Transfer Learning for Food Image Classification With EfficientNet

F. B. Zaman
22-47256-1@student.aiub.edu

M. T. Hassan
22-46481-1@student.aiub.edu

MD N. Rahman
22-46521-1@student.aiub.edu

S. Ahmed
22-46486-1@student.aiub.edu

***Abstract*— The rising prevalence of chronic diseases necessitates innovative tools to promote healthy dietary habits. Food image classification, a powerful application within computer vision, offers significant potential by enabling automated identification of food items in digital images. This study explores the application of transfer learning with EfficientNet architectures (B4 to B7) for food image classification on the Food-101 dataset. Transfer learning leverages pre-trained models to enhance accuracy and reduce training time. The evaluation compares the performance of these models with and without data augmentation, a technique that artificially expands the training data. Interestingly, the results revealed that models without data augmentation generally achieved higher accuracy. This research contributes to the development of effective food image classification systems for diverse applications in health and wellness.**

Keywords*— *EfficientNet, Food-101, Food image classification

I. INTRODUCTION

The global health landscape undergoes a significant transformation. The rising prevalence of chronic diseases like diabetes and obesity necessitates maintaining a balanced and healthy diet as a critical public health concern [1]. Fortunately, technological advancements offer solutions to empower individuals in making informed dietary choices. One such innovation is food image classification, a powerful tool within computer vision that facilitates the automated identification and categorization of food items in digital images.

Food image classification offers a vast array of applications beyond mere identification. In the domain of health and wellness, analyses of food images predict user dietary patterns and potential health risks. This information is leveraged to develop personalized guidance and promote healthier eating habits [1]. Additionally, food image classification applications empower visually impaired individuals by providing real-time information about the food in front of them, enhancing their independence in daily activities [2].

This paper investigates the application of deep learning techniques for food image classification. The research focuses on a particularly effective approach known as transfer learning. This technique leverages pre-trained models on similar tasks, significantly reducing training time and resource requirements compared to training a model from scratch. The study explores the impact of transfer learning on the performance of various deep learning architectures for food image classification. It evaluates the effectiveness of EfficientNetB4 to EfficientNetB7 models, recognized for their scalable precision.

The Food-101 dataset [3] is employed to assess the performance of these architectures with transfer learning. This research aims to identify the most suitable model for food image classification tasks on the Food-101 dataset, considering factors like accuracy, efficiency, and computational cost.

The remainder of the paper is structured as follows. Section 2 presents a review of relevant research in food image classification and transfer learning. Section 3 details the employed methodology, including the chosen deep learning architectures and datasets. Section 4 presents the evaluation of the results, comparing the performance of each model with transfer learning. Finally, Section 5 offers concluding remarks and explores potential avenues for future research in this exciting field.

II. RELATED WORKS

Deep learning techniques, particularly convolutional neural networks (CNNs), have revolutionized food image classification. This review explores the findings from several research papers, highlighting the advancements in CNN architectures, transfer learning, and the integration with Natural Language Processing (NLP) for tasks like recipe extraction.

The study by [Bossard et al., 2014] examines the effectiveness of various CNN models. They observed that simpler models achieved moderate accuracy but were susceptible to overfitting. However, incorporating transfer learning with the InceptionV3 architecture significantly improved performance, reaching a validation accuracy of 89.67% on the Food-101 dataset [Bossard et al., 2014]. This finding emphasizes the growing effectiveness of pre-trained models and fine-tuning for handling large and diverse datasets.

Further advancements are explored in the work by [Tripathi, 2021], who showcase the superiority of DenseNet-161 for food image classification. This architecture achieved a top-5 accuracy of 99.01% on Food-101, surpassing previous models [Tripathi, 2021]. This underlines the importance of utilizing pre-trained models and transfer learning for enhanced classification performance. Another approach is presented by

[Aa et al., 2023], who investigate the synergy between CNNs and web crawlers for extracting food information from images. They employed a pre-trained Inception v3 model with data augmentation techniques, achieving an accuracy of 97.00% for 20 classes. This integration demonstrates the potential of combining image processing with web data extraction for comprehensive food analysis tools [Aa et al., 2023].

