Human Regular Activities Recognition Using Convolutional Neural Network

Sharmin Akther, Ayat Ullah Nahid

Asian Journal of Research in Computer Science

Volume 15, Issue 1, Page 44-55, 2023; Article no.AJRCOS.95285
ISSN: 2581-8260

Human Regular Activities Recognition Using Convolutional Neural Network

Sharmin Akther [a*] and Ayat Ullah Nahid [b]

[a] Department of Information and Communication Engineering, Noakhali Science and Technology University, Noakhali-3814, Bangladesh.
[b] Department of Agriculture, Noakhali Science and Technology University, Noakhali-3814, Bangladesh.

Authors' contributions

This work was carried out in collaboration between both authors. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJRCOS/2023/v15i1314

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://www.sdiarticle5.com/review-history/95285

Original Research Article

Received: 02/12/2022
Accepted: 06/02/2023
Published: 07/02/2023

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. In this work, we recreated a dataset with five categories, including household activities, farming, exercise, sports, and occupation, to identify human daily actions. This collection has 4328 colored images in total, among them 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 paper, 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. We also have chosen two already-trained VGG16 and ResNet50 models. When we compare our model with the existing techniques, the investigation's findings demonstrate that the suggested network has a better recognition accuracy of 91%. Additionally, we have observed that accuracy varies throughout different epochs, and after 25 epochs we got better stable results from our model. The reader may find this article instructive in grasping CNN models for various recognizing applications.

**Corresponding author: E-mail: sharminakther6351@gmail.com;*

Keywords: Convolutional neural network (CNN); human activity recognition; deep learning; machine learning.

1. INTRODUCTION

Recognizing human interaction in the practice of Artificial Intelligence (AI) means identifying human actions from the raw records collected from a variety of sources [1]. On the other hand, 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. Choosing specific types, sometimes known as classes, of objects inside an image or video frame is vital work in the rapidly expanding field of computer vision [3]. 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. Image recognition systems are being quickly adopted by a wide range of industries, including security, healthcare, education, fintech, manufacturing, telecom, utility, and defense, to improve their visual data processing and analysis capabilities [3]. The top image recognition applications include object detection, optical character recognition (OCR), face recognition, fraud detection, visual search, and image captioning, as well as others [3]. For image recognition using deep learning at first, we have to classify the object from the images. Depending on how challenging the classification problem at hand is, there are two forms of image categorization [4]. Binary Categorization is the most prevalent recognition issue in supervised categorization and identification of images. 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 [4]. 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 [4].

Image recognition and categorization worked by accepting and analyzing a set of pixels from images. The device fulfills this by treating the sight as a series of matrices or vectors, the size of which is determined by the input image resolution. From a computer's perspective, the study of statistical data utilizing algorithms is simply speaking as picture classification. 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 [5]. In this study, we examine various image classification models for human pose recognition. We use a combination of transfer learning and traditional image-based approaches to achieve our goal. Moreover, we proposed a CNN based model to examine the same classification task. 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

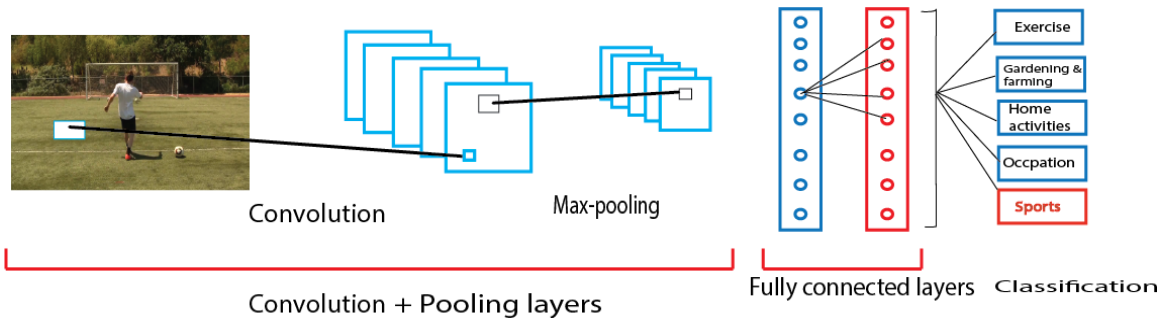


Fig. 1. Convolutional neural network layout, which accepts an image as input. Convolution, pooling, and an entirely integrated layer are applied to the image while CNN classifies it

convolution layer to create the output feature map in the backpropagation algorithm. Convolutions with numerous input maps may be combined in each output map [6]. The typical convolutional neural network that captures human typical actions is displayed in Fig. 1. Typically, for each convolutional layer we can get the output feature maps from the input feature maps using the Equation (1).

