling layer, a ReLU activation as a nonlinear function, three FC layers, and above all, it uses the dropout technique. The training process was such heavily complex that the computers could not train the AlexNet model, causing many limitations for using it. Finally, in 2000, this model’s training process was implemented parallel on two Nvidia GTX 580 GPUs. The procedure took nearly six days, considering the limitation of GPUs back then. This model was not utilized until 2012 due to the lack of high-speed hard drives to perform the process [60]. Figure 8 depicts the architecture of AlexNet in which the difference between this model and the LeNet can be seen.

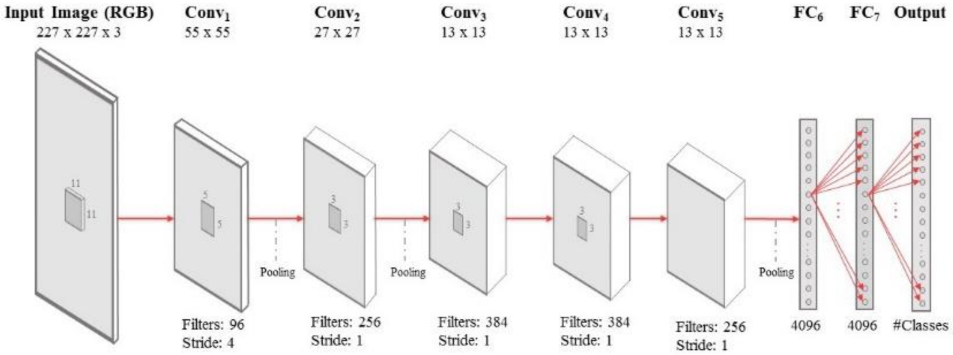

Figure 8: AlexNet architecture composed of 5 Conv layer following by pooling operations and three fully-connected layer

6.3 ZFNet

Until 2013, the training CNN network mechanism was trial and error. The performance of CNN on complex images was difficult and limited due to the lack of model improvement. In the ZFNet model, a modified version of AlexNet, the filters' size changed from 11 x 11 to 7 x 7, and the number of strides was reduced. Also, the number of filters increased in higher layers as well. The idea was that a smaller filter could hold a significant amount of input pixel size. Training first ZFNet, which was done by a 580 GTX GPU in 2013, took up to 12 days. The implementation of this model was to create a visualization technique so-called DE-convolutional network [61]. An architecture of a ZFNet is depicted in figure 9.

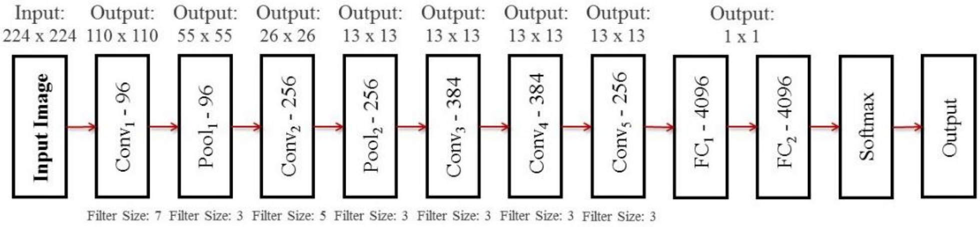

Figure 9: ZFNet consists of 5 Conv layer, two pooling layer, and three fully-connected layer

6.4 GoogleNet

GoogleNet, also known as Inception-V1, was introduced in 2014 and could win the 2014-ILSVRC competition. The architecture's main purpose was to achieve high accuracy by reducing computational costs. In GoogleNet architecture, the first layer is designed with an inception block, in which filters with sizes of 1 x 1, 3 x 3, and 5 x 5 are placed. The reason is to process different input sizes to improve classification tasks for the same class images containing different resolutions. GoogleNet consists of 22 layers that use a 1 x 1 concatenation filter to adjust input calculations before using the next layer's convolution kernels for processing. A pooling layer has also been implemented instead of FC layers to reduce the density of connections, decreasing the number of parameters from 138 million to 4 million. Like others, this model completed its training step with heavy GPUs within a week [62]. Figure 10 diagrams a part of GoogleNet architecture.

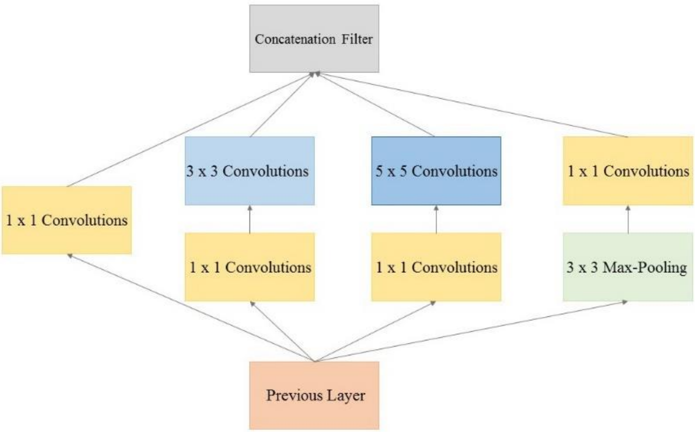

Figure 10: The output of the previous layer passes to a concatenation filter for transferring to the next layer after kernels processing

