A Comparative Study of CNN-based Approaches for Food Identification

Kaarthik Shrinivas V
Vellore Institute of Technology
Chennai, India
kaarthikofficial02@gmail.com

Braveen M
Vellore Institute of Technology
Chennai, India
braveen.m@vit.ac.in

Abstract— **The increasing concern over food allergies, food-related diseases, and food fraud has led to a growing interest in the development of accurate and efficient food identification systems. In this paper, we present a comprehensive comparative study of various Convolutional Neural Network (CNN)-based approaches for food identification. We compared 3-hidden layered CNNs, Resnet CNN, and transfer learning MobileNet CNNs using a dataset of food images that included a small variety of food types. Our results show that all of the CNN-based approaches we tested were able to achieve good levels of accuracy for food identification with such a small dataset. The transfer learning CNN approach, which uses a combination of pre-trained model CNN, performed the best, achieving an accuracy of 74.04%.**

Keywords— **CNN, Transfer Learning, ResNet architecture, MobileNet, Pre-trained model**

I. Introduction

Food identification is an essential task that has been gaining increasing attention in recent years. With the rise of food allergies, food-related diseases, and food fraud, the need for accurate and efficient food identification systems is more important than ever. One of the most promising approaches for food identification is the use of convolutional neural networks (CNNs). CNNs are a type of deep learning algorithm that are particularly well-suited for image classification tasks.

In this paper, we propose a food identification system using CNNs. The system is designed to accurately identify food items from images, and it is capable of identifying a small range of different food items. The system is based on a deep CNN that is trained using a small dataset of food images. The dataset includes a small variety of food items, such as dosa, bread, puri, and chapati, etc. Various systems are evaluated using accuracy. The results of the evaluation show that the transfer learning system is able to achieve good levels of accuracy for food identification for such a small dataset. Furthermore, the system is able to identify food items with high precision and recall, which is crucial for ensuring that food is properly identified, particularly in cases where food allergies or food-related diseases are a concern.

In summary, this paper compares various food identification systems using CNNs that are capable of accurately identifying a small range of different food items. The system is based on a deep CNN that is trained using a small dataset of food images, and it is evaluated using accuracy. The results of the evaluation show that the systems are able to achieve good levels of accuracy for food identification and can identify food items with high precision and recall. This system can be applied to a variety of different scenarios and can play an essential role in addressing food allergies, food-related diseases, and food fraud.

II. Background

CNNs are a type of deep learning algorithm that are modeled after the structure of the human visual system. They consist of multiple layers of artificial neurons, each of which performs a specific task, such as feature extraction or classification. The layers are connected in a hierarchical manner, with each layer building upon the features learned by the previous layer. This hierarchical structure allows

CNNs to learn increasingly complex features as the data flows through the network.

A convolutional neural network (CNN) is a type of deep learning neural network that is commonly used for image and video processing tasks. CNNs are designed to automatically and adaptively learn spatial hierarchies of features from input data.

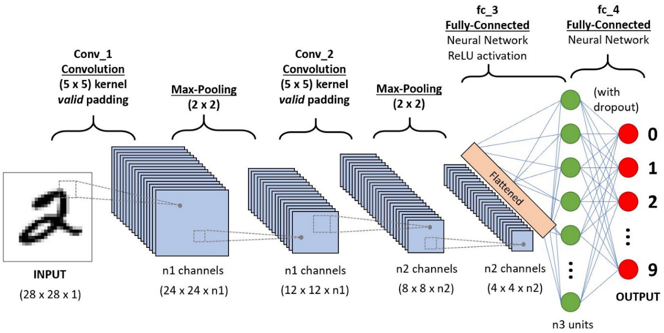
A CNN consists of an input and an output layer, as well as multiple hidden layers. The hidden layers of a CNN typically consist of convolutional layers, pooling layers, and normalization layers.

- Convolutional layers are responsible for extracting features from the input image, using a set of filters that scan the image and apply mathematical operations to the image pixels.
- Pooling layers are responsible for down-sampling the feature maps produced by the convolutional layers, in order to reduce their dimensionality and increase the robustness of the features to small translations of the input image.
- Normalization layers are responsible for normalizing the output of the previous layers, in order to reduce the internal covariate shift and improve the stability of the learning process.