$$X^{m^{\times}n} = Con(X^{M^{\times}N}, K^{q}) \quad (1)$$

Where the output feature dimension (m,n) can be calculated from input dimension (M,N) by following equation (2).

$$n = \left(\frac{N + 2p - 1}{s}\right) + 1 \quad (2)$$

where p indicates the padding, K indicates the kernel and S indicate the stride. In the each forward pass, the error is generated by a multiclass problem with c classes and N training examples is calculated by the given in Equation (3).

$$E^{N} = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{c}(t_k^n - y_k^n)^2 \quad (3)$$

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).

We use Sigmoid function at the last of the dense layer which is given by Equation (4):

$$f(s)_i = \frac{e^{s_i}}{\sum_j^c e^{s_j}} \quad (4)$$

Another loss function we also consider for our task categorical cross-entropy function is given by Equation (5):

$$CE = -\sum_i^c t_i \log(f(s)_i) \quad (5)$$

Figs. 2 and 3 respectively depicts the procedures of the convolutional layer and the pooling layer.

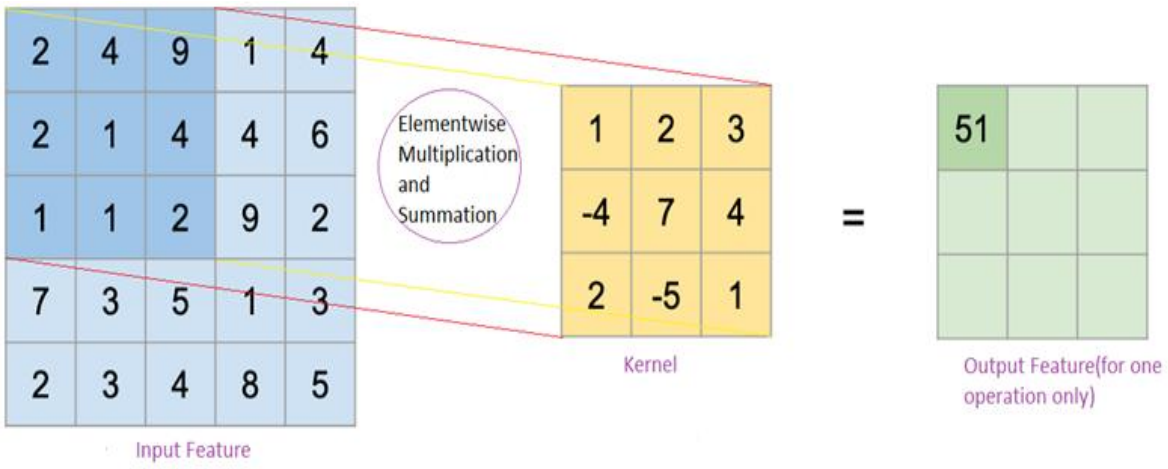


Fig. 2. The operation held in the convolution layer finding the convolved matrix by sliding the filter array over the image, and measuring the dot product

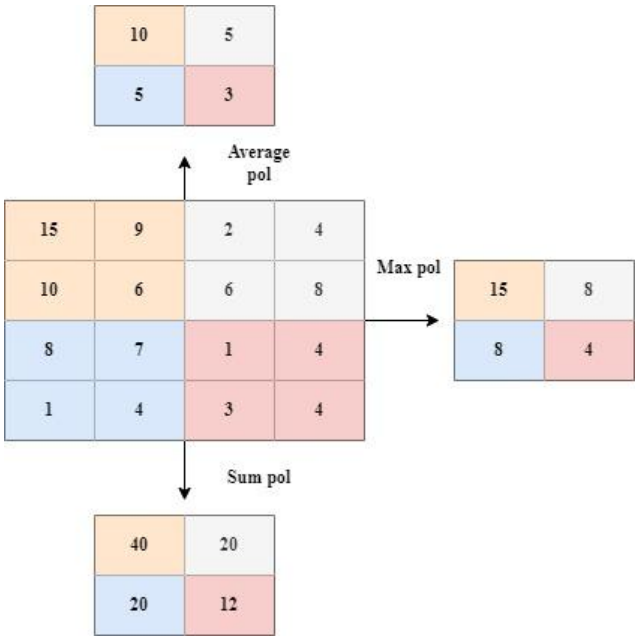


Fig. 3. Pooling operations with maximum, average, and sum pooling [public domain image