6.5 VGGNet

CNN's impressive image classification performance has guided researchers to look for more practical architectures for this neural network. That is why Oxford University researchers in 2014 introduced a model called Visual Geometry Group (VGG). VGG architecture consists of thirteen Conv layers, in which everyone is followed by a max-pooling layer and three FC layers. Filters with 11 x 11 and 5 x 5 have been replaced with ones of 3 x 3. Developers believed that using filters simultaneously with smaller sizes could do the same task as filters with 7 x 7 and 5 x 5. Computational complexity could also be reduced by using small-size filters. Implementing the VGGNet model on the image database, which had 14 million samples, included 1000 different classes, experimentally recorded its success with a return of 92.7% accuracy [63]. Figure 11 contains VGGNet architecture.

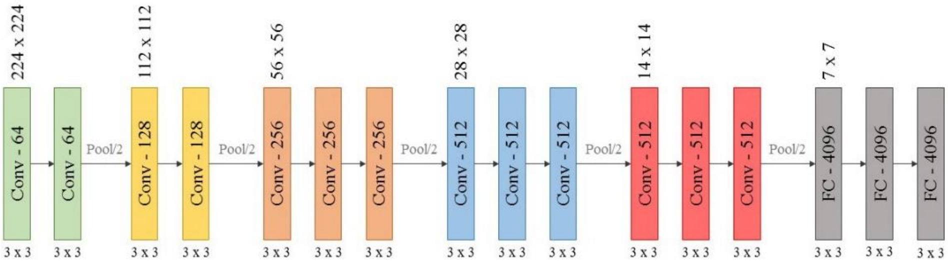

Figure 11: In VGGNet architecture, layers with a filter size of just 3x3 have been used, which caused its simplicity and high accuracy

6.6 ResNet

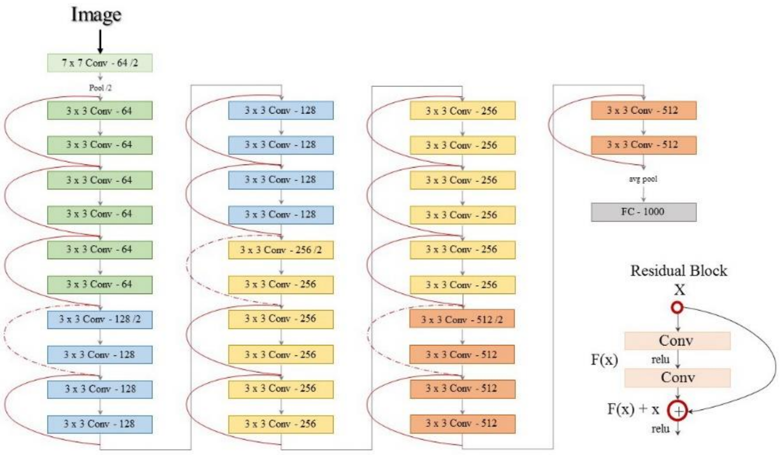

Figure 12: Residual neural networks perform by utilizing skip connections to jump over layers

Residual Neural Network (ResNet), introduced in 2015, changed the generation of CNNs. The theory behind designing such an architecture is that a higher layer should learn a new feature from the previous layer. Therefore, connections have been added to layers that can copy the next layer's input without considering feature extraction and identity from the previous layer. It has less computational complexity than other networks, although it is 20 and 10 times deeper than AlexNet and VGG, respectively. With 152 layers, ResNet showed an error of 3.57% after training and implementing on the ImageNet dataset, which was even less than a human error [64]. The ResNet architecture can be observed in figure 12.

6.7 Inception v3

Inception-v3, introduced in 2015, is a modified version of the GoogleNet. In inception-v3, the number of parameters is reduced to 24 million. This model also uses multi-level extracting in which convolutional layer factors were replaced by n x 1 and 1 x n asymmetries instead of n x n, making it performs much faster. Each filter with a size of 5 x 5 changed to two 3 x 3 filters in the architecture, leading to a significant reduction in the number of neurons and speeding up the process [65]. In figure 13, changes compared to GoogleNet are shown.