Moving beyond basic classification, [Abdul Kareem et al., 2024] explore a custom lightweight CNN model for fine-grained food image classification and recipe extraction. By integrating CNNs with NLP, they achieved significant accuracy improvements on challenging datasets, outperforming existing models like DeepFood and CNN-Food. This highlights the benefits of combining visual recognition and textual analysis for deeper understanding and enhanced accuracy in food image classification and related tasks [Abdul Kareem et al., 2024].

Finally, the work by [Rahman et al., 2024] delves into a customized deep learning network combined with NLP techniques. Their approach incorporates a modified ResNet-50 model with advanced NLP methods like Word2Vec and Transformers, leading to accuracy improvements of 2.4% and 7.5% on Food-101 and UECFOOD256 datasets, respectively. This showcases the effectiveness of combining deep learning with NLP for robust food image classification and complex data extraction in the culinary domain [Rahman et al., 2024].

In conclusion, these studies demonstrate the continuous advancements in food image classification using deep learning techniques. The use of advanced CNN architectures, transfer learning, and the integration with NLP open doors to progressively more accurate and sophisticated systems for various food-related applications.

III. METHODOLOGY

A. Data collection

The 101,000 images in the Food-101 dataset are split equally among 101 food categories, with 1,000 real-world photos in each category where 750 images are for training and 250 images are for testing purposes. This eclectic assortment includes dishes from many different international cuisines, including more specialized items like chocolate cake and chocolate mousse, as well as more commonplace items like pizza and apple pie. The dataset's diversity helps machine learning models develop more effectively and recognize a wider range of food items, meeting the needs of a global culinary community.

The Food-101 dataset's comprehensive representation of actual meal presentations is what makes it useful in real-world scenarios. These images, in contrast to staged photos, provide a more difficult challenge for image recognition models because they contain a variety of backgrounds, lighting conditions, and extraneous objects. This degree of intricacy is essential for creating reliable models that can function well under a variety of erratic and variable real-world circumstances.

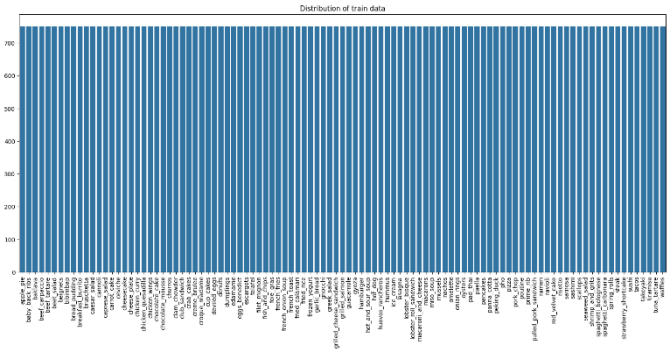


Figure 1: Distribution of train data

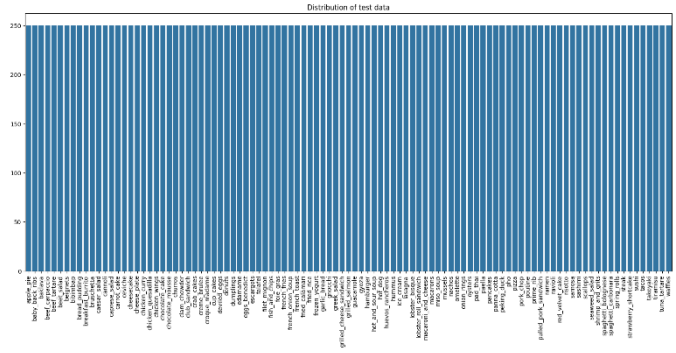


Figure 2: Distribution of test data

B. Data preprocessing

Using a data augmentation technique, a food image classification model's robustness and generalization abilities are strengthened by deliberately adding variability to the training dataset. TensorFlow and Keras libraries are used to implement a variety of transformations that accomplish this.

One approach is to rotate at random intervals of 20 degrees. This gives the model the ability to handle various food presentations where the item may be slightly skewed. Similarly, random horizontal and vertical shifts (up to 20% of the image dimension) simulate slight variations in camera position or object placement.

In order to help the model focus on particular regions of interest within the image and lessen the influence of background clutter, random zooming (up to 20%) is also incorporated into data augmentation.

Moreover, the method includes brightness corrections. Through arbitrary adjustments to image brightness between 80% and 120% of the initial value, the model gains proficiency in a variety of lighting scenarios. In order to strengthen the model's resistance to small geometric variations found in real-world scenarios, a random shearing transformation (up to 10%) is applied at the end.