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. The UCI HAR data set, which consists of six daily activities, was applied to evaluate the CNN model and got a better classification accuracy in [7]. In [8] 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 [9], it is proposed that CNNs classify human activities using unprocessed data from a cluster of sensing devices, and 16 lower limb behaviors have been observed inside this dataset. The potential for classification triple, double, and single sensor systems have 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 [10] the author 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. In [11] author has proposed to use of the Dynamical Convolutional Network architecture for recognizing the activities from the sensor data obtained from a smartphone. In [12] it has proposed a strategy for HAR dilemma characteristic learning that is periodic and exploits deep neural CNN to consistently automate feature learning via raw inputs.

In their paper [13] the authors have 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. In [14] the author has proposed a technique for Human Action Recognition (HAR) that uses a CNN. 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] authors 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 [16] convolutional neural networks (CNN) and bidirectional long short-term memory (BLSTM) were combined to propose a deep learning multi-channel architecture, and the suggested model was tested on two publicly accessible datasets. In [17] the author offered information about the

data, filtering techniques, feature extraction techniques, classification, and various performance measurements and illustrated an overview of machine learning and deep learning methodology in HAR. In [18] in order to identify and categorize human poses, it suggested the "DeneSVM," a cutting-edge deep transfer learning-based classification model. The lying, bending, sitting, and standing postures were the four main postures that the paradigm was meant to categorize. In [19] for IoHT applications, a public dataset that was gathered using two smartphones (held in wrist and pocket locations) has been taken into consideration. In this paper, we first investigate the performance using transfer learning-based method then by our own proposed model which 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. In most of the cases, we collect datasets from the public domain. In addition to this, we also capture some of the images by ourselves. 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. Fig. 4 offers six selections of pictures from each of the dataset's five classifications and Fig. 5 illustrates the general sequence for our assignment.

3.2 Implementation

We use Tensorflow framework to develop our model. 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. We preprocessed our datasets by using different types (e.g rescale) of the preprocessor. Moreover, we applied data augmentation to our preprocessed data to get more datasets and to get a better result from the model. To do data augmentation we apply shear, horizontal flip, and zoom range. Then our prescribed architecture as shown in Figs. 6 & 7 is used for the classification task.

Fig. 4. Dataset having five categories namely exercise, gardening & farming, home activities, occupation, and sports

Table 1. An 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 this task

3.3 The Architecture of the Envisioned Model

At first, we use pre-trained models VGG-16 and ResNet50 for checking the model performance. Then we create a CNN architecture to compare the model performances. To boost the network's expression ability and speed, a simple architecture has been created. However, our goal is to create a simple but good performing system. In the following section, we describe our proposed model parameters.

In the input layer of our model, we used (64×64×3) color image as input. The number of filters we used at this layer was 32, with a size of 3×3, and a stride of 1. After that, we use Max Pooling layer. the same structure is repeated twice. The overall architecture of our model parameter is depicted in Fig. 8.

