Food Image Classification Using Deep Learning

Author :
Lav Chaudhari (202101135) , Kush Patel (202101137) , Chaxu Patel (202101166)
Mentor :
Prof. Dr. Srimanta Mandal

***Abstract*—In recent years, the application of deep learning in image classification has shown significant promise across various domains, including food recognition. This report presents a comprehensive study on food image classification using deep learning techniques. The primary objective of this study is to develop an accurate and efficient model for classifying food images into their respective categories. We utilized a dataset named Food-101, which comprises a diverse collection of food images spanning a total of 101 distinct classes. The dataset was meticulously preprocessed to ensure the quality and consistency of the images. Several deep learning architectures were explored and implemented, including Convolutional Neural Networks (CNNs), to leverage their powerful feature extraction capabilities, along with transfer learning for continuous improvement in CNN performance. The implications of this study are significant for applications in the food industry, such as automated food recognition systems in restaurants.**

***Index Terms*—Deep Learning, Food image, Data augmentation, Convolutional Neural Networks (CNNs), Transfer learning, ResNet50, EfficientNetB0, InceptionNetV3, DenceNet201**

I. Introduction

IN recent years, the application of deep learning in image classification has shown significant promise across various domains, including food recognition. The advancements in computer vision and machine learning have made image classification more accessible and accurate due to the availability of vast datasets and powerful computational resources. This report presents a comprehensive study on food image classification using deep learning techniques, with a primary objective of developing an accurate and efficient model for categorizing food images.

Traditional image classification techniques such as K-Nearest Neighbors (KNN), Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), and Random Forests have been employed in the past. However, these methods often struggle when dealing with large datasets due to their limited scalability and efficiency. Convolutional Neural Networks (CNNs), on the other hand, have gained considerable attention for their ability to handle large amounts of data and provide high classification accuracy.

Training CNNs for image classification can be approached in two main ways: training the network from scratch or using transfer learning. Transfer learning is a deep learning technique where a pre-trained model, initially trained on a large dataset, is fine-tuned for a specific task. This approach is particularly beneficial when the available data for the new task is limited. In this study, transfer learning was employed by fine-tuning pre-trained models such as Inception V3, EfficientNetB0, DenceNet201 , and ResNet50, which are among the top-performing models in the annual ImageNet Large Scale Visual Recognition Challenge (ISLVRC). These models have demonstrated considerable accuracy and low validation loss, making them suitable candidates for food image classification.

The importance of food classification is underscored by rising global health concerns. According to the World Health Organization (WHO), more than 1.9 billion adults were overweight in 2016, with obesity rates doubling since 1980. Obesity is a major cause of chronic diseases such as diabetes, heart disease, high blood pressure, and certain types of cancer. Accurate food classification can aid in dietary monitoring for various demographics, including the elderly, patients with dietary restrictions, and fitness enthusiasts.Furthermore, from a marketing and economic perspective, food image classification can influence consumer choices and help in targeted advertising. For instance, if a restaurant owner wants to promote a particular dish, it is crucial that the advertisement reaches individuals inclined towards that type of food.

A. Relevant Works

Related Work

The emergence of deep learning has significantly advanced the field of image classification, particularly in handling large datasets. Numerous studies have applied these techniques to food image recognition, demonstrating substantial improvements over traditional methods.

Subhi and Ali [1] proposed a novel deep learning convolutional neural network (CNN) configuration for detecting and recognizing local food images. Using a dataset of Malaysian food images sourced from publicly available internet platforms, their study showed that CNNs outperform traditional image classification methods. The findings highlighted the importance of network depth in enhancing model performance, underscoring the potential of deep learning techniques in food image recognition.

Similarly, Hnoohom and Yuenyong [2] developed a prediction model to classify images of Thai fast food. Utilizing the Thai Fast Food (TFF) dataset and the Inception V3 model, they achieved an accuracy of 0.8833. However, the TFF dataset contained only 3960 images across 11 classes, which is relatively small for effective learning by CNNs according to best practices. Despite the limited dataset size, their work demonstrated the efficacy of deep learning models in food image classification.

