International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 06 | June - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Multiclass Classification using Enhanced MobileNet V2 Architecture

Ayesha Kazi[1], Prof. Sudhir Shikalpure[2], Prof. Vijayshri Injamuri[3]

[1]M . Tech, Department of Computer Science and Engineering, GECA
[2]Assistant Prof., Department of Computer Science and Engineering, GECA
[3]Assistant Prof., Department of Computer Science and Engineering, GECA

Abstract - In computer vision applications including object detection, face recognition, image classification, etc., the Convolutional Neural Network (CNN) predominates. MobileNet v2 is one such CNN architecture which can significantly cut down the parameters and cost of calculation while giving comparatively higher accuracy. In this paper, we propose the enhanced MobileNet architecture which shows an improvement in the accuracy. We have applied transfer learning techniques to fine-tune the MobileNet V2 model on the food dataset. The proposed model's generalization capacity, it's robustness and recognition accuracy shows a considerable amount of improvement with respect to the base architecture. The findings demonstrate that the enhanced MobileNet architecture's accuracy for food images recognition and classification has increased to 97.16%.

Key Words: Convolutional Neural Network (CNN), MobileNet v2, transfer learning, image recognition and classification

1.INTRODUCTION

The branch of artificial intelligence known as computer vision uses a number of technologies to help computers understand the physical world by translating human knowledge of visual understanding. Many facets of artificial intelligence have been transformed by neural networks, which enable superhuman accuracy for tasks including recognizing images, detecting faces or certain objects, analyzing traffic flows, medical imaging and the list goes on. As a substitute to the conventional method of employing manual features to analyze images, several deep learning models have been presented throughout the years. Among all the proposed architectures, CNNs have produced assuring outcomes.

MobileNet V2 is a prominent CNN architecture that is both efficient and effective for tasks like semantic segmentation, object identification, image classification, etc. It is an evolution of the original MobileNet architecture and was introduced by Google in 2018. MobileNetV2 is known for its exceptional balance between model size, computational efficiency, and accuracy, making it well-suited for a wide range of computer vision tasks on resource-constrained devices. MobileNet V2 has become a valuable asset in the field of deep learning, enabling the deployment of computer vision models on a wide range of devices, including smartphones, IoT devices, and edge computing systems. Its efficient architecture and customizable nature have contributed to its widespread adoption in real-world applications where resource constraints are a concern.

The MobileNet V2 has made it easier and more efficient to do image classification tasks like identifying different types of food. Multiclass food classification involves training a model to distinguish between various food items from input images. This task poses several challenges, including intra-class variations (different variations of the same dish), inter-class similarities (similar-looking dishes from different cuisines), and occlusions (partially visible or covered food items). MobileNetV2, with its lightweight yet powerful architecture, addresses these challenges by efficiently learning discriminative features from food images while minimizing computational overhead.

Food images in the same category are sometimes taken from diverse perspectives, with varying patterns, shapes, and sizes depending on the photographer; this presents a classification challenge. For instance, numerous items including glasses, bottles, and cutleries (eg. forks, knives and spoons) may be seen in the picture. Also, the similarity in the pattern and shape of different categories can contribute to the decrease in performance while classifying the images. For example, chapati and butter naan have almost the same shape and texture.

In this paper, we present an enhanced MobileNet V2 architecture which we then fine-tune on the food dataset to identify the image class and further categorize it as healthy or unhealthy.

2. RELATED WORKS

During the past several years, there has been a lot of activity in the field of deep neural architecture tuning to achieve the best possible balance between performance and accuracy. Numerous teams have improved training algorithms and conducted manual architectural searches, which have resulted in significant advancements over early designs like AlexNet, VGGNet, GoogLeNet, and ResNet. In recent times, there has been significant advancement in the research of algorithmic design, encompassing hyper-parameter optimization. With its novel architecture, MobileNetV2[1] transforms mobile convolutional neural networks (CNNs). Without sacrificing performance, it achieves exceptional efficiency by introducing linear bottlenecks and inverted residuals. The design ideas of MobileNetV2 are thoroughly described in this work, with a focus on accuracy, computing efficiency, and model size. MobileNetV2, with its excellent performance on COCO dataset detection and ImageNet classification tasks, is a major contribution to the field of deep learning for mobile and embedded devices. From edge computing to mobile vision, its small size and outstanding

performance make it an appealing option for a wide range of real-world applications.

