Human Regular Activities Recognition using Convolutional Neural Network

Abstract

Capturing commonly occurring behaviors is a tough issue in computer vision. A few of them are recreation, touring, leisure pursuits, and religious practice. A comprehensive effort has already been dedicated to this aspect to deal with this issue. We have created a dataset with five categories, including household activities, farming, exercise, sports, and occupation, to identify human daily actions. This collection still has 4328 colored images in total. 630 are set aside for testing, and 3698 for training. Deep learning and standard image-based strategies are being explored to address the issues. In this piece, we have designed a deep learning paradigm to classify the regular activities of human beings. To characterize people's daily chores, we use the CNN model, one of the greatest tools for visual identification. CNN is often used to indicate these massive operations. We also have chosen two already-trained VGG16 and ResNet50 models. When we compare it with the existing techniques, the investigation's findings demonstrate that the suggested network has a recognition accuracy of 91%. Additionally, we have discovered that accuracy varies throughout different epochs. The reader may find this article instructive in grasping CNN models for various recognizing applications.

Keywords: Convolutional Neural Network (CNN), Human activity recognition, Deep Learning, Machine Learning.

1. Introduction

Recognizing human interaction is the practice of applying Artificial Intelligence (AI) to recognize and name human actions from raw records collected from a variety of sources [1]. A mechanism for locating and verifying numerous pixels revealed in a figure is known as image detection and recognition. It is a strategy that entails image processing, segmentation, extraction of important features, and matching identification [2]. In artificial vision systems, identifying images is a critical topic. The area of computer vision included in artificial intelligence aims to give machines the same ability as people to comprehend information from images. Image, segmentation, localization, and object detection are examples of problems in computer vision. Recognizing Images is the most significant of these problems, and it is the foundation for each subsequent machine sight difficulty. Depending on how challenging the classification problem at hand is, there are two forms of image categorization [3]:

Binary Categorization: The most prevalent recognition issue in supervised categorization and identification of images is single-label categorization. For each image in single-label categorization, merely one notation or tag is used. As a result, the model generates a single value or forecast for each image it views. The model generates a vector with a length equal to the sum of classes and a value to determine whether a picture belongs to a specific class. Binary classification, which has only two classes, or multiclass categorization, which appears to comprise more than two, are examples of single-label characterization [3].

Multi-class Categorization: Multi-class categorization is a classification job in which each image might have several labels or annotations, with some figures bearing all of the labels simultaneously. While the issue statement appears to be comparable to single-label categorization in some ways, the problem statement is more complex. Multi-label assessment tasks are widely used in the diagnostic ultrasonography area, where a patient who may have multiple diseases can be diagnosed using visual data such as X-rays [3]. The convolutional neural network that captures human typical actions is displayed in broad strokes in Figure 1.

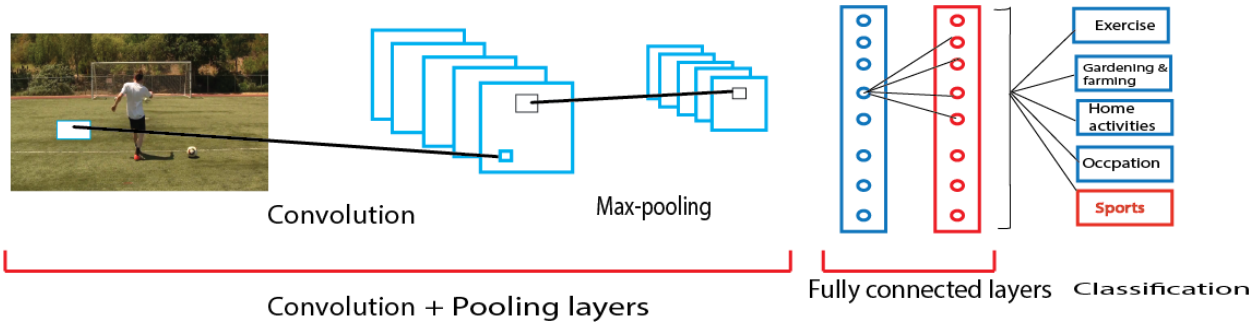


Fig 1: Convolutional neural network layout, which accepts an image as input. Convolution, pooling, and an entirely integrated layer are applied to the image while CNN classifies it.

Image recognition and categorization work by accepting a depiction built up of pixels that a device analyzes it. The device fulfills this by treating the sight as a series of matrices, the size of which is determined by the image resolution. From a computer's perspective, the study of statistical data utilizing algorithms is simply speaking, picture classification. In the processing of digital images, image categorization is achieved by grouping pixels into defined groups, or classes in an automatic manner [4]. These algorithms divide the photograph into a series of key attributes, reducing the workload for the final classifier. These factors aid the classifier to figure out what the image is about and which class it belongs to because the rest of the stages are dependent on it. The characteristic extraction process is the most critical step in categorizing an image. The data provided by the algorithm is also crucial in the categorization of images, especially supervised classification. As opposed to a terrible dataset with class-based data imbalance and low image and annotation quality, a better predictive set of data performs admirably [4]. Image categorization in still photos, after years of research, remains a difficult task due to the unconstrained scenario. This study examines how to estimate image recognition problems and how to solve them. We use a combination of transfer learning and traditional image-based approaches to achieve our goal. In computer vision problems, deep learning techniques have been widely deployed. CNN is one of these techniques, and conquering popularity in picture categorization day after day. CNN is particularly effective for tasks such as classification, processing, detection, and segmentation because it can extract patterns and representations from a given input image with greater precision and accuracy. CNN is a forward-feeding learning algorithm and is exceptionally the best suited for reducing the volume of parameters keeping the same model quality. Images are multidimensional, with each pixel having the criteria described above for CNNs.
The features maps from the preceding layer are convolved using learnable kernels and passed through the activation function at a convolution layer to create the output feature map in the backpropagation algorithm. Convolutions with numerous input maps may be combined in each output map[5]. Typically, we have that