Further advancements were made by Kawano and Yanai [3], [4], who explored methods to identify multiple food classes using deep convolutional neural networks (DCNNs) with data from the ImageNet database. In one approach, they fine-tuned a classifier pre-trained on a large-scale database using smaller food image datasets from the UECFOOD100 and UECFOOD256 databases. In another approach, they combined DCNN-learned features with conventional image features (such as RootHoG and RGB color values with 2x2 blocks) and trained the classifier using a one-vs-rest linear classification model. Their results indicated that integrating DCNN features with traditional image features could significantly boost classification performance.

These studies collectively illustrate the potential of deep learning models, particularly CNNs and DCNNs, in enhancing the accuracy and efficiency of food image classification. By leveraging pre-trained models and fine-tuning them for specific datasets, researchers have been able to achieve impressive results, highlighting the robustness and adaptability of deep learning techniques in this domain.

II. Proposed Approach

Proposed Methodology: Enhancing Food Image Classification Using Transfer Learning and Robust Model Architectures

Our approach to food image classification is structured to optimize performance through a carefully designed workflow. Beginning with a diverse dataset of food-101 images, our methodology emphasizes meticulous preprocessing and augmentation techniques to enhance both the quality and variability of the dataset. This initial step ensures that the models are trained on comprehensive and representative data.

The dataset is then divided into training and testing sets to facilitate model training and evaluation. Leveraging transfer learning, we capitalize on the knowledge stored in pretrained models like ResNet50, EfficientNetB0, InceptionNetV3, and DenseNet201. These models are adapted and fine-tuned specifically for our food recognition task, enabling us to achieve higher accuracy and efficiency in classification.

The selection of pretrained models is critical and is followed by a phase of further fine-tuning to optimize their performance on the food-101 dataset. This process involves selecting appropriate optimizers and loss functions tailored to the characteristics of our task. We adhere to best practices in batch processing and learning rate management during training to ensure stable convergence and effective learning.

Throughout the training and fine-tuning phases, continuous evaluation of the model's progress and performance is conducted. This iterative approach allows us to refine the models and address any challenges or limitations encountered during the training process.

The architecture of our food image classification system, depicted in Figure 1, illustrates the integration of preprocessing steps, model adaptation through transfer learning, and the iterative evaluation process. This holistic methodology aims to establish a robust and accurate food image classification system, combining technical sophistication with practical implementation considerations. By employing state-of-the-art pretrained models and rigorous training practices, we lay the groundwork for advanced applications in food recognition and related domains.

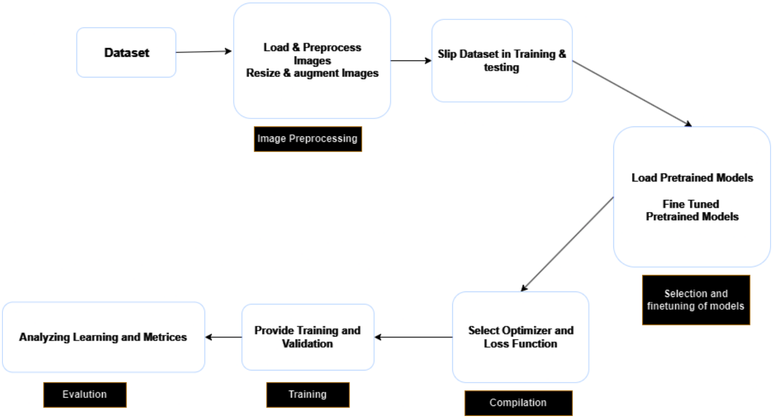
Dataset
Load & Preprocess Images
Resize & augment Images
Image Preprocessing
Slip Dataset in Training & testing
Load Pretrained Models
Fine Tuned Pretrained Models
Selection and finetuning of models
Select Optimizer and Loss Function
Compilation
Provide Training and Validation
Training
Analyzing Learning and Metrices
Evalution

Fig. 1. Visual Representation of the Food Image Classification Model

Fig. 2. Food101 dataset

A. Dataset Description

The Food-101 dataset is a widely used benchmark in the field of food image classification, consisting of 101,000 images across 101 food categories. Each category contains 1,000 high-resolution images sourced from various online platforms. The dataset is meticulously curated to ensure diversity and quality, with images typically resized to a uniform resolution (e.g., 512x512 pixels). In standard usage, the dataset is split into 75% training data (75,750 images) and 25% testing data (25,250 images), facilitating rigorous evaluation of classification models. This structured division helps in training robust models capable of accurately identifying diverse types of food items from digital images.