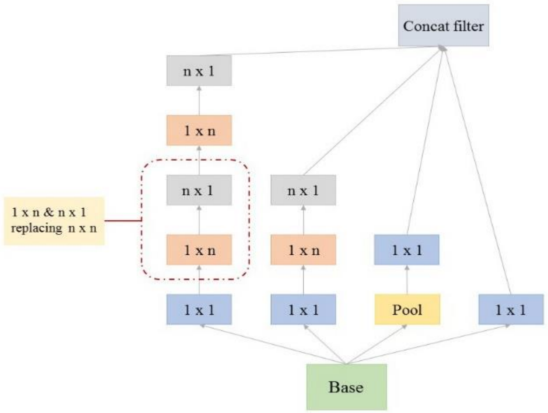

Figure 13: Inception-v3 architecture shows n x n factors replaced by n x 1 and 1 x n, causing better performance

6.8 ResNeXt

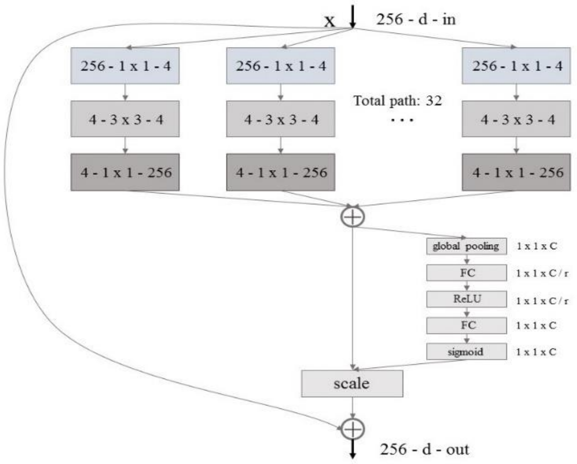

Figure 14: ResNeXt architecture is like a highly modularized network

ResNeXt, introduced in 2017, is a combination of ResNet and VGGNet models. This model won the ILSVRC 2017 reward, basically uses a simple modular architecture to classify images. ResNeXt is built by repeating a constructive block that collects a set of transformations with the same topology. In this architecture, which is an extension of the deep residual network model, a ‘split-transform-merge’ strategy is replaced instead of the standard residual block [66]. In Figure 14, the implementation and architecture of ResNeXt are depicted.

6.9 SENet

Squeeze and excitation Network (SENet), introduced in 2017, on which experimental researches have shown that SE blocks can be used on all CNN networks before the Conv layer. Placing SE block before the Conv layer allows the network to adjust each feature map’s weight adaptively. As the model name implies, this block’s operation consists of two operations; squeeze and excitation. SENet focuses mainly on relationships between channels. This new block is proposed to select feature maps (commonly known as channels) related to distinct objects. It identifies important feature maps by stopping mapping insignificant features and increasing the classes’ weight [67]. In figure 15, a SE block is more understandable.

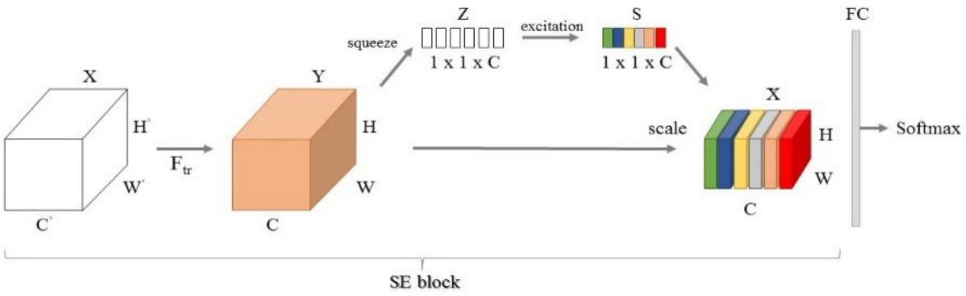

Figure 15: A SE block architecture can be trained by variants of nonlinear interactions such as spatial normalization

6.10 MobileNet

MobileNet is a small, low-consumption, low-delay model introduced in 2017 by google researchers, which can solve resource constraints problems. In MobileNet, a deep Conv layer was replaced by the normal Conv layer. Deep Conv layers process individually on each color channel instead of processing on all the colors. MobileNet is an architecture composed of 28 layers, including deep Conv layers, 1 x 1 point Conv layers, batch norm, ReLU, average collecting layer, and softmax, which is more suitable for mobile-based vision programs in the absence of computing power. This model can be used to classify, detect, embed, and segmentation similar to other popular large-scale models such as Inception [68]. A simple form of a MobileNet architecture is shown in figure 16.

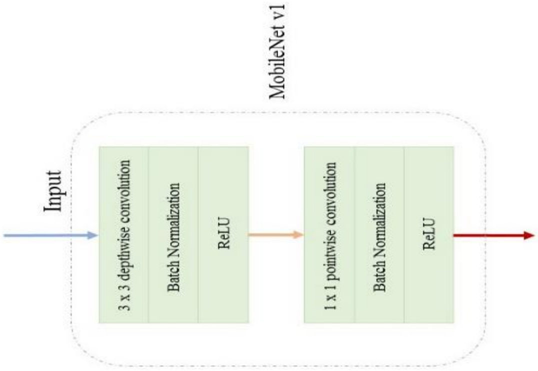

Figure 16: Depth wise convolution layers have been replaced by convolution layers, which speeded up computations

6.11 DenseNet

DenseNet is one of the newest and most powerful neural networks for recognizing visual objects introduced in 2017. This model is quite similar to ResNet but has fundamental differences that could solve the ResNet problems. ResNet problem was keeping each layer's information inside the layers without transferring the output to the next layer, making the next layers not having enough information and data to be trained. DenseNet transfers layers' previous information to the next layer. Therefore, the features are transferred from all previous layers to all subsequent layers [69]. It could reduce the gradient calculations problem, minimize the number of parameters, and reuse features. Figure 17 shows how the information is transferred from the previous layers to each next layer.