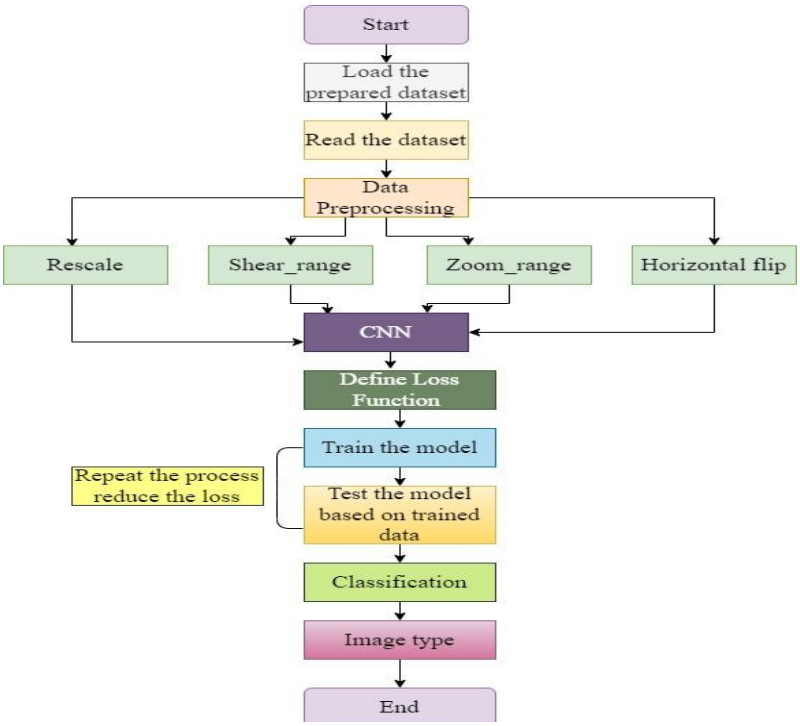


Fig. 6. The implementation procedure of the postulated model

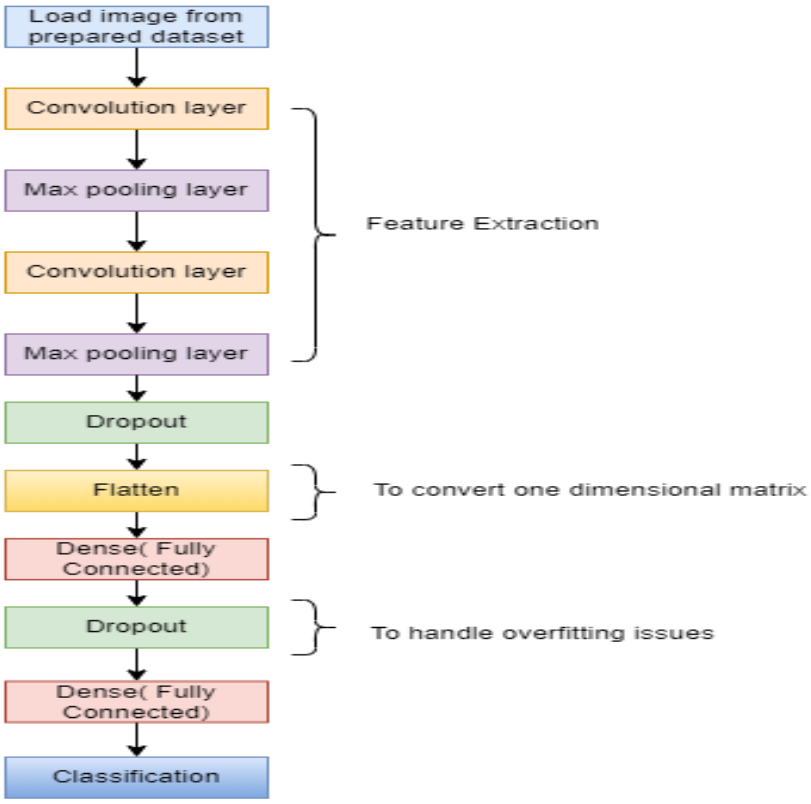


Fig 7. Convolution, max-pooling for feature extraction, dropout for mitigating over-fitting concerns, and flattening for converting one-dimensional arrays are all layers in our model's heterogeneous network