The field of image recognition is altered by the groundbreaking work on deep residual learning by He et al[2]. It solves the difficulty of training deep convolutional neural networks (CNNs) by introducing residual connections, which allow the gradients to flow effectively. This method mitigates the vanishing gradient issue and makes training much deeper networks easier. The authors show improved performance over earlier techniques through a series of experiments conducted on ImageNet. The influence of Deep Residual Learning is not limited to image recognition; it also affects other areas of deep learning research. Its sophisticated approach to deep network training continues to be fundamental, spurring improvements in CNN topologies and model training methodologies.

The feasibility of employing digital food photographs for dietary evaluation among nutrition professionals is investigated in a study by Fatehah et al [3]. They evaluate this method's correctness and dependability through a thorough research. According to their results, analyzing digital food photos shows promise as a useful technique for dietary assessment that is effective and convenient. Nonetheless, issues including limitations in image quality and the accuracy of chunk size estimation are recognized. Notwithstanding these challenges, the study highlights the potential of digital food image evaluation as a useful tool for nutritionists to have in their toolbox to help with nutritional assessment and encourage better eating practices.

In order to classify food images,[4] this research presents a novel method that combines global structural information with local appearance. It uses shape context descriptors to encode the structural information of food objects and non-redundant local binary pattern (NRLBP) to accurately represent their looks. To combine these features for improved food image classification, two strategies are suggested. Experiments conducted on two datasets show notable gains in classification performance, confirming the effectiveness of combining structural and local appearance information. This triple contribution provides a promising path towards robust and reliable food picture categorization, which could find use in food identification systems and dietary evaluation.

The deep convolutional neural network architecture "Inception" [5], which made headlines in 2014 for its cutting-edge results in the ImageNet Large-Scale Visual Recognition Challenge, is described in the abstract. Its creative design emphasizes multi-scale processing for better classification and detection tasks, optimizing computational resources by increasing depth and width while keeping a consistent computational budget.

Big data analysis is used in this study [6] to improve food packaging optimization. Fuzzy directivity classification is used to extract packaging parameters by adaptive fuzzy matching, and the best packaging techniques for different goods are chosen. When paired with food attributes, directional clustering maximizes packaging choice, enhancing classification precision and encouraging the creation of clever food packaging.

Wei et al [8] use PyTorch and a Predictive REcurrent Neural Network (PReNet) to pioneer food-specific image recognition. Its convolutional neural network (CNN) architecture outperforms conventional models in terms of food classification accuracy. It enhances scholarly debate and finds applications in nutrition and health by utilizing a rich food dataset. The study promotes future developments in the field by showcasing the possibilities of PyTorch and PReNet.

In light of the COVID-19 pandemic and the increasing demand for real-time food and non-food classification, the research by Salma et al [9] introduces a lightweight Convolutional Neural Network (CNN) designed for effective classification. It solves computational complexity issues and exceeds earlier approaches with an impressive accuracy of 96.875%. This contribution highlights the significance of accessible solutions in contemporary global contexts and holds great potential for applications needing quick and accurate classification.