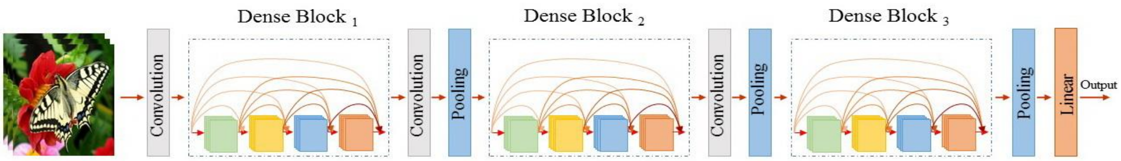

Figure 17: In a dense block, each next layer receives all previous feature maps and is obtusely connected to each previous layer

6.12 Xception

Xception can be considered as extra inception, introduced in 2017. It is based entirely on deep recognizable Conv processing, followed by a point Conv layer and generally composed of 36 Conv layers that form the network feature extraction basis. In the Xception, the first inception block has been expanded, and instead of the various dimensions of 1 x 1, 3 x 3, 5 x 5, a 3 x 3 dimension and a 1 x 1 Conv layer have been used to adjust the computational complexity. In short, Xception architecture is a linear stack of deep detachable Conv segments with residual connections. The theory was that spatial correlations and cross-channel correlations could be sufficiently separated. This idea, which was more effective than Inception-V3, ResNet-50, ResNet-101, ResNet-152, VGGNet., showed the network has become more efficient in computing [70]. Figure 18 shows the Xception architecture.

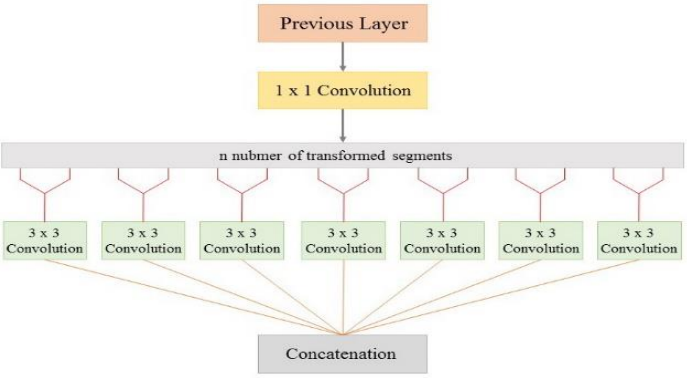

Figure 18: The architecture of the Xception model

7. CNN’s Applications

Convolutional neural networks have been implemented in many different machine learning parts with acceptable performance, namely object detection, identification, classification, regression, and recognition. Bearing in mind CNN has achieved such successful performance with a large number of labeled data, and generally, the training step, which is the most important part, needs variants of sufficient labeled data [71]. Fortunately, there are many various datasets with sufficient labels, including recognizing traffic signs, shared medical images and texture datasets, large scales of human face’s pictures, and so on, to train neural networks. Some of the important and interesting applications are mentioned below.

7.1 CNN Application Based-on Computer Vision

Computer vision’s main purpose is to create artificial systems that can process data such as videos, images, and text to extract the required information and features from inputs. Face recognition is one of the most complex tasks in computer vision requiring heavy processing. The procedure always faces two major challenges. Firstly, it is difficult to extract features from images with crowded backgrounds for face recognition, and secondly, as the search space expands, it also gets more difficult to accurately identify the location and determine the size of the face. Face recognition systems should detect and recognize changes such as lighting, changes in posture, angles, and different facial expressions [72]. Convolution neural networks, considered a class of deep neural networks, have high capabilities for detecting image patterns widely used in computer vision algorithms and could solve many problems related to mentioned tasks [73].

7.2 CNN Application Based-on Image Classification

With the emergence of CNNs and performing in image classification, the world has undergone a great change. The CNN models represent a breakthrough in image recognition and similar tasks [74]. They are commonly used to analyze visual images and often work behind-the-scenes for image classification purposes. One of the largest and most important uses of CNNs is in medical science. Much research has been conducted on diagnosing Alzheimer's disease by classifying brain images [75]. Researches have shown by training CNN models with brain scans, and brain patterns, Alzheimer's disease or cancer could be diagnosed with a low error rate [76].

7.3 CNN Application Based-on Detection and Segmentation

The main goal in detecting objects is to identify various objects that have already been introduced to the system. The main difference between image classification and object detection is that the image classification task is to classify an image to a specific class; however, object detection aims to determine the exact location or count the number of samples of an object in the image. The most important and challenging issue is calculating and determining an object's dimensions and size in the image. Fortunately, by introducing a CNN model, the so-called fast R-CNN, this problem has been solved [77].