The output of the last hidden layers are then passed through fully connected layers that perform the final classification or regression.

CNNs have been successful in various image classification tasks such as object recognition, facial expression recognition, facial recognition, and so on. They are also used in other applications such as natural language processing, speech recognition and many other areas of computer vision.

CNNs have been shown to be highly effective for image classification tasks, and have been used for a wide range of applications, including object detection, image segmentation, and facial recognition. In the context of food identification, CNNs can be used to extract features from food images, such as color, texture, and shape, and use these features to classify the food items.



The use of CNNs for food identification has been widely studied in recent years, and a variety of different CNN-based approaches have been proposed. These approaches include single-model CNNs, ResNet CNNs, and transfer learning CNNs. Single-model CNNs use a single CNN to classify food images. Ensemble CNNs use a combination of multiple CNNs to classify food images, and transfer learning CNNs use a pre-trained CNN as a starting point and fine-tune it on a food identification dataset.

The use of these various CNN-based approaches has shown promising results for food identification, but a comprehensive comparison of the different approaches has not been conducted. In this paper, we aim to fill this gap by presenting a comparative study of various CNN-based approaches for food identification. This study will provide a thorough evaluation of the performance of different CNN-based approaches and will inform the choice of the best approach for different applications.

III. RELATED WORKS

In this section, few of the previous studies on machine learning-based Image classification methods are briefly discussed.

Reddy et al [1] proposes a food recognition system that uses image processing and a convolutional neural network (CNN) to classify food items and estimate their calorie content. The system first segments the food images using image processing techniques, and then uses a CNN to classify the food items. The calorie content of the food items is estimated using a separate calorie estimation model. This paper does not compare various methods of CNN that are used for image classification and this model is trained on a huge dataset and does not perform well in small dataset.

Kagaya et al [2] used traditional image classification methods, SVM, and CNN to identify and classify the food into 10 different categories and concluded that CNN is best for classification. This paper gives us a comparison between traditional classification algorithms such as SVM and CNN but doesnot incorporate different architectures of CNN.

Thai et al [3] proposes a method where they separate the image into many sub-images based on the features of images. Each sub-image is classified by an ANN. Finally, SVM has been used to compile all the classified results of ANN. The classification model has brought together many ANN and one SVM. This paper gives us a new method on how to combine an ANN and SVM to classify images. But the use of CNN gave better accuracy on the same dataset.

Gontumukkala et al [4] proposes a method where they use SVM to classify Images and they are trying to understand SVM and then understand how to draw a decision boundary and try to make it optimal and use it for classification. Two Datasets “Dogs Vs Cats” and “Color Classification” are used in this paper. This paper gives a detailed explanation on how SVM is used for classification of images.

Kim et al [5] used a general Bag of Words model in order to compare two different classification methods. Both K-Nearest-Neighbor (KNN) and Support-Vector-Machine (SVM) classification are well known and widely used. We were able to observe that the SVM classifier outperformed the KNN classifier. This paper gave us an understanding that SVM classifiers are better at classification than KNN.

Chandra et al [6] performed a detailed survey on SVM and how it is used in classification and regression problems. SVM has the strongest mathematical model for classification and regression. This paper reviews the different computational models of SVM and key processes for the SVM system development.

Ragusa et al [7] compared various image classification especially for binary classification problems namely (Food vs Non-Food Images) and compared them using parameters such as Accuracy, True Positive Rate (TPR), True Negative (TNR). They compared various methods such as CNN, SVM, Softmax Classifier, AlexNet CNN. This paper concludes that Binary SVM and fine-tuned AlexNet model gives the best results for binary classification of images.

Abu et al [8] conducted a study on image classification by using the deep neural network (DNN). They used 5 categories of flowers and used them in different sizes of MobileNet and obtained the results. They observed that smaller models take less time to train and have slightly less accuracy compared to bigger models.

Aguilar et al [9] aims to solve the overfitting problem that occurs when CNNs are trained over a small dataset. They aim to solve this problem by using fusion classifiers where they combine 2 or more CNN outputs to obtain the results. They concluded that Fusion of CNN can give better results when less data is available but Fusion of more CNN models can also lead to Overfitting problems.