B. Data Preprocessing

The preprocessing of the Food-101 dataset begins with unzipping the archive, revealing a structured organization of 101 folders, each corresponding to a specific food category. Within each category folder, there are separate subfolders designated for training and testing data, facilitating clear delineation of data for model development and evaluation. All images are initially resized to a standardized format; for most pretrained models like ResNet50, EfficientNetB0, and DenseNet201, this involves resizing from the original 512x512 resolution to 224x224 pixels. However, for models such as Inception V3, which require a 299x299 pixel input, images are resized accordingly to meet the specific input requirements. Following resizing, a series of data augmentation techniques are applied to enhance the diversity and robustness of the training dataset. These techniques include zooming in and out to focus on different image details, adjusting brightness and contrast levels to simulate varying lighting conditions, adding random noise

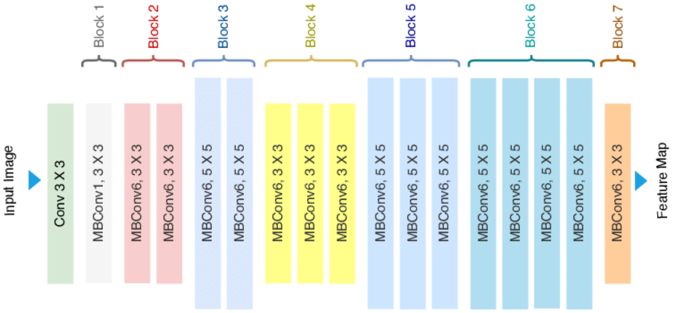

Fig. 3. Architecture of EffientNet-B0

to images to increase model resilience to noise, horizontally flipping images to augment training samples without altering class labels, and applying slight rotations to simulate different viewpoints. These preprocessing steps collectively ensure that the dataset is well-prepared to train deep learning models effectively for accurate food image classification, accommodating variations that may be encountered in real-world applications.After augmentation, the dataset is split into two parts with a ratio of 25:75 for testing and training, respectively. These preprocessing steps collectively ensure that the dataset is well-prepared to train deep learning models effectively for accurate food image classification, accommodating variations that may be encountered in real-world applications.

C. Pretrained Models for Food Image Classification

1) EffientNetB0: EfficientNetB0 is part of the EfficientNet family of models, known for their ability to achieve high accuracy with fewer parameters compared to traditional convolutional neural networks. The architecture is built upon a compound scaling method that uniformly scales all dimensions of depth, width, and resolution. EfficientNetB0 begins with a standard convolution layer followed by multiple mobile inverted bottleneck (MBConv) blocks, which are efficient and effective in capturing spatial features. These MBConv blocks include depthwise separable convolutions and squeeze-and-excitation optimization to enhance feature recalibration. The architecture concludes with a fully connected layer and a softmax layer for classification. EfficientNetB0 strikes a balance between accuracy and computational efficiency, making it well-suited for tasks like food image classification where both performance and resource usage are critical. The architecture of EfficientNet-B0 is shown in figure3.

2) InceptionNetV3: InceptionV3 is a deep convolutional neural network architecture known for its innovative use of inception modules, which allow the network to efficiently capture multi-scale features. The architecture consists of a series of these modules, each containing parallel convolutional layers with different kernel sizes (1x1, 3x3, and 5x5) and a pooling layer, followed by concatenation of their outputs. This design enables the model to extract rich and varied feature representations at each layer. InceptionV3 also employs factorized convolutions to reduce the number of parameters and computational cost, as well as auxiliary classifiers to improve convergence and combat vanishing gradients. The network culminates in a global average pooling layer, followed by a fully connected layer and a softmax layer for classification. With its efficient and effective design, InceptionV3 achieves high accuracy while maintaining manageable computational demands, making it a strong choice for image classification tasks, including food image classification.