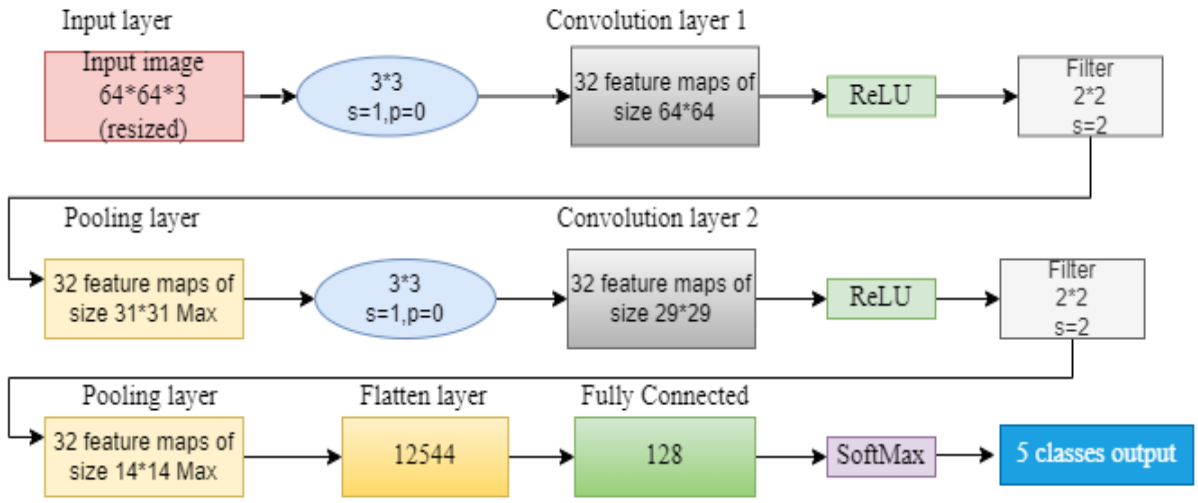


Fig. 8. Layer description of the structure, where p denotes padding and s implies stride

4. MODEL EVALUATION AND RESULTS ANALYSIS

A disparity of the suggested approach with previously reported approaches for the sorting of images and recognition are 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 4328 color images in 5 major categories, with 3698 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.

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

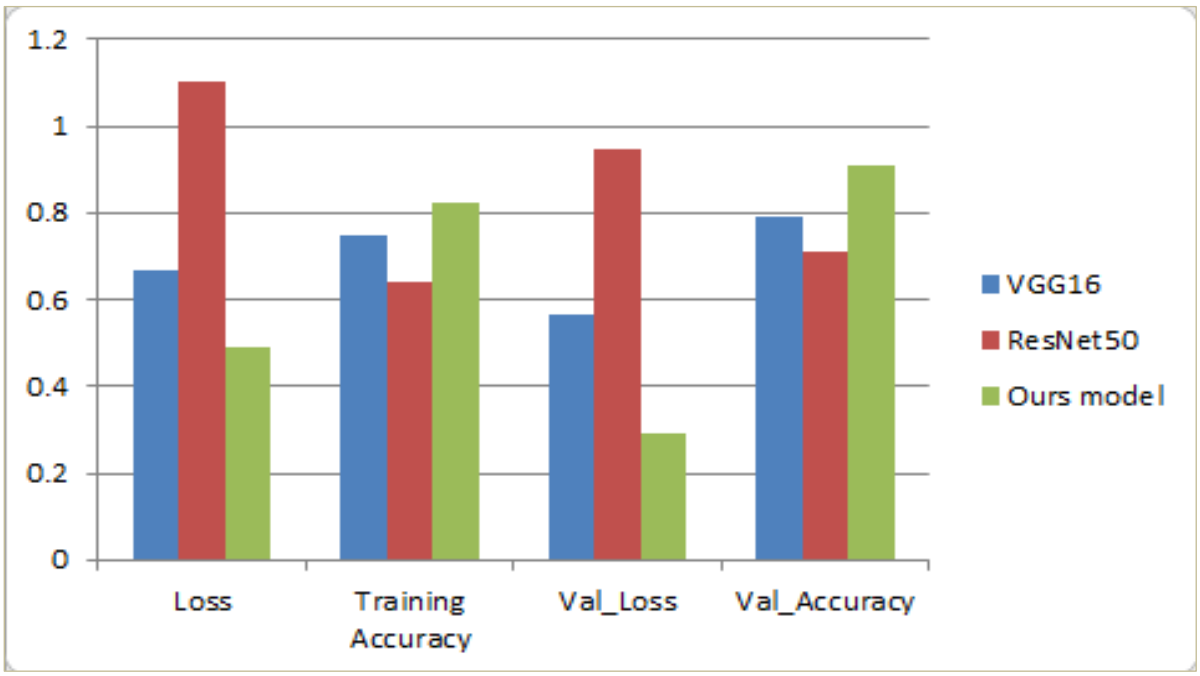


Fig. 9. The comparison of three models (VGG16, ResNet50, and our model)