The research by Bu et al [10] proposes a solution based on transfer and ensemble learning to address issues in food image recognition. When compared to individual models, it achieves greater accuracy (96.88%) by utilizing pre-trained CNN models and fine-tuning on food datasets. Initially, convolutional neural network models (VGG19, ResNet50, MobileNet V2, AlexNet) pre-trained on the ImageNet dataset were used to extract generic picture features. Second, the food picture dataset was used to fine-tune the four previously trained models. Ultimately, many fundamental learner combination techniques were employed to create the ensemble model and categorize feature data. The outcomes show how effective ensemble learning is at improving recognition accuracy, highlighting the method's potential for real-world use in recommendation systems, food retrieval, and nutrition monitoring.

The study by Nanni et al [11] introduces a versatile computer vision system employing deep CNN features fused with traditional hand-crafted features. It offers three substructures for various image classification tasks: remapping CNN output, using penultimate layer features, and merging deep layer outputs. Feature transform techniques enhance dimensionality reduction. Tested across diverse datasets, the system exhibits robust performance, supported by statistical analysis and method independence assessment, validating its generalizability.

3. METHODOLOGY

Our proposed MobileNet architecture that we suggested is as follows. Initially, we employed the MobileNet V2 architecture pre-trained model. Three layers from the original network—the completely linked, average pooling, and softmax layers—were chosen to be eliminated. Second, three more layers are attached: the batch normalization (BN), the softmax, and the global average pooling (GAP) layers. Our suggested MobileNet architecture's primary goal is to accelerate the network's training process and increase accuracy. The dropout technique is then suggested as a way to stop overfitting. Overfitting is a common issue when a model learns to represent the noise in the training data instead of the underlying patterns. As a result, the model works effectively with training data but not well with fresh, untested data. Overfitting may be prevented or reduced using a number of strategies:

1) Cross-validation: By dividing the data into many train-validation-test sets, it is possible to assess how well the model performs when applied to new data and identify instances of overfitting.

2) Regularization: By introducing a penalty term to the loss function, techniques like L1 and L2 regularization punish overly complicated models and promote improved model generalization.

3) Feature Selection: You may lessen the complexity of the model and lessen overfitting by choosing just the most pertinent characteristics and eliminating any unnecessary or duplicate features.

4) Model simplification: Reducing the likelihood of overfitting can be achieved by using simpler models with fewer parameters, such as linear models or decision trees with restricted depth.

Additionally, the network trains more quickly thanks to the batch normalization layer. The rectified linear unit (ReLU), an activation function, is calculated in between the dropout layer and the batch normalization layer. ReLU provides assistance in resolving the vanishing gradients issue that may arise while deep neural networks are being trained. When gradients become incredibly tiny during training and propagate backward through the layers of the network, it is known as the vanishing gradients issue. This has the potential to impede learning, particularly in deep networks. By offering a straightforward, computationally effective activation function that prevents saturation—that is, the issue of gradients being too small—ReLU helps to reduce this problem. If the input value is positive, the ReLU activation function simply outputs that value; if not, it produces zero. In terms of math, ReLU(x) = max(0, x). This indicates that ReLU sets negative values to zero while interpreting the positive portion of its input. In doing so, it adds sparsity to the network, which can help with regularization and feature selection, enhancing the model's capacity for generalization.

A method for creating new training picture data that is related to the same image is called data augmentation. When numerous data augmentation techniques are used to solve image identification problems, accuracy performance is increased. These techniques include rotation, flipping, horizontal and vertical shifting, width shifts, and height shifts. The data augmentation methods used in this paper's studies include rescaling, rotation, horizontal flip, height shift, width shift, shear, and zoom. In addition, based on the range of the parameters, the image is randomly altered to produce a new image in every training session.

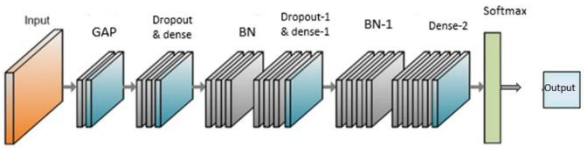
Moreover, cropping at random is used. This procedure, which is illustrated in Figure 5, starts with random point positions (x, y), then automatically crops and resizes to the desired size. The image used in this experiment has 224 by 224 pixel dimensions.