3) ResNet50: ResNet50, or Residual Network with 50 layers, is a deep convolutional neural network architecture known for its revolutionary use of residual learning to address the vanishing gradient problem. The architecture is composed of an initial convolutional layer followed by a series of convolutional blocks, each containing multiple convolutional layers. The key innovation is the introduction of shortcut connections, or skip connections, which bypass one or more layers and directly add the input of one layer to the output of a subsequent layer. This enables the network to learn identity mappings, allowing gradients to flow more easily during backpropagation and facilitating the training of much deeper networks. ResNet50 consists of 48 convolutional layers along with 1 max pooling and 1 average pooling layer, ending with a fully connected layer and a softmax layer for classification. The architecture's depth, combined with its ability to maintain robust gradient propagation, makes ResNet50 highly effective for complex image classification tasks, including food image classification.

4) DenseNet201: DenseNet201, or Densely Connected Convolutional Network with 201 layers, is a deep learning architecture that introduces dense connections to enhance information flow between layers. In this architecture, each layer receives input from all preceding layers and passes its own feature maps to all subsequent layers, ensuring maximum information and gradient flow. This is achieved through dense blocks, where layers are connected in a feed-forward fashion, and transition layers, which consist of batch normalization, 1x1 convolution, and 2x2 average pooling. The network starts with a convolution and pooling layer, followed by multiple dense blocks and transition layers. This design results in efficient feature reuse, reducing the number of parameters while maintaining high performance. DenseNet201 culminates in a global average pooling layer, a fully connected layer, and a softmax layer for classification. The dense connectivity pattern of DenseNet201 makes it highly efficient and effective for complex image classification tasks, including food image classification, by promoting feature reuse and alleviating the vanishing gradient problem.

D. Model Trainind and Evalution

In our model training process, we first utilize different pretrained models including EfficientNetB0, DenseNet201, ResNet50, and InceptionV3. The images from the Food-101 dataset are initially resized to meet the input requirements of these models, with most being resized to 224x224 pixels and InceptionV3 images to 299x299 pixels. After resizing, data augmentation techniques such as zooming in and out, brightness and contrast adjustments, adding noise, horizontal flipping, and rotations are applied to increase the diversity and robustness of the training dataset.

Each pretrained model processes the input images and passes the output to a series of custom fully connected layers.

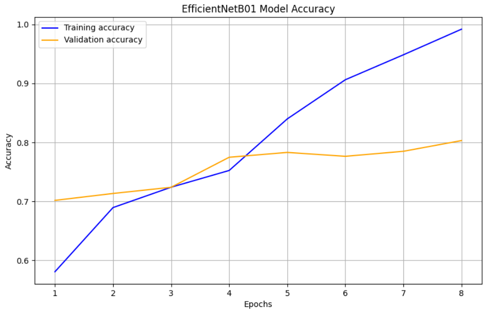

Fig. 4. EfficientNet-B0 training accuracy

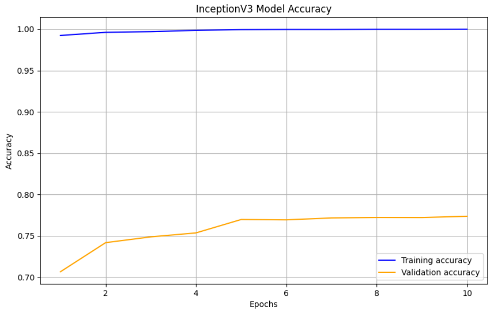

Fig. 5. InceptionV3 training accuracy

These custom layers include a hidden layer with a suitable number of neurons and an activation function like ReLU. The final output layer contains 101 nodes, corresponding to the 101 food categories, and uses a softmax activation function to generate probabilities for each class. The image is classified into the category with the highest probability.