Rastegari et al [10] uses face and non-face images for classification. They propose two efficient approximations to standard convolutional neural networks: Binary-Weight-Networks and XNOR-Networks. In Binary-Weight-Networks, the filters are approximated with binary values resulting in 32× memory saving. In XNOR-Networks, both the filters and the input to convolutional layers are binary. XNOR-Networks approximate convolutions using primarily binary operations. This paper focuses on memory reduction as CNN requires high memory to attain high accuracy.

Kamavisdar et al [11] compared various algorithms namely Decision Tree, Fuzzy Measures, SVM, ANN with each other and analysed their advantages and disadvantages. This paper provides us with a theoretical comparison between these image classification methods.

Pasolli et al [12] proposed a novel method for classification of Spatial images such as satellite images. They suggested combining spectral and spatial information directly in the iterative process of sample selection. We can even extend it to other images. They are using SVM to classify the images and do not include CNN.

Korytkowski et al [13] presents a novel approach to visual objects classification based on generating simple fuzzy classifiers using local image features to distinguish between one known class and other classes. It outperformed the bag of features method in terms of accuracy and speed.

Jiang et al [14] aims to classify noisy food images correctly and classify them using various CNN models and compare the CNN models with each other. They succeeded in classification of noisy images.

IV. DATASET

The dataset used in this comparative study is collected from google and the dataset contains a total of 90 images belonging to 9 different food items. These 90 images are divided into 63 training images and 27 testing images. The reason for a small dataset is due to limitations in processing power of the device and to check the performance of various CNN architectures in small datasets.

The images belong to various classes such as Bread, Chapati, Dosa, Ice cream, Idly, Maggi, Pizza, Pongal, Puri. We have included images such that some images from different classes look almost the same so that it will be difficult for CNNs to classify them. These images also include the external background details such as the plate the dish is served in, table, etc.

V. COMPARISON

Now we can learn about the various architectures of CNN that are compared to each other.

A. 3-hidden layered CNN

This CNN contains 5 layers namely 1 input layer, 3 hidden layers that contain convolutional and pooling layers, 1 Output layer that contains a flatten layer and dense layer to classify them into classes.

The architecture is a feedforward convolutional neural network with the following layers:

- A convolutional layer (Conv2D) with 32 filters, a kernel size of 3x3, and a ReLU activation function. The output shape is (254, 254, 32)
- A max pooling layer (MaxPooling2D) with a pool size of 2x2. The output shape is (127, 127, 32)
- A convolutional layer (Conv2D) with 64 filters, a kernel size of 3x3, and a ReLU activation function. The output shape is (125, 125, 64)
- A max pooling layer (MaxPooling2D) with a pool size of 2x2. The output shape is (62, 62, 64)
- A convolutional layer (Conv2D) with 128 filters, a kernel size of 3x3, and a ReLU activation function. The output shape is (60, 60, 128)
- A max pooling layer (MaxPooling2D) with a pool size of 2x2. The output shape is (30, 30, 128)
- A flatten layer (Flatten) to reshape the 3D output from the previous max pooling layer into a 1D array. The output shape is (None, 115200)
- A dense layer (Dense) with 512 units and a ReLU activation function.
- A dense layer (Dense) with 9 units and a softmax activation function. This is the output layer, with 9 neurons corresponding to 9 classes.

This model gave us a testing accuracy of 40.74%.

B. 3-hidden layered CNN trained on edges of images

This CNN contains 5 layers namely 1 input layer, 3 hidden layers that contain convolutional and pooling layers, 1 Output layer that contains a flatten layer and dense layer to classify them into classes. This CNN architecture is the same as the architecture of CNN (A) . The difference is that this CNN is trained on edges of the images in the dataset. We are able to get an testing accuracy of 18.52%.

C. ResNet

ResNet50 is a deep convolutional neural network architecture that won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2015. The architecture is characterized by its deep residual connections, which allow the network to learn more complex representations of the input data. ResNet50 is widely used in various computer vision tasks and is a commonly used architecture for transfer learning.

The architecture of ResNet50 is made up of several convolutional and pooling layers, followed by fully connected layers.