7.4 CNN Application Based-on Video Processing

Image processing techniques usually standardize image features as a two-dimensional signal, analyzing them using signal processing techniques. However, video processing is a special case of signal processing where the input and output signals are video files or streams [78]. In video processing, a series of frames must be consecutively inserted into neural networks to identify ongoing features. That is why a Long Short Term Memory (LSTM) model was introduced, which can perform well for video processing in CNN networks [79]. In general, this approach comprehends by two parts. In the first one, features are extracted from the sixth frame of the videos, and in the second part, ordinal information between features of the frame is elicited using the bi-directional LSTM neural network [80].

7.5 CNN Application Based-on Natural Language Processing (NLP)

Natural Language Processing is a procedure in which languages and words are transformed into new forms so that computers could be understood and extract their important features. The problem with traditional models was that the systems could not understand the core and purpose of a text, failing to extract conceptual information correctly. Today, RNN models easily achieve this goal. Besides, CNNs have shown good results performing in this regard as well. A variant of results has shown CNN can understand and analyze emails, texts, books, and tweets [81].

7.6 CNN Application Based-on Speech Recognition

As mentioned earlier, deep CNNs are the best way to analyze images. In addition to this success, CNNs have shown excellent results in speech recognition. The main purpose is to extract required features from speech sounds transformed into signals and considered input [82]. Therefore, the task can be translating speech, turning sounds into writing, extracting the general and main meaning of sounds, classifying and distinguishing different sounds from each other. Interestingly, CNNs are now so advanced in speech recognition that networks can even detect emotions of speeches from signals [83].

7.7 CNN Application Based-on 1D-Data

Not only have CNNs performed well in clear images, but they have also been very successful in processing 1D-data. The success was since CNN has a high ability to be trained and extract data features, unlike all the old ML techniques. 1D-CNN is very effective for extracting fixed-length input features from a set of general data. 1D-CNN is usually used where the exact location of a feature in images is not important. Experiences have shown this model, which receives inputs in the form of 1D-signal, is mostly used in real-time processing where processing speed is an important criterion [84].

7.8 CNN Application Based-on Low-Resolution Images

One of the most important and efficient uses of CNN is classifying low-quality images. CNN also showed its ability and strength when a high-resolution dataset of images is available for training purposes, but the testing dataset is full of low-resolution images. CNN implements the process by converting small grains into larger images. For instance, identifying objects in satellite images that are often very small, detecting license plates with low resolution from a distance, or even identifying bird species [85].

8. How CNN Could Solve Feature Extracting Problems

Based on what is discussed in this piece, it is found that image classification had many difficulties before the introduction and development of today's advanced CNN. Problems were nothing more than accurately extracting key features [86]. Image classification algorithms could not extract small and fine-grained features that determine the input class at that time. Another issue with classifying algorithms was about the samples belonging to several classes [87]. The algorithms could not correctly identify the data assigned to more than one class in multi-class classifications [88]. With the emergence of CNNs and their ability to be trained deeply, such problems have been easily solved [89].

9. Conclusion

The unsupervised machine learning procedure is based on human behavior. The machine can learn basic relationships and features from input data to make decisions independently without needing clear instructions or direct human intervention. Primary ML algorithms have been replaced with several models, e.g., an artificial neural network to conduct specified tasks like image classification, decision-making, and prediction. This evolution arose from the failure of old models regarding efficiency and accuracy for recognition tasks compared to human beings. Computer vision's main focus is to create an artificial intelligible system that functions like the human brain. It focuses more on how the computer can automatically achieve a high-level understanding without any human interventions. On the other hand, classification, which is the most pivotal task in machine learning, is mostly used for images. To achieve higher accuracy in classification, deep learning methods and artificial neural networks have been introduced. They have shown high performance and convincing efficiency in various applications such as image classification, segmentation, and object detection. The most famous and useful model in the artificial neural network is Convolutional Neural Network. CNN is a Deep Neural Network class that consists of an architecture containing several different layers, each with a specific function in processing. The CNN models have been widely utilized in several approaches, such as diagnosing brain diseases, including Alzheimer's. In the future, the main focus might be on enhancing algorithms for diagnosing details of such diseases and even predicting how they occur in human brains.

References

[1]. W.-L. Chao, "Machine learning tutorial," Digit. Image Signal Process., 2011.

[2]. E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 4, pp. 607–626, 2008.

[3]. M. E. Cintra, M. C. Monard, H. A. Camargo, and T. P. Martin, "A comparative study on classic machine learning and fuzzy approaches for classification problems," 2005.

[4]. C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," arXiv Prepr. arXiv1611.03530, 2016.

[5]. F. Altenberger and C. Lenz, "A non-technical survey on deep convolutional neural network architectures," arXiv Prepr. arXiv1803.02129, 2018.