$$x_j^p = f\left(\sum_{i \in Mj} x_i^{p-1} * k_{ij}^p + b_j^p\right)$$

where p indicates the current layer and Mj is a collection of input maps. The bias b is applied to each output map, however, for a given output map, separate kernels are utilized to convolve the input maps. In other words, if output map j and map k both sum over input map i then the kernels used to transform map i differ for output map j and output map k.
In the feedforward pass, the error is generated by a multiclass problem with c classes and N training examples.

$$E^N = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{c}(t_k^n - y_k^n)^2$$

Where y_k^n is the value of the k-th output layer unit and t_k^n is the matching k-th dimension of the target for the n-th pattern (label).

Sigmoid function:

$$f(s)_i = \frac{e^{s_i}}{\sum_j^c e^{s_j}}$$

Categorical cross-entropy function:

$$CE = -\sum_i^c t_i \ log(f(s)_i)$$

Figures 2 and 3 respectively depict the procedures of the convolutional layer and the pooling layer.

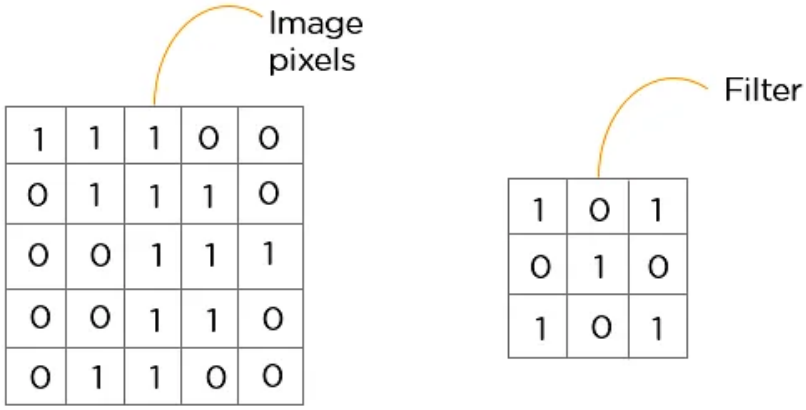


Fig 2: The operation held in the Convolution Layer finding the convolved matrix by sliding the filter array over the image, and measuring the dot product.

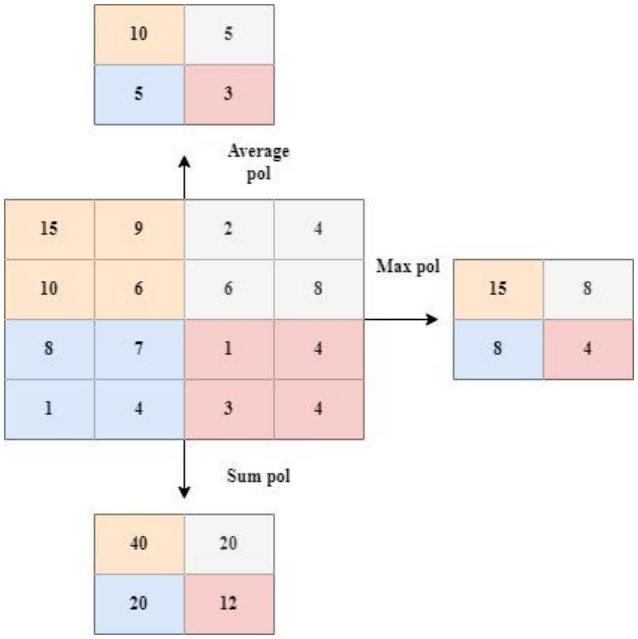


Fig 3: Pooling operations with maximum, average, and sum pooling

2. Literature Review

One of the vital areas of study and the most contentious issues in computer vision is categorization. The reason for its importance is the abundance of applications. Image recognition systems have advanced quickly in response to ever-increasing technical advancements, and have achieved tremendous progress in recent years. In [6], has used a neural net model to describe human behaviors using interaction and custom attributes. The UCI HAR data set, which consists of six daily activities, was applied to evaluate the model, and it had a classification accuracy score of 95.79 %. In [7], has developed a platform for hand movement recognition by using the CNN from the image. It has primarily targeted blind and deaf people who are unable to interact with others. They want to rely on a kind of visual correlation for that. The fine verbal exchange platform provided by sign language allows hearing-impaired men or women to bring their minds and connect with a regular character.