Fig -1: Sample food images from the dataset

The architecture has following layers -

Global Average Pooling (GAP) -

The primary reason for including Global Average Pooling (GAP) in the base model is that it prepares it for the final classification layer by calculating the average output of each feature map of the layer that comes before it. This layer, which lacks any trainable parameters like Max Pooling, significantly aids in data reduction. GAP, an indicator of overfitting, aids in the stability of validation accuracy. Thus, GAP in conjunction with the basic model aids in lowering the model's overfitting as



well as the CNN model's total computation time.

Fig -2: The proposed architecture of MobileNet V2

Regularization (Dropout) -

Dropout is a well-liked and effective regularization method, in conjunction with L2 and L1 regularization. By adding a penalty term to the loss function that is dependent on the absolute value of the coefficients, L1 regularization, also known as Lasso regression, encourages sparsity and performs automated feature selection, which helps to minimize overfitting and enhances the model's generalization performance. A penalty component depending on the squared magnitude of the coefficients is added to the loss function using L2 regularization, also known as ridge regression. By discouraging overfitting and promoting simpler solutions with smaller coefficient values, this penalty helps the model perform better when it comes to generalization. However, during CNN training, dropout simply disables certain neurons with a probability of P. A dropout of 0.5 or 0.25 is often employed in most CNNs. Half of the neurons are dormant and not included in CNN when P = 0.5. P = 0.25 indicates that 25% of neurons are dormant. Dropout reduces the complexity of the neural network and aids in preventing the CNN model from overfitting.

Dense Layer -

In a model, every neuron in the layer above the dense layer contributes to the output of the dense layer neurons, which multiply matrix-vectors. The process known as matrix vector multiplication occurs when the column vector of the dense layer and the row vector of the output from the previous layers are equal. When multiplying two vectors in a matrix, the row vector needs to have an equal number of columns as the column vector. The dense layer is often referred to as a completely linked layer, a basic neural network construction component. The phrase "dense" refers to a layer in which every neuron, or node, gets input from every other neuron in the layer above. This indicates that every neuron in the layer is linked to every other neuron in the layer above it.

Batch Normalisation (BN) -

The idea behind batch normalization is to segregate the input data into discrete groups, or batches, and process each batch in parallel using a normalization layer that is applied in-between each pair of network layers. The primary purpose of this approach is to enhance deep neural network training.Addressing the issue of internal covariate shift is one of Batch Normalisation's main goals. The phenomenon known

as "internal covariate shift" describes how, during training, the parameters of previous layers' parameters affect the distribution of layer inputs. The network's convergence may be hampered by this instability. BatchNorm reduces internal covariate shift by normalizing the inputs, which results in training dynamics that are more stable and smooth.

Softmax -

Softmax is a mathematical function that may be used to take an input vector of K real values and output another vector of K real values that adds up to 1. Its main objective is to translate logits or raw scores into probabilities. These probabilities represent the likelihood or degree of confidence that each class is correctly classified. The softmax function accepts input values from any range, including zero, positive, negative, and larger than one.Nonetheless, these values are converted into probabilities that fall between 0 and 1 once the softmax function is applied.To be more precise, the softmax function exponentiates every input value. It then divides each exponentiated value by the total of all exponentiated values to normalize the output values.

4. RESULT

The development and evaluation of our model have shown promising outcomes in terms of correctly identifying the input image. With little loss of accuracy, the suggested model works well on the food image dataset. The improved model has an accuracy of 97.16%. Below are the figures showing Accuracy and Loss against the number of epochs during training and validation phases for enhanced MobileNet V2. The model takes an image as the input, identifies the class of the input image and then classifies the output as healthy or unhealthy.