[6]. J. A. E. Anderson et al., "Task-linked diurnal brain network reorganization in older adults: A graph theoretical approach," J. Cogn. Neurosci., vol. 29, no. 3, pp. 560–572, 2017.

[7]. T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," Comput. Vis. image Underst., vol. 81, no. 3, pp. 231–268, 2001.

[8]. H. Zhang, J. E. Fritts, and S. A. Goldman, "Image segmentation evaluation: A survey of unsupervised methods," Comput. Vis. image Underst., vol. 110, no. 2, pp. 260–280, 2008.

[9]. R. J. Radke, "A survey of distributed computer vision algorithms," in Handbook of Ambient Intelligence and Smart Environments, Springer, 2010, pp. 35–55.

[10]. O. Marques, E. Barenholtz, and V. Charvillat, "Context modeling in computer vision: techniques, implications, and applications," Multimed. Tools Appl., vol. 51, no. 1, pp. 303–339, 2011.

[11]. X. Yang, S. Sarraf, and N. Zhang, "Deep learning-based framework for Autism functional MRI image classification," J. Ark. Acad. Sci., vol. 72, no. 1, pp. 47–52, 2018.

[12]. J. C. S. J. Junior, S. R. Musse, and C. R. Jung, "Crowd analysis using computer vision techniques," IEEE Signal Process. Mag., vol. 27, no. 5, pp. 66–77, 2010.

[13]. D. Lu and Q. Weng, "A survey of image classification methods and techniques for improving classification performance," Int. J. Remote Sens., vol. 28, no. 5, pp. 823–870, 2007.

[14]. S. Sarraf, "Binary Image Segmentation Using Classification Methods: Support Vector Machines,

Artificial Neural Networks and K th Nearest Neighbours," Int. J. Comput., vol. 24, no. 1, pp. 56–79, 2017.

[15]. P. Babaniamansour, M. Ebrahimian-Hosseinabadi, and A. Zargar-Kharazi, "Designing an optimized novel femoral stem," J. Med. Signals Sens., vol. 7, no. 3, p. 170, 2017.

[16]. S. Sarraf, "Hair color classification in face recognition using machine learning algorithms," Am. Sci. Res. J. Eng. Technol. Sci., vol. 26, no. 3, pp. 317–334, 2016.

[17]. A. Sarraf, "Binary Image Classification Through an Optimal Topology for Convolutional Neural Networks," Am. Sci. Res. J. Eng. Technol. Sci., vol. 68, no. 1, pp. 181–192, 2020.

[18]. S. Sarraf, "French Word Recognition Through a Quick Survey on Recurrent Neural Networks Using Long-Short Term Memory RNN-LSTM," Am. Sci. Res. J. Eng. Technol. Sci., vol. 39, no. 1, pp. 250–267, 2018.

[19]. D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari, "A survey of active learning algorithms for supervised remote sensing image classification," IEEE J. Sel. Top. Signal Process., vol. 5, no. 3, pp. 606–617, 2011.

[20]. P. Babaniamansour, M. Mohammadi, S. Babaniamansour, and E. Aliniagerdroudbari, "The relation between atherosclerosis plaque composition and plaque rupture," J. Med. Signals Sens., vol. 10, no. 4, pp. 267–273, 2020.

[21]. C. Dhaware and K. H. Wanjale, "Survey on image classification methods in image processing," Int. J. Comput. Sci. Trends Technol., vol. 4, no. 3, pp. 246–248, 2016.

[22]. S. Sarraf and J. Sun, "Advances in functional brain imaging: a comprehensive survey for engineers and physical scientists," Int. J. Adv. Res., vol. 4, no. 8, pp. 640–660, 2016.

[23]. S. Sarraf, "EEG-based movement imagery classification using machine learning techniques and Welch's power spectral density estimation," Am. Sci. Res. J. Eng. Technol. Sci., vol. 33, no. 1, pp. 124–145, 2017.

[24]. C. Saverino, Z. Fatima, S. Sarraf, A. Oder, S. C. Strother, and C. L. Grady, "The associative memory deficit in aging is related to reduced selectivity of brain activity during encoding," J. Cogn. Neurosci., vol. 28, no. 9, pp. 1331–1344, 2016.

[25]. A. Sarraf, A. E. Jalali, and J. Ghaffari, "Recent Applications of Deep Learning Algorithms in Medical Image Analysis," Am. Sci. Res. J. Eng. Technol. Sci., vol. 72, no. 1, pp. 58–66, 2020.

[26]. Y. Li, H. Zhang, X. Xue, Y. Jiang, and Q. Shen, "Deep learning for remote sensing image classification: A survey," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 8, no. 6, p. e1264, 2018.

[27]. S. Sarraf, C. Saverino, and A. M. Golestani, "A robust and adaptive decision-making algorithm for detecting brain networks using functional mri within the spatial and frequency domain," in 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Apr. 2016, pp. 53–56, doi: 10.1109/BHI.2016.7455833.