In [8], it is proposed that CNN's classify human activities using unprocessed data from a cluster of sensing devices, and 16 lower limb behaviors have been observed inside this dataset, a bundle of individuals with five distinct sensors. The potential for classification triple, double, and single sensor systems has been probed using a diverse range of combinations of activities and sensors, demonstrating how motion signals can be modified to be fed into CNNs using different access architectures, and contrasting the performance of various groups of sensors. In [9], debuted the first method for HAR based on deep learning models and produced an image of a spectrogram from an inertial signal feeding actual images to a network of convolutional neurons. [10], has further expanded to categorize normal duties employing data from an expansive dataset of topmost moves using a multi-channel multilayer system that utilizes signals including both circular velocity and speed. In [11], has proposed to use of the Dynamical Convolutional Network architecture for recognizing the activities from the sensor data obtained from a smartphone. Its potential architecture has a better ability to record enduring interactions and variable-length input sequences. In [12], had proposed a strategy for HAR dilemma characteristic learning that is periodic and exploits deep neural CNN to consistently automate feature learning via raw inputs. The taught features had got additional discriminative strength by utilizing the labeled data through supervised learning.

In [13], had shared weights for the entirety of the input signals in the convolutional layer (full weights sharing), and extracted the same features without separating modalities for multi-modal data, which could cause interferences between characteristics produced by accelerometers and gyroscopes for capturing modality-specific features. In [14], has proposed a technique for Human Action Recognition (HAR) that uses a CNN. The first data was gathered from 20 initiatives undertaken by 7 subjects from the MSR 3D Action data set as well as 10 actions conducted by 6 subjects and acquired by the v2 sensor for the Kinect. It was carried out on a total of 39715 images and 97.23% accuracy was achieved on the Kinect data set, and 87.1% on the MSR data set. In [15], have employed the trained dataset and an improved architecture for categorizing images using convolutional neural networks to find and identify tasks. It aids scientists in better understanding how CNN models work for different image categorization tasks.

In this essay, we construct a deep CNN that can effectively handle the activity recognition assignment, and then we compare the research findings to the pre-trained classifiers.

3. Materials and Methods

3.1 Dataset Preparation

We collect samples of five different types of people namely exercise, gardening and farming, home activities, occupation, and sports. These samples are divided into two parts: The training dataset the and Test dataset. There is a total of 4328 colored images in this collection, which are assembled into five categories: exercise, gardening and farming, home activities, occupation, and sports. 3698 of these images have already been separated for training, with the remaining 630 being used for testing. Figure 4 offers six selections of pictures from each of the dataset's five classifications and figure 5 illustrates the general sequence for our assignment.

Fig 4: Dataset having five categories namely Exercise, Gardening & farming, Home activities, Occupation, and Sports

Table 1An overall description of our dataset having five categories

Name of class	Image per class	Number of training images	Number of testing images
Exercise	1200	1000	200
Gardening and farming	832	732	100
Home Activities	657	577	80
Occupation	439	389	50
Sports	1200	1000	200