Table 2. 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

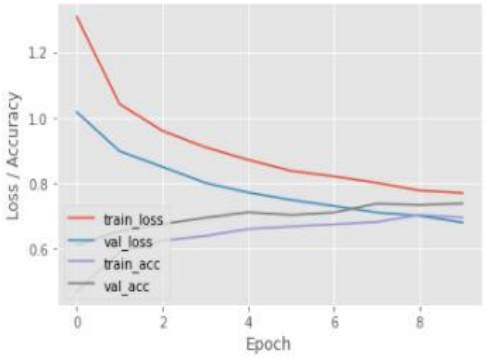
Table 3. The relationship between epochs and accuracy

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%

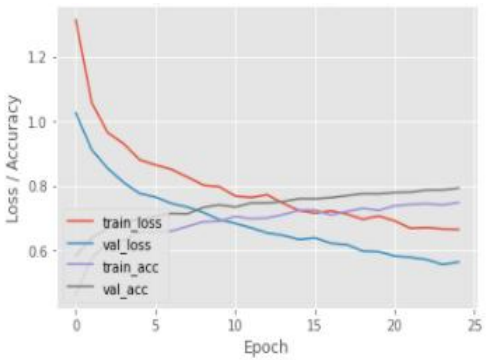
5. DISCUSSION

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 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 Fig. 10.

VGG16 model

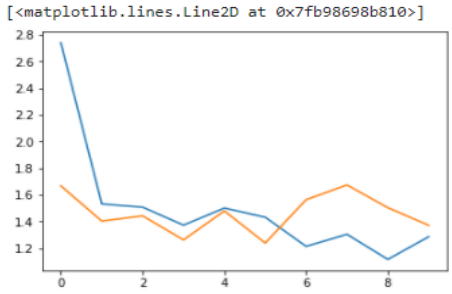


Epoch =10

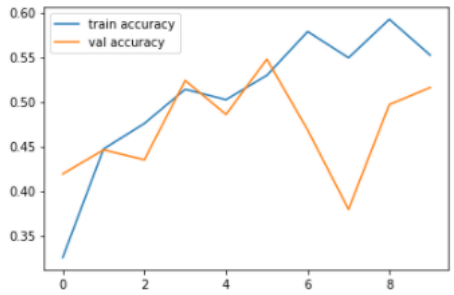


Epoch =25

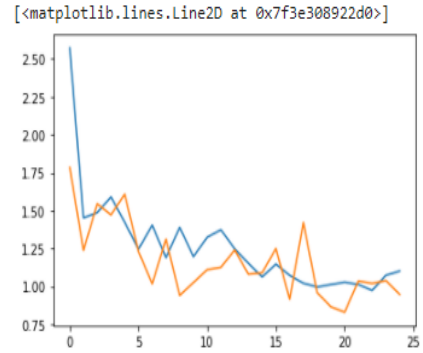
ResNet50 model



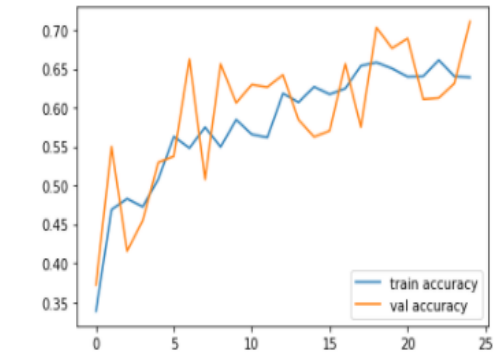
Loss Plot (Epoch=10)



Accuracy Plot (Epoch=10)

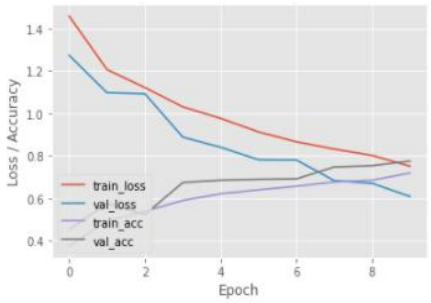


Loss Plot (Epoch=25)

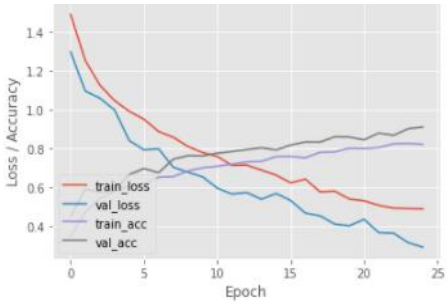


Accuracy Plot (Epoch=25)

Proposed Model



Loss Plot, Accuracy Plot (Epoch=10)



Loss Plot, 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. 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.