The core of the architecture is the residual block, which contains two or more layers of convolutional filters. These blocks are connected by "skip connections" that add the input of a block to its output, allowing the network to learn residual functions rather than the original functions. This allows the network to learn much deeper representations of the input data and helps to alleviate the vanishing gradient problem.

The ResNet50 architecture consists of several of these residual blocks, arranged into four groups. The first group contains one convolutional layer and one max pooling layer, followed by four groups of residual blocks. Each of these groups contains several residual blocks, with the number of blocks increasing as the network progresses towards the output layer.

The final layers of the network are fully connected layers, which are used to make the final predictions. The output of the final fully connected layer is fed into a softmax activation function, which produces the final probability scores for each class.

Overall, ResNet50 is a deep network with 50 layers. The architecture is designed to take advantage of the residual connections to enable training of deep networks.

We are able to get a testing accuracy of 11.11% due to a small dataset.

D. MobileNet_v2

MobileNet_v2 is a deep convolutional neural network architecture that is designed to be efficient and lightweight, making it suitable for mobile and embedded devices. It was developed by Google and introduced in 2018 as a successor to the original MobileNet architecture.

Like the original MobileNet, MobileNet_v2 uses depth wise separable convolutions, which factorize a standard convolution into a depthwise convolution and a 1x1 pointwise convolution. This reduces the number of parameters and computation required, making the network more efficient.

The architecture of MobileNet_v2 also includes several other optimizations, such as linear bottlenecks between the layers and inverted residual blocks, which further reduce the computation required. MobileNet_v2 also includes a feature called "linear bottleneck", which improves the model accuracy by decreasing the number of channels in the bottleneck layer.

MobileNet_v2 is pre-trained on the ImageNet dataset and can be easily fine-tuned for a variety of image classification and object detection tasks. Due to its lightweight architecture, it's often used as a base model for real-time object detection in mobile and embedded devices.

We are able to get a testing accuracy of 74.04% after fine-tuning the model for our dataset.

E. *MobileNet_v2 trained on the edges extracted from the images.*

The MobileNet_v2 used is a pre-trained model on the ImageNet dataset, and we trained this model on the edges extracted from the images in our dataset.

The edges are extracted using the Soble edge detection algorithm. We are able to attain a testing accuracy of 22.22%.

F. *Combined model of MobileNet_v2 and ResNet50*

Two different models are combined using ensemble methods.

The architecture has 4 layers:

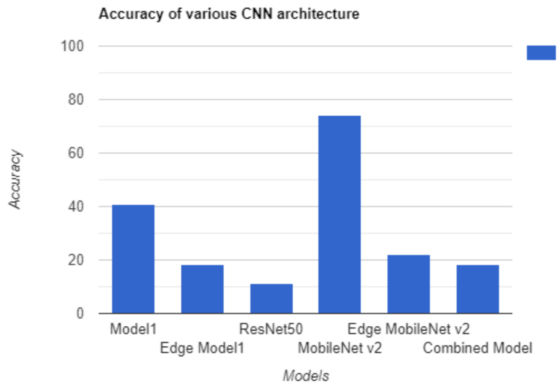
- input_8: This is an Input layer, it serves as the entry point for the data into the model. It defines the shape of the input data, in this case, it's a tensor of shape (None, 224, 224, 3) which means it's a batch of images with 224x224 pixels and 3 color channels (RGB)
- model: this layer is a function call, MobileNet. The output shape is (None, 9) which means the model is outputting 9 class scores.
- resnet50: this is another function call of a pre-trained model, in this case, it's ResNet50. The output shape is also (None, 9)
- concatenate_2: This is a concatenate layer, it concatenates the output from the model and resnet50 along the last axis (axis=-1). The output shape is (None, 18) which means it's combining the 9 class scores from model and resnet50
- dense_a: This is a Dense (Fully connected) layer, it has 9 units and a softmax activation function. It's connected to the concatenate_2 layer, taking the concatenated 18 class scores and outputting the final 9 class scores which will be used as the final predictions of the model.

We obtained a testing accuracy of 18.52%.

These are the various architectures of CNN that we have compared using a small dataset. We have performed edge detection to see how the results change if we focus only on edges of the images. The below bar graph compares the accuracies of the various models of CNN.