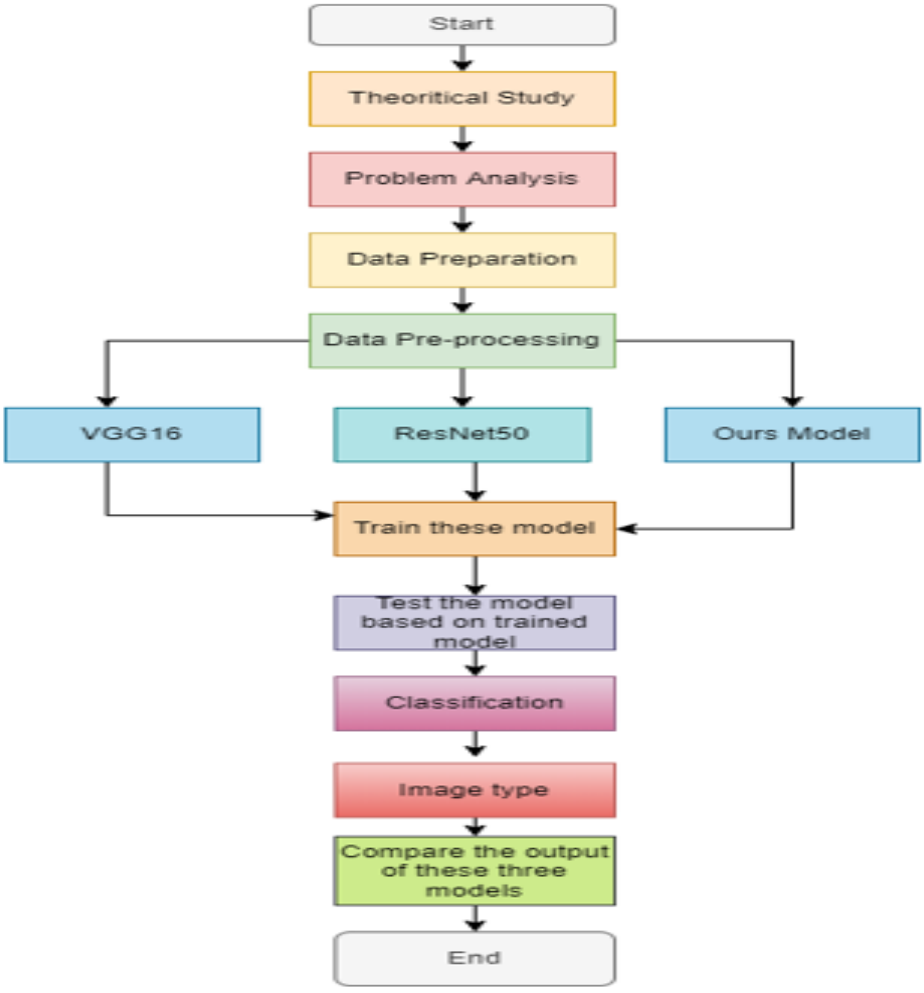


Fig 5: The overall workflow of our task

3.2 Implementation

We have implemented our model as follows:
Important packages such as TensorFlow, Keras, NumPy, and Matplotletc are imported. The prepared dataset of images is uploaded to google drive and then imported to Colab. Reading the dataset, data is preprocessed using the ImageDataGenerator() function that is imported from the preprocessing section of the package, Keras. An orderly model is created, then CNN layers like Conv2D, MaxPooling2D, and Dense are added to it. The completely interconnected layer is referred to be dense. The input shape of the images must be defined in the first layer. Flatten and Dropout functions are also included in the model, with flatten function converting the input matrix into an array that is only one dimension and the dropout function dealing with overfitting. The model is compiled and built by using the compile () and fit () functions. As an optimizer “adam”, as loss function “categorical_crossentropy”, and “accuracy” metrics during compiling are used.

The testing accuracy of the model is evaluated in this section, and after the prediction, the expected output is determined. Figure 6 demonstrates how our strategy is implemented.

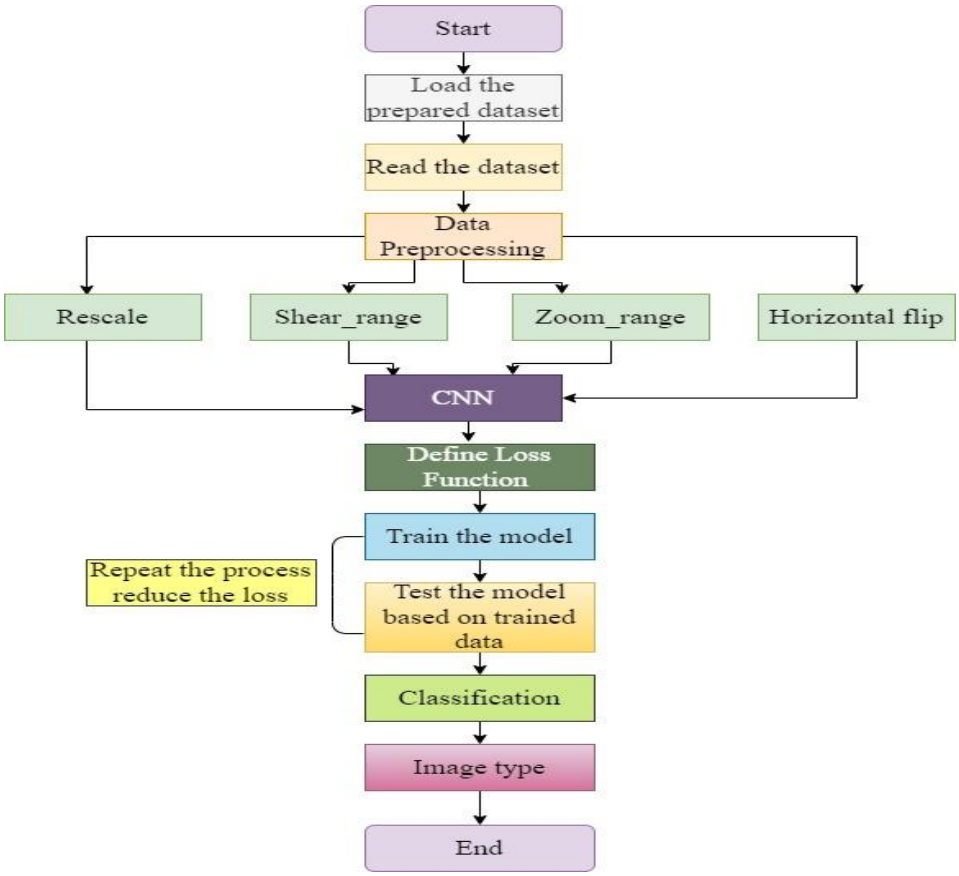


Fig 6: The implementation procedure of the postulated model

The architecture of the envisioned model:

To boost the network's expression ability and speed, a simple architecture has been created. However, our goal is to create a simple optical system. The following is the model's layout:

Convolutional input layer

The layer absorbs information from the input image in a variety of ways. The number of filters is 32, with a size of 3×3, a stride of 1, padding of 0, a rectifier activation function, and a max norm weight restriction of 3.

Max Pool layer

The pooling Layer is usually applied after a Convolutional Layer. The target is to scale down the convolved characteristic chart of this layer to save on computational costs. By reducing the linkages between layers and conducting independent operations on each featured grid, this is

established. This obtains the largest element from the characteristic chart in Max Pooling. The volume of the filter is 2×2.