This data augmentation method essentially produces a wider range of training examples. This makes it possible for the model to function more accurately on previously unseen images with small variations in rotation, position, brightness, or other

factors and helps prevent it from overfitting to the unique characteristics of the original dataset.

C. Technology used

The foundation of the project rested on the rich data provided by the Food-101 dataset. Encompassing 101,000 images across 101 food categories, this diversity offered a strong training ground for the CNN model. To ensure optimal training conditions, the images underwent preprocessing steps. This included resizing them for consistency (224x224 pixels) and applying normalization techniques to align them with the requirements of the pre-trained models that were incorporated.

The fundamental idea influencing all CNN architectures was transfer learning. This effective method makes use of the information extracted from previously trained models. Various pre-trained models were used as base architecture such as EfficientnetB7, EfficientnetB6, MobileNetV2 etc. The architectures took advantage of the learned features of these pre-trained models while avoiding overfitting in the initial training stage by freezing their initial weights.

Following the base model, a series of layers specifically designed for feature extraction and classification were implemented. Global Average Pooling condensed the feature maps into a more manageable format, while subsequent dense layers with ReLU activation (512, 256, and 128 units) progressively extracted higher-level features from the data. Dropout layers with a rate of 0.2 were strategically placed after each dense layer. This technique combats overfitting by randomly deactivating a portion of units during training, promoting model robustness.

The final layer of the CNN was designed for multi-class classification. It comprised a dense layer with 101 units, corresponding to the 101 food categories. The SoftMax activation function, commonly used in such problems, was employed in this layer.

To optimize the training process, the Adam optimizer was used to compile the model which is known for its effectiveness in handling noisy datasets like image collections. The learning rate was set at 0.0001. The sparse categorical crossentropy function served as the loss function, suited for the multi-class classification task at hand. Accuracy was the primary metric used to gauge the model's performance during training and testing.

Further optimization came through batch training, where the model was trained in batches of 16 images for efficiency. The total training steps per epoch were dynamically calculated based on the training dataset size. Regular validation checks were conducted using the validation dataset to assess the model's performance on unseen data. To prevent overfitting and unnecessary training time, Early Stopping was implemented. This technique halts the training process if the validation accuracy fails to improve for a predefined number of epochs.
Below is a brief introduction of the architecture of models used:

a)EfficientNetB4: This model uses ϕ=4 and is significantly larger in terms of depth and width. It aims at applications requiring very high accuracy, where computational constraints are less of a concern, making it ideal for high-end devices or cloud computing.

Here the value of ϕ determines how much to scale each of these dimensions to increase the model's capacity and accuracy systematically.

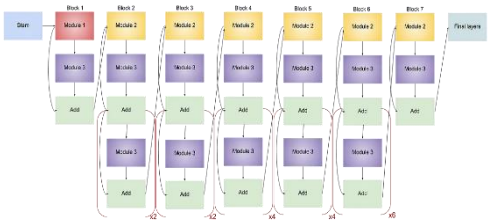


Figure 3: EfficientNetB4

b)EfficientNetB5: With ϕ=5, B5 expands the capabilities of the network to capture finer details through increased resolution, making it suitable for complex image classification tasks that involve very detailed visuals or varied classes.

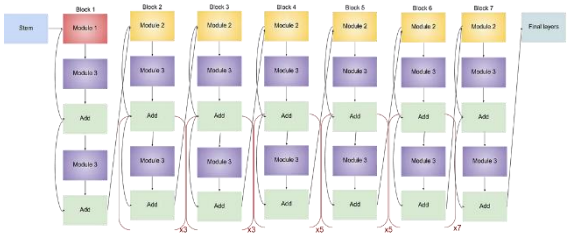


Figure 4: EfficientNetB5

c) EfficientNetB6: At ϕ=6, B6 continues to scale up the dimensions and is optimized for high-performance computing environments. It delivers exceptional performance on challenging datasets, benefiting significantly from advanced hardware.

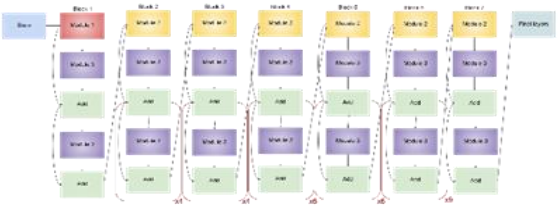


Figure 5: EfficientNetB6

d) EfficientNetB7: The largest model in the series, with ϕ=7, B7 represents the peak of scaling in this architecture. It is designed for cutting-edge applications that demand the highest accuracy and can leverage substantial computational resources, such as specialized AI accelerators.