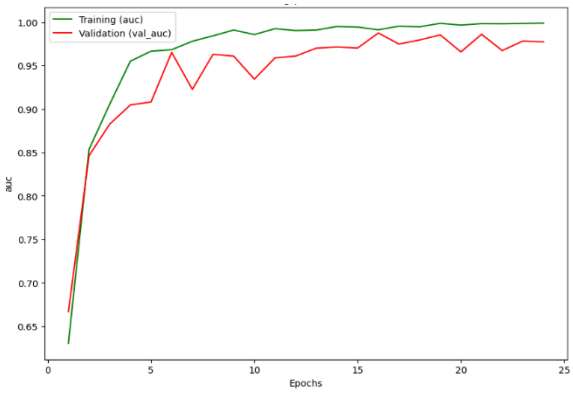


Fig -3: Accuracy during training and validation

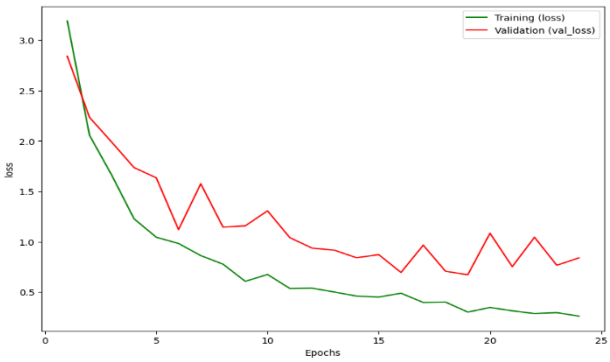


Fig -4: Loss during training and validation

From the figure 5, the model assesses the input image during the categorization process, identifying its features and attributes to place it in the most appropriate classifications. The model precisely analyzes the image to extract relevant patterns and properties that describe the image's content. It determines the three most pertinent classes that best capture the visual components shown in the image through this thorough examination. Each of these candidate classes is given a probability score, which represents the model's confidence in its classification, in order to decide the final classification. These probability scores provide important information about the degree of certainty the model has for each classification candidate by encapsulating the chance that the input picture belongs to each relevant class. On calculating these probability scores, the model selects the class with the highest probability as the final classification for the input image. This class is deemed the most representative and indicative of the content portrayed in the image, serving as the final classified label assigned to the input.

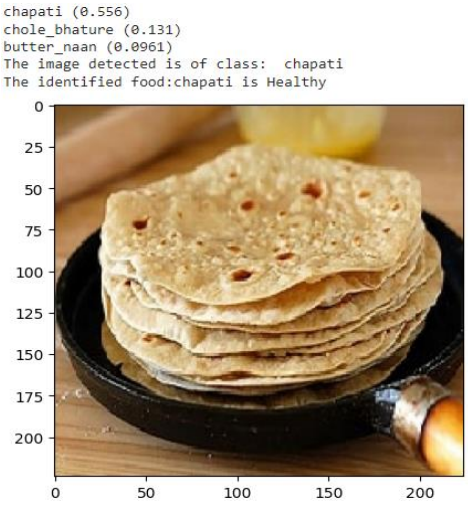


Fig -5: Identified image class (Chapati) and category (Healthy)

Accuracy	Training	Validation
Previous model	94.59	96.86
Enhanced model	99.88	97.16

Table -1: Accuracy comparison

As the table above illustrates, comparing the performance of the improved model to the prior model offers an overview of the accuracy about the improvements made.

Making use of the homemade recipe collection greatly expands the breadth and depth of our research findings. Through the integration of various culinary settings included in the dataset, we are able to validate the effectiveness and dependability of our model in a wide range of cooking situations. This thorough evaluation provides solid confirmation of our model's functionality and performance, as well as bolstering the validity of the study findings. The knowledge gained from this extensive testing phase will be extremely helpful in the future for more improvements and adjustments. These improvements are essential for maximizing our model's performance and guaranteeing its smooth integration and applicability in actual cooking settings. By continuously enhancing and refining our model using the

insights gained from this evaluation, we open the door to its wider acceptance and more significant use in real-world culinary contexts. Below image depicts the outcome for the same –