Convolutional layer
The number of filters is 64, with a size of 3×3, a stride of 1, padding of 0, a rectifier activation function, and a max norm weight restriction of 3.

Max Pool layer
Another max pooling layer with size 2×2.

Dropout
Dropout is set to 25% to handle over-fitting issues.

Flatten layer
Flatten is used to convert the matrix into a one-dimensional array.

The Fully connected layer
The input images of the previous layers are flattened and supplied to the FC layer in this step. The normal arithmetic workable calculations are then performed on the flattening variable via a few additional fully connected layers in our model, which has 128 units and a rectifier activation function dropout with a second dropout set to 50%. Our model has 128 units and a feature for activating rectifiers.

Dropout
The second dropout is set to 50%.

The fully connected output layer
It has 5 units and a SoftMax activation function.

Figure 7 represents the layers that were employed, and figure 8 describes these layers in our framework.

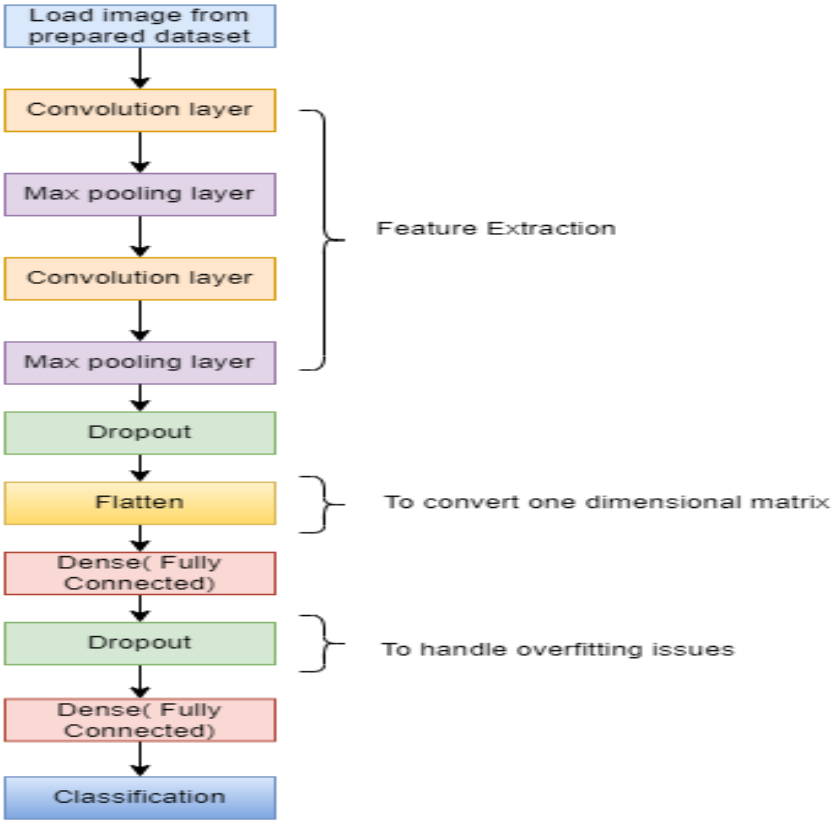


Fig 7: Convolution, Max-pooling for feature extraction, Dropout for mitigating over-fitting concerns, and Flattening for converting one-dimensional arrays are all layers in our model's heterogeneous network.

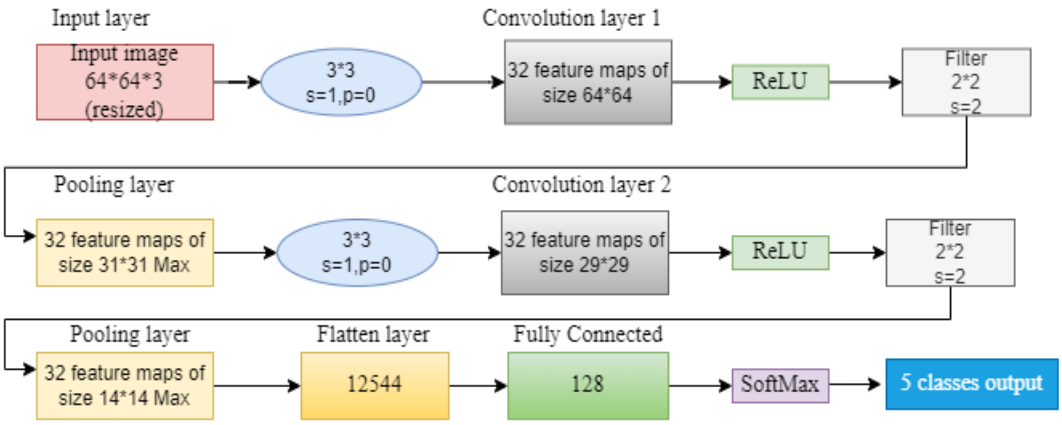


Fig 8: Layer description of the structure, where p denotes padding and s implies stride

4. Model Evaluation and Results Analysis

A disparity of the suggested approach with previously reported approaches for the sorting of images and recognition is offered in this portion to establish that the proposed network has a good performance for current objectives. The suggested architecture's performance is assessed using our dataset for recognition and classification tasks. The dataset bears 4528 color images in 5 major categories, with 3898 images as a training dataset and 630 images as a test dataset: exercise, gardening and farming, home activities, occupation, and sports. Each image is 64*64 pixels in size. The network is taught using the Adam with the proportion of items built during a production run.