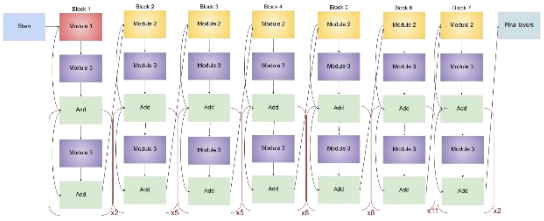
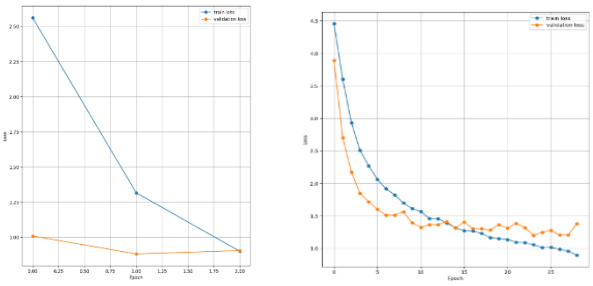


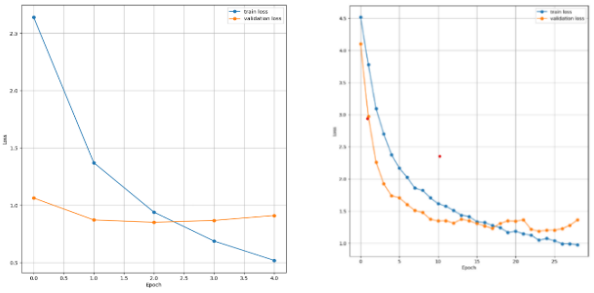
Figure 6: EfficientNetB7

IV. RESULTS



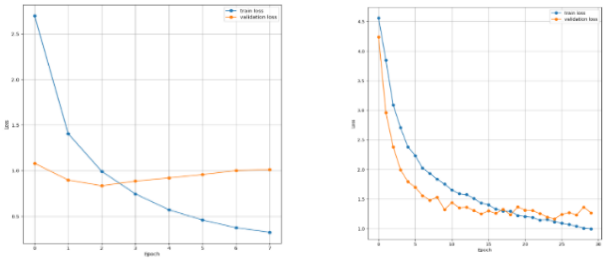
a)on Food 101 b)on augmented data

Figure 7: EfficientNetB7 Training and Validation loss



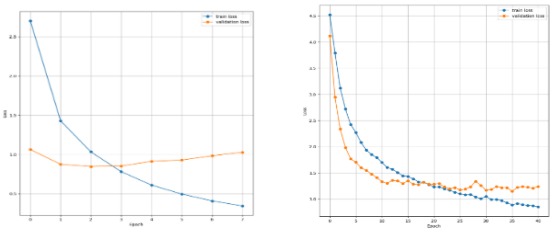
a)on Food 101 b)on augmented data

Figure 9: EfficientNetB5 Training and Validation loss



a)on Food 101 b)on augmented data

Figure 10: EfficientNetB5 Training and Validation loss



a)on Food 101 b)on augmented data

Figure 10: EfficientNetB4 Training and Validation loss

A. *Evalution Result on FOOD-101*

According to the original FOOD-101 dataset, the best accuracy result was achieved with EfficientNet-B6 at 81.09%, while the lowest result was obtained with EfficientNet-B4 at 80.68%. EfficientNet-B5 achieved an accuracy of 80.93%, and EfficientNet-B7 achieved 80.07%.

B. *Evalution Result on Augmented FOOD-101*

In the evaluations using the augmented FOOD-101 dataset, the accuracy decreased across all models. The highest accuracy was obtained by EfficientNet-B5 at 74.82%, while the lowest result was obtained by EfficientNet-B6 at 72.51%. EfficientNet-B4 achieved 74.17%, and EfficientNet-B7 achieved 72.67%.

V. DISCUSSION

This study investigated the effectiveness of transfer learning and data augmentation for EfficientNet models in classifying food images from the Food-101 dataset. Our findings revealed an unexpected result: models trained without data augmentation achieved consistently higher accuracy across all EfficientNet architectures compared to those with augmentation. The EfficientNet-B6 model, for instance, reached a peak accuracy of 81.09% without augmentation, significantly outperforming the 72.67% accuracy achieved with augmentation. Similar trends were observed for other EfficientNet models.