We perform fine-tuning on the pretrained models by initially freezing the convolutional layers and training the custom fully connected layers. Gradually, the pretrained layers are unfrozen, and fine-tuning is applied to the entire network to adjust the weights and better fit the specific characteristics of the food images. Appropriate optimizers and loss functions are selected to ensure efficient training. This combination of pretrained models and fine-tuning enables us to leverage existing knowledge while adapting to the nuances of the Food-101 dataset, resulting in a robust and accurate food image classification system.

III. Results

We compile and train a deep learning model tailored for food image classification using TensorFlow. The model is configured with ‘sparse_categorical_crossentropy‘ as the loss function, optimized by Adam with a learning rate set to 0.001. Training extends over 10 to 30 epochs based on pre-trained model, with validation conducted on 15% of the test data per epoch.

The training duration of our models varies significantly based on the pretrained architecture selected. Training times range from approximately 3.5 to 7 hours, contingent on the specific model utilized. We will comprehensively compare the performance results achieved by the pretrained models: EfficientNetB0, DenseNet201, ResNet50, and InceptionV3. This comparative analysis will assess each model’s accuracy, validation loss, and computational efficiency. Such evaluation is crucial for determining the most suitable architecture for our food image classification task, balancing training time with performance metrics to achieve optimal results. This approach ensures a thorough examination of model capabilities and informs strategic decisions in deploying deep learning solutions for practical applications.

Per epoch, the training and validation accuracies for the pretrained models EfficientNetB0, DenseNet201, ResNet50, and InceptionV3 were monitored and recorded.

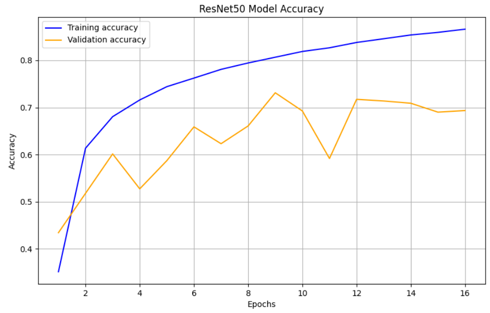

Fig. 6. ResNet50 training accuracy

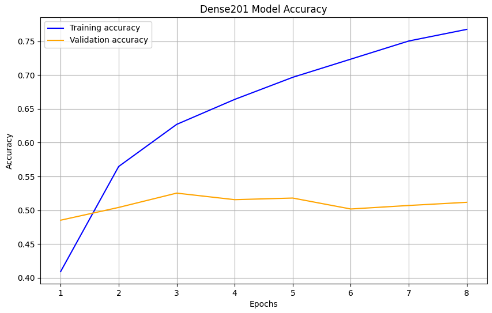

Fig. 7. DenseNet201 training accuracy

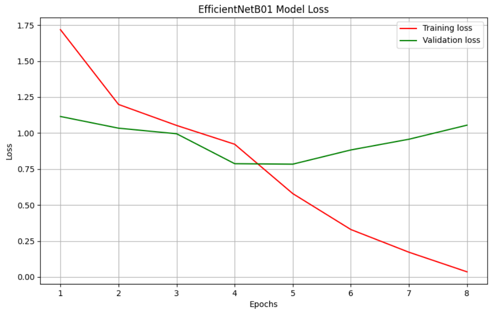

Fig. 8. EfficientNet-B0 training loss

IV. Conclusion

This study introduces advanced techniques leveraging transfer learning to fine-tune deep learning networks for food