Table2 : The suggested network's overall accuracy in comparison to current techniques

Model	Loss	Training Accuracy	Validation Loss	Validation Accuracy
VGG16	0.6649	0.7480	0.5639	0.7928
ResNet50	1.1007	0.6391	0.9474	0.7109
Our model	0.4896	0.8217	0.2922	0.9112

The table is represented by a histogram where a comparison among used three models is given in figure 9:

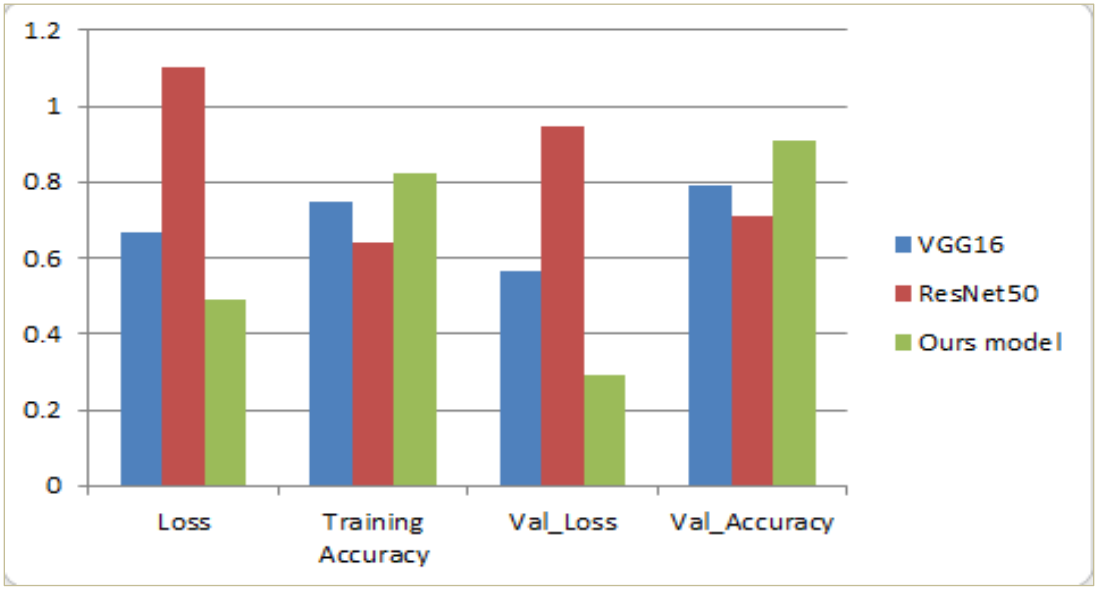


Fig 9: The comparison of three models (VGG16, ResNet50, and our model)

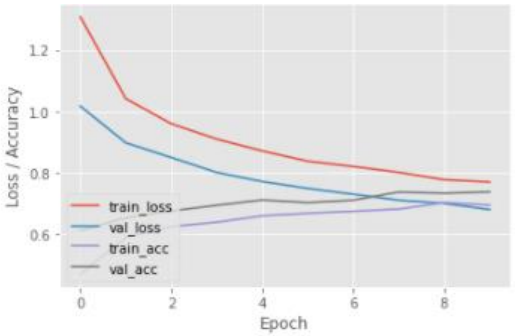
Table 3: The relationship between epochs and accuracy

Model	Number of epochs	Validation Accuracy
VGG16	10	73.73%
VGG16	25	79.28%
ResNet50	10	51.59%
ResNet50	25	71.09%
Our model	10	77.53%
Our model	25	91.12%