[28]. Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "A survey on deep learning for big data," Inf. Fusion, vol. 42, pp. 146–157, 2018.

[29]. L. Deng, "Three classes of deep learning architectures and their applications: a tutorial survey," APSIPA Trans. signal Inf. Process., 2012.

[30]. J. Gehring, Y. Miao, F. Metze, and A. Waibel, “Extracting deep bottleneck features using stacked auto-encoders,” in 2013 IEEE international conference on acoustics, speech and signal processing, 2013, pp. 3377–3381.

[31]. G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” Neural Comput., vol. 18, no. 7, pp. 1527–1554, 2006.

[32]. S. Sarraf and J. Sun, “Functional brain imaging: A comprehensive survey,” arXiv Prepr. arXiv1602.02225, 2016.

[33]. S. Sarraf, “Current Stage of Autonomous Driving Through A Quick Survey for Novice,” Am. Sci. Res. J. Eng. Technol. Sci., vol. 73, no. 1, pp. 1–7, 2020.

[34]. X. Chen, X. Liu, Y. Wang, M. J. F. Gales, and P. C. Woodland, “Efficient training and evaluation of recurrent neural network language models for automatic speech recognition,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 24, no. 11, pp. 2146–2157, 2016.

[35]. B. Zhao, J. Feng, X. Wu, and S. Yan, “A survey on deep learning-based fine-grained object classification and semantic segmentation,” Int. J. Autom. Comput., vol. 14, no. 2, pp. 119–135, 2017.

[36]. B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro, “Exploring generalization in deep learning,” arXiv Prepr. arXiv1706.08947, 2017.

[37]. K. Krishnan, T. Schwering, and S. Sarraf, “Cognitive dynamic systems: A technical review of cognitive radar,” arXiv Prepr. arXiv1605.08150, 2016.

[38]. D. T. Mane and U. V Kulkarni, “A survey on supervised convolutional neural network and its major applications,” in Deep Learning and Neural Networks: Concepts, Methodologies, Tools, and Applications, IGI Global, 2020, pp. 1058–1071.

[39]. C. Grady, S. Sarraf, C. Saverino, and K. Campbell, “Age differences in the functional interactions among the default, frontoparietal control, and dorsal attention networks,” Neurobiol. Aging, vol. 41, pp. 159–172, 2016.

[40]. A. A. M. Al-Saffar, H. Tao, and M. A. Talab, “Review of deep convolution neural network in image classification,” in 2017 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET), 2017, pp. 26–31.

[41]. Z. Qin, F. Yu, C. Liu, and X. Chen, “How convolutional neural network see the world-A survey of convolutional neural network visualization methods,” arXiv Prepr. arXiv1804.11191, 2018.

[42]. S. Sarraf and G. Tofighi, “Classification of alzheimer’s disease using fmri data and deep learning convolutional neural networks,” arXiv Prepr. arXiv1603.08631, 2016.

[43]. N. Aloysius and M. Geetha, “A review on deep convolutional neural networks,” in 2017 International Conference on Communication and Signal Processing (ICCSP), 2017, pp. 588–592.

[44]. M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, “Xnor-net: Imagenet classification using binary convolutional neural networks,” in European conference on computer vision, 2016, pp. 525–542.

[45]. C.-Y. Lee, P. W. Gallagher, and Z. Tu, “Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree,” in Artificial intelligence and statistics, 2016, pp. 464–472.

[46]. M. D. Zeiler and R. Fergus, “Stochastic pooling for regularization of deep convolutional neural networks,” arXiv Prepr. arXiv1301.3557, 2013.

[47]. K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for

visual recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 9, pp. 1904–1916, 2015.

[48]. W. Ouyang et al., "Deepid-net: multi-stage and deformable deep convolutional neural networks for object detection," arXiv Prepr. arXiv1409.3505, 2014.

[49]. S. Sarraf and M. Ostadhashem, "Big data application in functional magnetic resonance imaging using apache spark," in 2016 Future Technologies Conference (FTC), 2016, pp. 281–284.

[50]. R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.

[51]. N. Ershadinia, N. Mortazavinia, S. Babaniamansour, M. Najafi-Nesheli, P. Babaniamansour, and E. Aliniagerdroudbari, "The prevalence of autoimmune diseases in patients with multiple sclerosis: A cross-sectional study in Qom, Iran, in 2018," Curr. J. Neurol., vol. 19, no. 3, pp. 98–102, 2020.

[52]. S. Babaniamansour, M. Hematyar, P. Babaniamansour, A. Babaniamansour, and E. Aliniagerdroudbari, "The Prevalence of Vitamin D Deficiency Among One to Six Year Old Children of Tehran, Iran," J. Kermanshah Univ. Med. Sci., vol. 23, no. 4, 2019.

[53]. D. Misra, "Mish: A self regularized non-monotonic neural activation function," arXiv Prepr. arXiv1908.08681, vol. 4, 2019.

[54]. S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in International conference on machine learning, 2015, pp. 448–456.

[55]. G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," arXiv Prepr. arXiv1207.0580, 2012.