These results suggest a potential drawback of data augmentation in this specific case. The inherent diversity within the Food-101 dataset might already be sufficient for the models to generalize well. It's possible that excessive data augmentation introduced noise or irrelevant variations that confused the learning process. This highlights an important point: while data augmentation is a powerful technique to improve model robustness in many scenarios, its effectiveness can be highly dependent on the specific characteristics of the dataset.

The performance variations between the EfficientNet models also point to the role of model complexity in classification accuracy. The superior performance of the EfficientNet-B6 model signifies the advantage of using deeper and more complex architectures in capturing intricate details present in food images.

Models	Without Augmentation	With Augmentation
EfficientNet-B4	80.68 %	74.17 %
EfficientNet-B5	80.93 %	74.82 %
EfficientNet-B6	81.09 %	72.51 %
EfficientNet-B7	80.07 %	72.67 %

Table 1: Comparison between models

VI. CONCLUSION

This study examined the performance of EfficientNet models for food image classification on the Food-101 dataset, focusing on the effects of transfer learning and data augmentation. Surprisingly, models trained without data augmentation yielded higher accuracy. The EfficientNet-B6 model stood out, demonstrating the benefits of using deeper architectures.

These findings suggest that data augmentation, while valuable for improving generalization in many cases, needs to be carefully tailored to the dataset. The natural variability within theFood-101 dataset might be sufficient for effective model training.

In summary, this research contributes to the development of more accurate and efficient food image classification systems. It underscores the importance of evaluating augmentation strategies in the context of the specific dataset and model architecture. Future work will focus on refining augmentation techniques and investigating their impact on other datasets and classification tasks, aiming to optimize the balance between model robustness and accuracy.

REFERENCES

[1] B. M. Popkin, L. S. Adair, and S. W. Ng, “Global nutrition transition and the pandemic of obesity in developing countries,” Nutrition Reviews, vol. 70, no. 1, pp. 3–21, Jan. 2012, doi: https://doi.org/10.1111/j.1753-4887.2011.00456.x.

[2] World Health Organization, “Noncommunicable Diseases,” World Health Organisation, 2023. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

[3] N. Martinel, G. Foresti, and C. Micheloni, “Wide-Slice Residual Networks for Food Recognition.” Available: https://arxiv.org/pdf/1612.06543

[4] P. Tripathi, “TRANSFER LEARNING ON DEEP NEURAL NETWORK: a CASE STUDY ON FOOD-101 FOOD CLASSIFIER,” International Journal of Engineering Applied Science and Technology, vol. 5, no. 9, Jan. 2021, doi: 10.33564/ijeast.2021.v05i09.037.

[5] Y. Zhang et al., “Deep learning in food category recognition,” Information Fusion, vol. 98, p. 101859, Oct. 2023, doi: 10.1016/j.inffus.2023.101859.

[6] G. Suddul and J. F. L. Seguin, “A comparative study of deep learning methods for food classification with images,” Food and Humanity, vol. 1, pp. 800–808, Dec. 2023, doi: 10.1016/j.foohum.2023.07.018.

[7] V. G, P. Vutkur, and V. P, “Food classification using transfer learning technique,” Global Transitions Proceedings, vol. 3, no. 1, pp. 225–229, Jun. 2022, doi: 10.1016/j.gltp.2022.03.027.

[8] R. S. A. Kareem, T. Tilford, and S. Stoyanov, “Fine-grained food image classification and recipe extraction using a customised deep neural network and NLP,” Computers in Biology and Medicine, p. 108528, Apr. 2024, doi: 10.1016/j.compbiomed.2024.108528.

[9] L. Touijer, V. P. Pastore, and F. Odone, “Food image Classification: The benefit of In-Domain Transfer Learning,” in Lecture notes in computer science, 2023, pp. 259–269. doi: 10.1007/978-3-031-43153-1_22.

[10] A. Chaitanya, J. Shetty, and P. Chiplunkar, “Food image classification and data extraction using convolutional neural network and web crawlers,” Procedia Computer Science, vol. 218, pp. 143–152, Jan. 2023, doi: 10.1016/j.procs.2022.12.410.