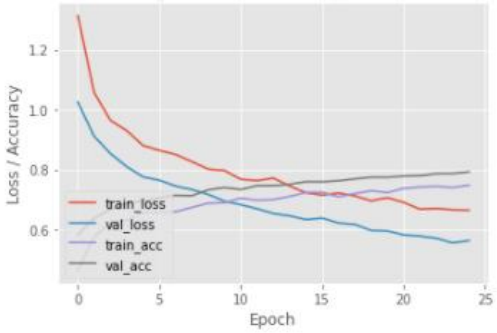
Discussion

Based on findings from our dataset's picture categorization study, this paper provides a fundamental deep neural network. The suggested technique outperforms growth models in terms of enhancing the total precision of the living item identification procedure. It is difficult to decide which classifier is the best since classifier selection can vary according to the requirements. The classifier chosen will depend on the classification challenge because numerous factors can reduce classification accuracy. From our experiment, we have achieved 79.28% accuracy from the VGG16 model, 71.09% accuracy from the ResNet50 model, and 91.12% accuracy from our proposed model. Among VGG16 and ResNet50, our network is faster to train, uses less memory, runs smoothly, and has simple parameters to tweak. Similarly, our system’s speed is adequate for situations involving an abundance of samples or features. We have noticed from our experiment output that both the training and testing loss of the ResNet50 model is comparatively high. For that, the accuracy of this model is also low. We also have found that both training and testing loss of VGG16 is medium than ResNet50 and our model. And its accuracy is also medium compared to others. For our model, we have gained a lower loss than the other two models. And its accuracy is better than ResNet50 and VGG16. It is observed that the accuracy differs with the number of epochs. By improving the number of epochs, the accuracy increases. For these models, we selected 10 and 25 as the epoch sizes. Regarding the VGG16 model, we discovered accuracy in 73.73% of the 10 epochs and 79.28% of the 25 epochs. We have accomplished an accuracy of 51.59% of the amount of epochs 10 and 71.09% of the epoch count 25 for the ResNet50 model. In terms of precision, we have covered 91.12% of epochs 25 and 77.53% of epochs 10. We can summarize that our model has provided lower loss and is better than the VGG16 and the ResNet50 model. The plot of loss and accuracy for the proposed model, the ResNet50, and the VGG16 are indicated in Figure 10.

VGG16 model

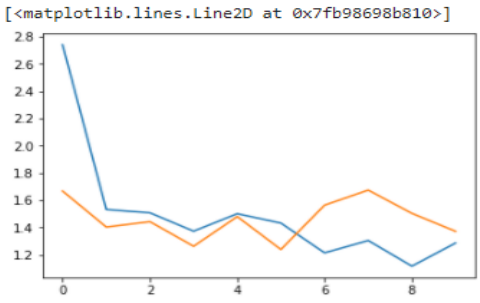


Epoch =10

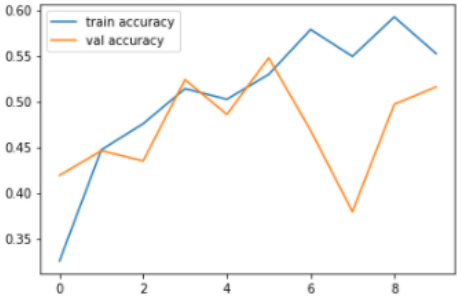


Epoch =25

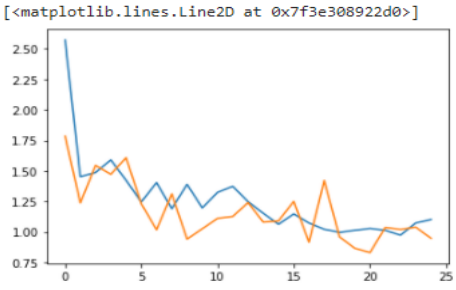
ResNet50 model



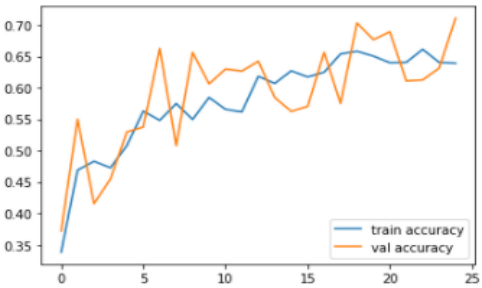
Loss Plot (Epoch=10)



Accuracy Plot (Epoch=10)

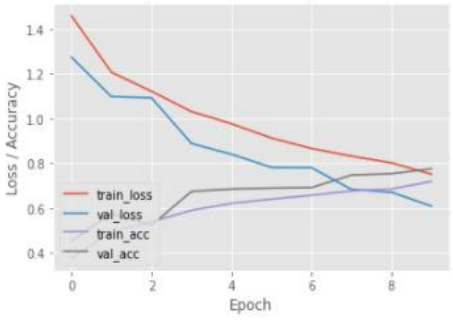


Loss Plot (Epoch=25)

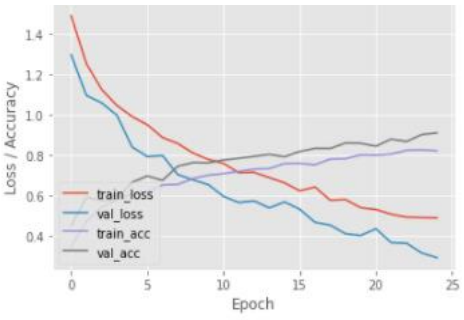


Accuracy Plot (Epoch=25)

Proposed Model



Loss Plot Epoch=10



Accuracy Plot Epoch=25

Fig 10: Representation of Loss and Accuracy plot for VGG16, ResNet50, and our Proposed model

We have also noticed that VGG16 and ResNet50 require more computational time than our model. These two models are comparatively slow because these two models are pre-trained. These were trained on the “ImageNet” dataset. These two models take additional time because of that.
Each of the existing techniques has pros and cons because the basic goal of this study is to assess grouping accuracy.
However, the framework's performance in terms of classification accuracy is strong.
Depending on these research trends, it’s clear that the above-mentioned methodologies’ categorization accuracy varies for different challenges in different settings.
To summarize, a challenging issue is to fix which classifier performs properly because it is entirely based on the type of data, image size, parameter adjustment, and other factors.