[56]. A. Dhillon and G. K. Verma, "Convolutional neural network: a review of models, methodologies and applications to object detection," Prog. Artif. Intell., vol. 9, no. 2, pp. 85–112, 2020.

[57]. A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," arXiv Prepr. arXiv1704.06857, 2017.

[58]. Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," Neurocomputing, vol. 187, pp. 27–48, 2016.

[59]. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[60]. M. Z. Alom et al., "The history began from alexnet: A comprehensive survey on deep learning approaches," arXiv Prepr. arXiv1803.01164, 2018.

[61]. M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in European conference on computer vision, 2014, pp. 818–833.

[62]. C. Szegedy et al., "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[63]. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv Prepr. arXiv1409.1556, 2014.

[64]. C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in Proceedings of the AAAI Conference on Artificial Intelligence, 2017, vol. 31, no. 1.

[65]. X. Xia, C. Xu, and B. Nan, “Inception-v3 for flower classification,” in 2017 2nd International Conference on Image, Vision and Computing (ICIVC), 2017, pp. 783–787.

[66]. S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492–1500.

[67]. J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[68]. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.

[69]. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[70]. F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.

[71]. S. C. Strother, S. Sarraf, and C. Grady, “A hierarchy of cognitive brain networks revealed by multivariate performance metrics,” in 2014 48th Asilomar Conference on Signals, Systems and Computers, 2014, pp. 603–607.

[72]. H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, “A convolutional neural network cascade for face detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5325–5334.

[73]. S. Yang, P. Luo, C.-C. Loy, and X. Tang, “From facial parts responses to face detection: A deep learning approach,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 3676–3684.

[74]. D. CireAan, U. Meier, J. Masci, and J. Schmidhuber, “Multi-column deep neural network for traffic sign classification,” Neural networks, vol. 32, pp. 333–338, 2012.

[75]. S. Sarraf and G. Tofighi, “Deep learning-based pipeline to recognize Alzheimer’s disease using fMRI data,” in 2016 Future Technologies Conference (FTC), 2016, pp. 816–820.

[76]. S. Sarraf, “5g emerging technology and affected industries: Quick survey,” Am. Sci. Res. J. Eng. Technol. Sci., vol. 55, no. 1, pp. 75–82, 2019.

[77]. S. Gidaris and N. Komodakis, “Object detection via a multi-region and semantic segmentation-aware cnn model,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1134–1142.

[78]. X. Wang, L. Gao, J. Song, and H. Shen, “Beyond frame-level CNN: saliency-aware 3-D CNN with LSTM for video action recognition,” IEEE Signal Process. Lett., vol. 24, no. 4, pp. 510–514, 2016.

[79]. A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, “Action recognition in video sequences using deep bi-directional LSTM with CNN features,” IEEE access, vol. 6, pp. 1155–1166, 2017.

[80]. S. Sarraf, D. D. Desouza, J. A. E. Anderson, and C. Saverino, “MCADNNet: Recognizing stages of cognitive impairment through efficient convolutional fMRI and MRI neural network topology models,”

IEEE Access, vol. 7, pp. 155584–155600, 2019.

[81]. N. Kalchbrenner, E. Grefenstette, and P. Blunsom, “A convolutional neural network for modelling sentences,” arXiv Prepr. arXiv1404.2188, 2014.

[82]. O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn, “Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition,” in 2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP), 2012, pp. 4277–4280.

[83]. K.-Y. Huang, C.-H. Wu, Q.-B. Hong, M.-H. Su, and Y.-H. Chen, “Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 5866–5870.

[84]. O. Abdeljaber, O. Avci, S. Kiranyaz, M. Gabbouj, and D. J. Inman, “Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks,” J. Sound Vib., vol. 388, pp. 154–170, 2017.

[85]. X. Peng, J. Hoffman, X. Y. Stella, and K. Saenko, “Fine-to-coarse knowledge transfer for low-res image classification,” in 2016 IEEE International Conference on Image Processing (ICIP), 2016, pp. 3683–3687.

[86]. S. Sarraf, C. Saverino, H. Ghaderi, and J. Anderson, “Brain network extraction from probabilistic ICA using functional Magnetic Resonance Images and advanced template matching techniques,” in 2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE), 2014, pp. 1–6.

[87]. S. Sarraf, “Analysis and Detection of DDoS Attacks Using Machine Learning Techniques,” Am. Sci. Res. J. Eng. Technol. Sci., vol. 66, no. 1, pp. 95–104, 2020.

[88]. S. Sarraf, D. D. DeSouza, J. Anderson, and G. Tofighi, “DeepAD: Alzheimer’s Disease Classification via Deep Convolutional Neural Networks using MRI and fMRI,” bioRxiv, p. 70441, 2017.

[89]. S. H. Sarraf, M. Soltanieh, and H. Aghajani, “Repairing the cracks network of hard chromium electroplated layers using plasma nitriding technique,” Vacuum, vol. 127, pp. 1–9, 2016.