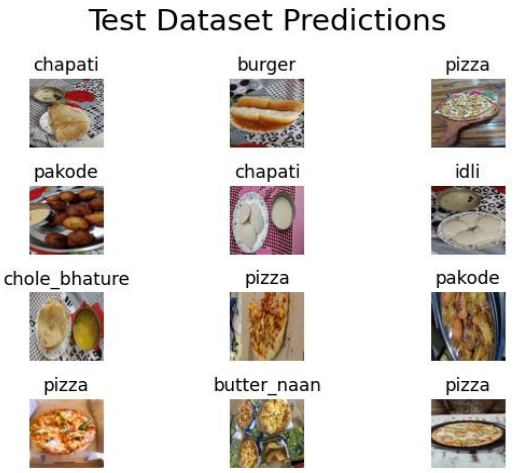


Fig -4: Testing the model on homemade recipes

5. CONCLUSIONS

In this work, we used the enhanced MobileNet V2 architecture to the dataset of food images. We presented the enhanced architecture of MobileNet V2 which helps to solve the overfitting issue. By implementing the global average pooling (GAP) layers, the number of parameters in this suggested MobileNet design is reduced. Additionally, the rectified linear unit (ReLU), dropout layers, and batch normalization (BN) layers are integrated. The softmax is the final layer which calculates the probabilities of possible output classes and the one with the maximum probability is the final recognized class. This architecture gave an accuracy of 97.16%.

We intend to build deep ensemble convolutional neural network (CNN) architectures in the future. These architectures are a synthesis of the most advanced deep CNN architectures. Since the feature vector extracted from the convolutional layers may perform better than the deep neural network design used individually, this is of interest to us. Also, a system can be generated to accumulate almost all the existing types of food classes to experiment the model's accuracy on a computationally very large dataset.

ACKNOWLEDGEMENT

I would like to thank my guide, all of the faculty members and anonymous reviewers for their advice and recommendations, without which I could not have finished this task. Additionally, I would like to thank all of the researchers whose work was made publicly available and which proved to be an invaluable source of information for our effort.

REFERENCES

1. Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. CVPR 2018
2. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. CoRR, abs/1512.03385, 2015. 1, 3, 4, 8
3. Fatehah, A., Poh, B., Shanita, S., and Wong, J. 2018. Feasibility of Reviewing Digital Food Images for Dietary Assessment among Nutrition Professionals. Nutrients 10, 8 (July 2018).
4. Nguyen, D., Zong, Z., Ogunbona, P., Probst, Y., and Li, W. 2014. Food image classification using local appearance and global structural information. Neurocomputing. 140, 242– 251.
5. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pages 1–9. IEEE Computer Society, 2015.
6. Tian Wei, Song Xiangbo. Selection of Optimal Packaging Methods for Different Food Based on Big Data Analysis. 2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS)
7. Kaggle dataset - https://www.kaggle.com/datasets/l33tc0d3r/indian-food-classification
8. Wei Zuo, Weiwei Zhang, Zengchao Ren.Food Recognition and Classification Based on Image Recognition: A Study Utilizing PyTorch and PReNet. 2023 IEEE International Conference on Image Processing and Computer Applications (ICIPCA)
9. Salma Zakzouk; Aya Saafan; Menna-allah Sayed; Mustafa A. Elattar; M. Saeed Darweesh. Light-Weight Food/Non-Food Classifier for Real-Time Applications. 2022 4th Novel Intelligent and Leading Emerging Sciences Conference (NILES)
10. Le Bu,Caiping Hu ,Xiuliang Zhang. Recognition of food images based on transfer learning and ensemble learning. https://doi.org/10.1371/journal.pone.0296789
3. Nanni L., Ghidoni S., and Brahnam, S. 2017. Handcrafted vs. non-handcrafted features for computer vision classification. Pattern Recognition. 71, 158–172.