5. Conclusion

In this document, we have used our dataset which has been categorized into five classes and we have taken two pre-trained models namely VGG16 and ResNet50. And we have built a different model. From our experiments, we have gained better accuracy from our models than the VGG16 and the ResNet50 model. This paper also briefly reviews the basics of CNNs and their spectacular growth across a broad spectrum of computer vision applications over the recent years, such as object detection, posture prediction, scene interpretation, visual segmentation, and so on. These findings suggest that classifying photos using deep learning can produce accurate results. However, a few problems still need to be rectified. A core network has been presented as a solution to the identification issue with photographs. The suggested strategy calls for less memory and processing power. In comparison to traditional approaches, the model improves categorization accuracy and delivers good recognition outcomes. Besides, the network's execution assessment appears that it can be utilized to develop a significantly speedier classifier. With a portrait as the input, the specified organization can deal with countless obstacles for a variety of applications. To summarize, this article aims to help searchers, masters, and readers in a way better realize the request for pictures, and categorization and discover a worthy arrangement. Small and 2D images are employed in the training process. When we compared to typical JPEG images, the processing time for these images is extremely long. Using clusters of GPUs to stack the model with additional layers and develop the prototype with more picture data will result in more accurate image classification results. The next update will emphasize locating large-scale 3D images, as these will be key for the image segmentation process.

Compliance with Ethical Standards

Ethical Approval: This article does not contain any studies with human participants or animals performed by any of the authors.

REFERENCES

[1] N. Gupta, S. K. Gupta, R. K. Pathak, V. Jain, P. Rashidi, and J. S. Suri, "Human activity recognition in artificial intelligence framework: a narrative review," *Artif Intell Rev*, vol. 55, no. 6, pp. 4755–4808, Aug. 2022, doi: 10.1007/s10462-021-10116-x.

[2] K. S. Assistant Professor and D. H. Inbarani Assistant Professor, "A comparative analysis of Convolution Neural Network structures for image classification." [Online]. Available: http://adalyajournal.com/

[3] H. Bandyopadhyay, "Image Classification in Machine Learning [Intro + Tutorial]." https://www.v7labs.com/blog/image-classification-guide (accessed Oct. 26, 2022).

[4] G. Boesch, "A Complete Guide to Image Classification in 2022 - viso.ai." https://viso.ai/computer-vision/image-classification/ (accessed Oct. 26, 2022).

[5] J. Bouvrie, "Notes on Convolutional Neural Networks," 2006.

[6] S. K. Bashar, A. al Fahim, and K. H. Chon, *Smartphone Based Human Activity Recognition with Feature Selection and Dense Neural Network*. 2020. doi: 10.0/Linux-x86_64.

[7] X.-S. Yang and Institute of Electrical and Electronics Engineers, *Proceedings of the World Conference on Smart Trends in Systems, Security and Sustainability (WS4 2020) : July 27-28, 2020, virtual conference*.

[8] A. Bevilacqua, B. Caulfield, K. Macdonald, A. Rangarej, V. Widjaya, and T. Kechadi, "Human Activity Recognition with Convolutional Neural Networks CATCH: Connected Health for Cancer View project Utilization of inertial measurement units to analyse lower limb movement in athletes with chronic ankle instability during sports related tasks View project Human Activity Recognition with Convolutional Neural Networks," 2018. [Online]. Available: https://www.researchgate.net/publication/327667610

[9] M. A. Alsheikh, A. Selim, D. Niyato, L. Doyle, S. Lin, and H.-P. Tan, "Deep Activity Recognition Models with Triaxial Accelerometers," Nov. 2015, [Online]. Available: http://arxiv.org/abs/1511.04664

[10] IEEE Computational Intelligence Society, International Neural Network Society, Institute of Electrical and Electronics Engineers, and B. C.) IEEE World Congress on Computational Intelligence (2016: Vancouver, *2016 International Joint Conference on Neural Networks (IJCNN): 24-29 July 2016, Vancouver, Canada.*

[11] N. Nair, C. Thomas, and D. B. Jayagopi, “Human activity recognition using temporal convolutional network,” in *ACM International Conference Proceeding Series*, Sep. 2018. doi: 10.1145/3266157.3266221.

[12] J. Bo Yang, M. Nhut Nguyen, P. Phyo San, X. Li Li, and S. Krishnaswamy, “Deep Convolutional Neural Networks On Multichannel Time Series For Human Activity Recognition.”

[13] S. Ha, J. M. Yun, and S. Choi, “Multi-modal Convolutional Neural Networks for Activity Recognition,” in *Proceedings - 2015 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2015*, Jan. 2016, pp. 3017–3022. doi: 10.1109/SMC.2015.525.

[14] Z. Ahmad, K. Illanko, N. Khan, and D. Androutsos, “Human action recognition using convolutional neural network and depth sensor data,” in *ACM International Conference Proceeding Series*, Aug. 2019, pp. 1–3. doi: 10.1145/3355402.3355419.

[15] M. Aamir, Z. Rahman, W. Ahmed Abro, M. Tahir, and S. Mustajar Ahmed, “An Optimized Architecture of Image Classification Using Convolutional Neural Network,” *International Journal of Image, Graphics and Signal Processing*, vol. 11, no. 10, pp. 30–39, Oct. 2019, doi: 10.5815/ijigsp.2019.10.05.