**Food 101 Object Classification**

**Ferhan Ahmed Tanuj Pandey Shikha Roy**

**Student-PGDCSAI Student-PGDCSAI Student-PGDCSAI**

**Great learning/IIIT-Delhi Great learning/IIIT-Delhi Great learning/IIIT-Delhi**

**ferhanahmed02@gmail.com codename.tanuj@gmail.com rshikha29@ymail.com**

**(1) Abstract**

Food picture categorization is a major computer vision challenge with applications ranging from automated food business services to dietary monitoring and food recommendation systems. In order to increase the precision of food image categorization, this study investigates three different approaches: a basic CNN ensemble, an ensemble of pretrained VGG16 models, and a single pretrained VGG16 model. The dataset used in the study is made up of food photos that have been divided into training and testing subsets and classified into 17 different groups. The first method, which used a basic CNN ensemble, showed only modest success in classifying and extracting features. The second method produced better results by combining features from different models and using transfer learning to create an ensemble of pretrained VGG16 models. Out of all the strategies examined, the third and final strategy had the highest classification accuracy since it only used one pretrained VGG16 model. Extensive pretraining on a large-scale dataset improved this model's performance overall and its ability to extract features. The study illustrates the benefits of pretrained models for challenging picture classification problems and shows how sophisticated architectures can significantly improve accuracy. These findings imply that performance in food picture classification applications can be greatly enhanced by utilizing transfer learning and adding pretrained models.

**Keywords**: Food Image Classification, Convolutional Neural Networks, Ensembling Techniques, VGG16, Data Augmentation, Early Stopping

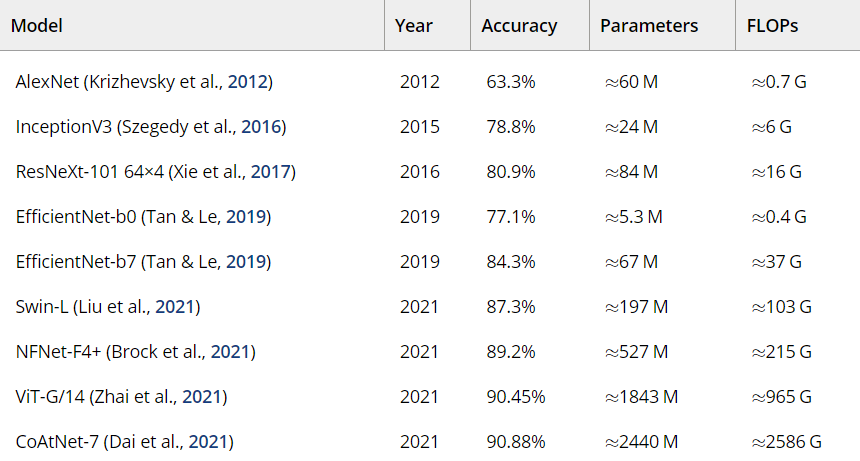
**(2) Introduction**

Food image classification is a crucial problem for many applications such as nutritional analysis and meal recommendation. It entails classifying photographs of food items into predetermined groups. The intricacies between identical foods and the variations in food appearances make it difficult to classify food photographs accurately. The objective of this study is to assess and contrast several model architectures in order to identify the best method for classifying food images. The study aims to determine how model complexity and transfer learning affect classification accuracy by examining a single pretrained VGG16 model, an ensemble of pretrained VGG16 models, and a basic CNN ensemble. The results of this study may help design more precise and effective classification algorithms as well as best practices for the categorization of food images.

**(3) Literature Review**

**Related Work**

Deep learning techniques have significantly improved food image classification compared to previous machine learning techniques. Earlier studies, such the one conducted by Bossard et al. (2014), classified food photos using Random Forests with an accuracy of 50.76%. This method exposed the shortcomings of previous approaches and illustrated the difficulties in reliably identifying food items from photographs.



Performances of some relevant previous architectures are shown above.

A big advancement was made with the development of convolutional neural networks, or CNNs. Although this study's dataset size hindered it, Lu (2015) used CNNs and data augmentation to increase accuracy to 90%. With the use of Inception modules, Liu et al. (2016) improved the field with their DeepFood model, which achieved a top-1 accuracy of 77.4% on the Food-101 dataset. The ability of deep learning models to handle challenging image categorization tasks was demonstrated by this work. Classification performance has been significantly improved by the application of ensembling techniques. Because combining numerous models addresses the shortcomings of individual models, accuracy and robustness can be increased. To improve performance and generalization, various models' strengths have been combined through the use of techniques like bagging, boosting, and stacking. In order to improve model performance even further, recent research has investigated the integration of attention processes with Generative Adversarial Networks (GANs), as demonstrated by the works of Kim et al. (2022) and Huang et al. (2023).