6. 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.

ETHICAL APPROVAL

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

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Gupta N, Gupta SK, Pathak RK, Jain V, Rashidi P, Suri JS. Human activity recognition in artificial intelligence framework: A narrative review. Artif Intell Rev. 2022;55(6):4755-808.
   DOI: 10.1007/s10462-021-10116-x, PMID 35068651.
2. KS. 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. Sadekov K. AI image recognition–applications and business benefits; May 30, 2022 [cited Feb 01, 2023].
   Available:https://mindtitan.com/resources/blog/ai-image-recognition-applications-and-benefits/.
4. Bandyopadhyay H. Image classification in machine learning [intro + tutorial] [cited Oct 26, 2022].
   Available:https://www.v7labs.com/blog/image-classification-guide.
5. Boesch G. A Complete Guide to Image Classification in 2022 – viso.ai [cited Oct 26, 2022].
   Available:https://viso.ai/computer-vision/image-classification/
6. Bouvrie J. Notes on convolutional neural networks; 2006.
7. Bashar SK, al Fahim A, Chon KH. Smartphone based human activity recognition with feature selection and dense neural network. Annu Int Conf IEEE Eng Med Biol Soc. 2020;2020:5888-91.
   DOI:10.1109/EMBC44109.2020.9176239:10.0/Linux-x86_64
8. X-S. Yang and institute of electrical and electronics engineers. proceedings of the world conference on smart trends in systems, security and sustainability, virtual conference. 2020;WS4(Jul 27-28).
9. Bevilacqua A, Caulfield B, Macdonald K, Rangarej A, Widjaya V, Kechadi T. 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.
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). Canada; 2016.
11. Nair N, Thomas C, Jayagopi DB. Human activity recognition using temporal

convolutional network in ACM International Conference Proceeding Series; 2018.
DOI: 10.1145/3266157.3266221.

12. Yang JB, Nhut Nguyen M, Phyo San P, Li XL, Krishnaswamy S. Deep convolutional neural networks on multichannel time series for human activity recognition.
13. Ha S, Yun JM, Choi S. Multi-modal convolutional neural networks for activity recognition. In: Proceedings IEEE International Conference on Systems, Man, and Cybernetics, SMC 2015, Jan. 2016; 2015:3017-22.
DOI: 10.1109/SMC.2015.525
14. Ahmad Z, Illanko K, Khan N, Androutsos D. Human action recognition using convolutional neural network and depth sensor data in *ACM* International Conference Proceeding Series. 2019:1-5.
DOI: 10.1145/3355402.3355419.
15. Aamir M, Rahman Z, Ahmed Abro WA, Tahir M, Mustajar Ahmed S. An optimized architecture of image classification using convolutional neural network. IJIGSP. 2019;11(10):30-9.
DOI: 10.5815/ijigsp.2019.10.05
16. Ihianle IK, Nwajana AO, Ebenuwa SH, Otuka RI, Owa K, Orisatoki MO. A deep learning approach for human activities recognition from multimodal sensing devices. IEEE Access. 2020;8: 179028-38.
DOI: 10.1109/ACCESS.2020.3027979
17. Alhumayani m, Monir M, Ismail r. machine and deep learning approaches for human activity recognition. IJICIS. Sep 2021;0(0): 1-9.
DOI: 10.21608/ijicis.2021.82008.1106
18. Ogundokun RO, Maskeliūnas R, Misra S, Damasevicius R. A novel deep transfer learning approach based on depth-wise separable CNN for human posture detection. Information. 2022;13(11): 520.
DOI: 10.3390/info13110520
19. Issa ME, Helmi AM, Al-Qaness MAA, Dahou A, Abd Elaziz MA, Damaševičius R. Human activity recognition based on embedded sensor data fusion for the Internet of healthcare things. Healthcare (Basel). 2022;10.
DOI: 10.3390/healthcare10061084, PMID 35742136

Peer-review history:
The peer review history for this paper can be accessed here:
https://www.sdiarticle5.com/review-history/95285