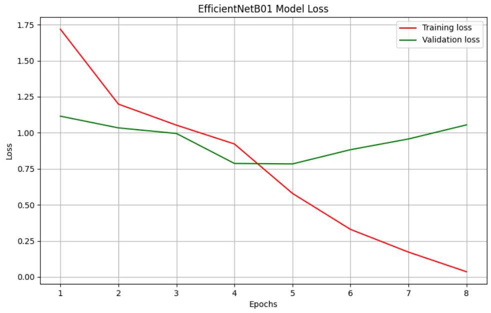

Fig. 9. InceptionV3 training loss

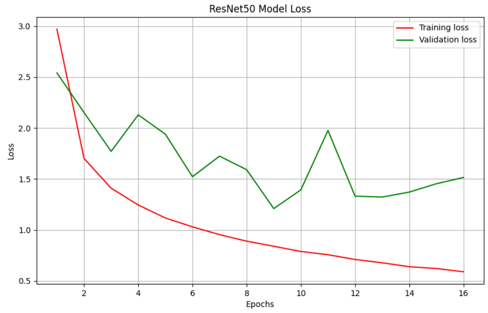

Fig. 10. ResNet50 training loss

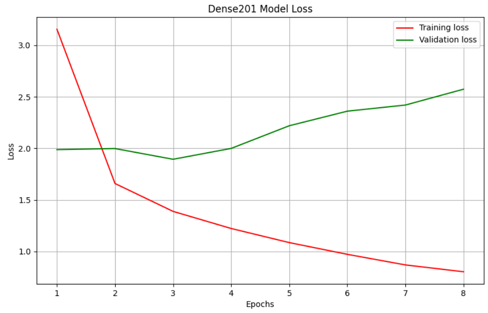

Fig. 11. DenseNet201 training loss

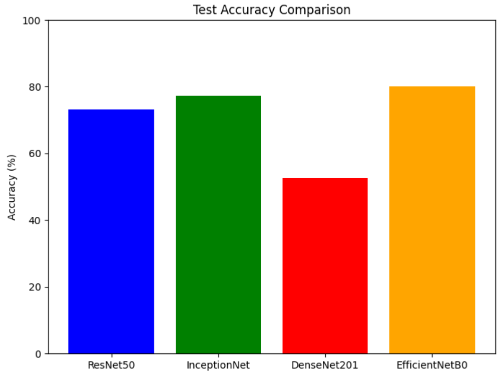

Fig. 12. Testing accuracy for different models

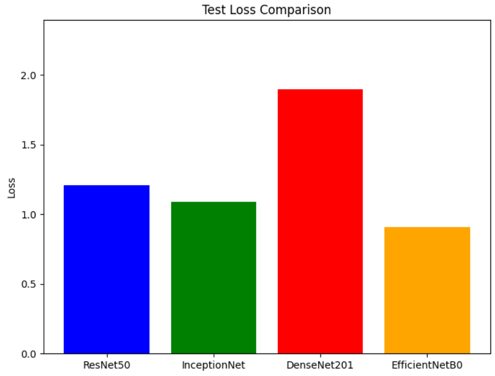

Fig. 13. Testing loss for different models

classification using the Food-101 dataset. Additionally, the dataset underwent augmentation by carefully selecting additional samples to increase class representation and enhance existing images, particularly addressing issues like distorted or crowded images that may obscure food details.

The augmented dataset significantly improved network training outcomes, highlighting the benefit of increased sample diversity in enhancing model performance. Looking ahead, future efforts aim to integrate these findings into a dedicated mobile application capable of real-time food recognition using smartphone cameras. Such an application would serve to assist visually impaired individuals and travelers navigating buffet-style dining environments.

V. REFERENCES

[1] A Deep Convolutional Neural Network for Food Detection and Recognition by Mohammed A. Subhi and Sawal Md.Ali

[2] Thai Fast Food Image Classification Using Deep Learning by Narit Hnoohom and Sumeth Yuenyong

[3] K. Yanai and Y. Kawano, "Food image recognition using deep convolutional network with pre-training and fine-tuning," in Multimedia Expo Workshops (ICMEW), 2015 IEEE International Conference on. IEEE, 2015, pp. 1–6.

[4] Y. Kawano and K. Yanai, "Food image recognition with deep convolutional features," in Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication. ACM, 2014, pp. 589–593.

[5] Food Classification from Images Using Convolutional Neural Networks David J. ttokaren, Ian G. Fernandes, A. Sriram, Y.V. Srinivasa Murthy, and Shashidhar G. Koolagudi

[6] Basrur, Ankit, Dhrumil Mehta, and Abhijit R. Joshi. "Food Recognition using Transfer Learning." 2022 IEEE Bombay Section Signature Conference (IBSSC). IEEE, 2022.