Accuracy - Accuracy is found by dividing the correctly predicted samples by the total number of samples.

$$\text{accuracy} = \frac{\text{correct predictions}}{\text{all predictions}}$$



VI. CONCLUSION

In this study, we evaluated the performance of various CNN architectures and methods commonly used for image classification using a small dataset. The aim was to determine the effectiveness of these CNNs when dealing with limited data. As we can observe, the transfer learning model (MobileNet_v2) that is pre-trained on the ImageNet dataset performed the best for a small dataset. As we can expect, only edges don’t give enough information for CNN to classify they performed the least compared to those where the image is given. ResNet50 was not able to perform as expected as it requires a large dataset to predict the image accurately as it has many layers. Hence we can conclude that transfer learning models are the best when we have a limited dataset. The future works are discussed in the below section.

VII. FUTURE WORKS

Future research will expand the comparison to include additional CNN architectures, as well as evaluate the models with a larger dataset. Additionally, this study only trained the models on original images and their edges, future studies can investigate the effect of various image modifications, such as converting to grayscale or cropping to only include the food, on model performance.

REFERENCES

[1] Reddy, V. H., Kumari, S., Muralidharan, V., Gigoo, K., & Thakare, B. S. (2019, May). Food Recognition and Calorie Measurement using Image Processing and Convolutional Neural Network. In 2019 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT) (pp. 109-115). IEEE.

[2] Kagaya, H., Aizawa, K., & Ogawa, M. (2014, November). Food detection and recognition using convolutional neural network. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 1085-1088)

[3] Thai, L. H., Hai, T. S., & Thuy, N. T. (2012). Image classification using support vector machine and artificial neural network. International Journal of Information Technology and Computer Science, 4(5), 32-38.

[4] Gontumukkala, S. S. T., Godavarthi, Y. S. V., Gonugunta, B. R. R. T., Subramani, R., & Murali, K. (2021, July). Analysis of Image Classification using SVM. In 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 01-06). IEEE.

[5] Kim, J., Kim, B. S., & Savarese, S. (2012, January). Comparing image classification methods: K-nearest-neighbor and support-vector-machines. In Proceedings of the 6th WSEAS international conference on Computer Engineering and Applications, and Proceedings of the 2012 American conference on Applied Mathematics (pp. 133-138).

[6] Chandra, M. A., & Bedi, S. S. (2021). Survey on SVM and their application in image classification. International Journal of Information Technology, 13(5), 1-11

[7] Ragusa, F., Tomaselli, V., Furnari, A., Battiato, S., & Farinella, G. M. (2016, October). Food vs non-food classification. In Proceedings of the 2nd International workshop on multimedia assisted dietary management (pp. 77-81).

[8] Abu, M. A., Indra, N. H., Rahman, A. H. A., Sapiee, N. A., & Ahmad, I. (2019). A study on Image Classification based on Deep Learning and Tensorflow. International Journal of Engineering Research and Technology, 12(4), 563-569.

[9] Aguilar, E., Bolaños, M., & Radeva, P. (2017, September). Food recognition using fusion of classifiers based on CNNs. In International Conference on Image Analysis and Processing (pp. 213-224). Springer, Cham.

[10] Rastegari, M., Ordonez, V., Redmon, J., & Farhadi, A. (2016, October). Xnor-net: Imagenet classification using binary convolutional neural networks. In European conference on computer vision (pp. 525-542). Springer, Cham.

[11] Kamavisdar, P., Saluja, S., & Agrawal, S. (2013). A survey on image classification approaches and techniques. International Journal of Advanced Research in Computer and Communication Engineering, 2(1), 1005-1009.

[12] Pasolli, E., Melgani, F., Tuia, D., Pacifici, F., & Emery, W. J. (2013). SVM active learning approach for image classification using spatial information. IEEE Transactions on Geoscience and Remote Sensing, 52(4), 2217-2233.

[13] Korytkowski, M., Rutkowski, L., & Scherer, R. (2016). Fast image classification by boosting fuzzy classifiers. Information Sciences, 327, 175-182.

[14] Jiang, M. (2019). Food Image Classification with Convolutional Neural Networks. CS230: Deep Learning, Fall.