**(4) Materials and Methods**

**4.1 Proposed Work**

Two approaches were first tried in our research for classifying food images, however we ran into issues with accuracy. The first technique used three identical CNN models in a basic CNN ensembling strategy. Prior to categorization, the features from these models were concatenated. Although this strategy has potential, its accuracy on the test data was only 27.7%, suggesting that the complexity and feature extraction capabilities of the model were not sufficient. The second technique used pretrained VGG16 models for ensembling. This method increased accuracy on validation data to 45.6%. VGG16 offered improved feature extraction capabilities with its pretrained weights and deep architecture. But more improvement was required, which led us to investigate a more sophisticated strategy.   
In the end, we only employed one pretrained VGG16 model. Pretrained weights were used by the TensorFlow and Keras-trained VGG16 model to take advantage of enhanced feature extraction capabilities. To strengthen the resilience of the model, data augmentation approaches were used. On the test data, this approach yielded the greatest accuracy of 60.6%. Improved performance was mostly attributed to the pretrained features and depth of the VGG16 architecture.

**4.2 Methodology**

The suggested work aimed to use cutting edge deep learning techniques to increase the classification accuracy of food images. In order to do this, a methodical strategy was used, which included optimizing a pretrained VGG16 model—well-known for its performance in picture classification tasks—to do this. The approach concentrated on using the pretrained model's strengths and customizing it for the task of classifying food images.

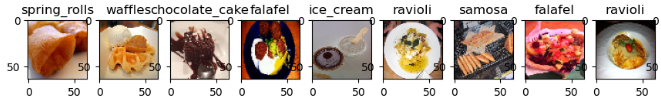
**4.3 Data Details**

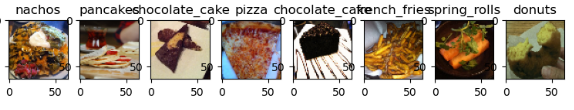
Pictures from 17 distinct cuisine groups were included in the dataset utilized for this investigation. A 70:30 ratio was used to divide the total number of photos into training and testing sets.



Food images from dataset

Data augmentation approaches were used to improve the model's capacity for generalization and to facilitate efficient model training.





Sample images visualized After EDA

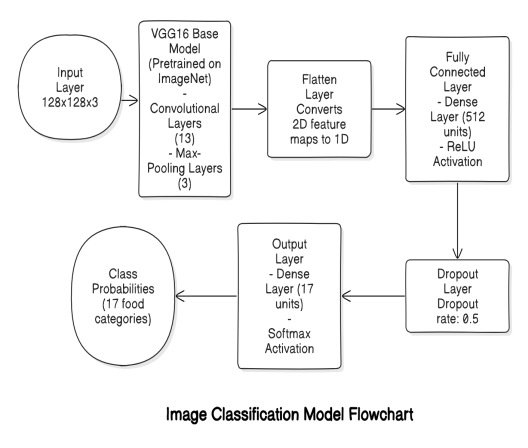
Among these methods were: Rotation: To provide the food items in the model a variety of orientations. Flipping: Adding both vertical and horizontal variances. Scaling: To take into consideration the various food item sizes in photos. Cropping: To make the photographs appear to have various focal points. Color corrections: To bring color variances and lighting situations into balance. During training, these augmentations were used to make sure the model was resilient and could handle a wide variety of food photos.

**4.4 Method**

The VGG16 model with pretrained weights was the technique used to classify food-related images. The VGG16 model's capacity to extract rich feature representations from images made it a solid foundation because it was pretrained on the ImageNet dataset. Important actions comprised:   
Changing Fully linked Layers: To better categorize the 17 food groups, custom layers were added to the VGG16 in place of the original fully linked layers. The model is able to specialize in the task of classifying food images thanks to this adaptation.   
Fine-Tuning: Using the enhanced training data, the model was adjusted. To boost performance, the pretrained model was fine-tuned by varying its weights and retraining it on the particular food dataset.

**4.5 Architecture**

The components of the VGG16 architecture are as follows: The task of extracting features from the input photographs is assigned to thirteen convolutional layers. These layers are configured with limited receptive fields and homogeneous architecture to capture hierarchical features. Three max-pooling layers are used to reduce the spatial dimensions of the feature maps while maintaining significant features.   
The original VGG16 model has three fully connected layers. In this investigation, the last fully linked layer was a custom output layer that was configured to classify images into 17 different food groups. The architecture was selected because it has a history of success in numerous picture classification applications and deep feature extraction capabilities.

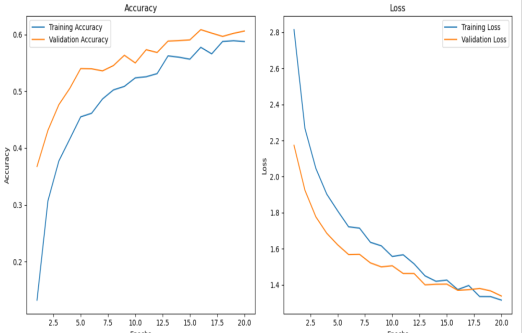


**4.6 Implementation**

Two well-known deep learning frameworks, TensorFlow and Keras, were used in the implementation. The final layer of the VGG16 model was adjusted to fit the requirements of the food categorization task once it was imported with pretrained weights. The ensuing implementation specifics were observed: Preprocessing of the data: The images were resized to 224 × 224 pixels, which is the smallest size that VGG16 requires for input. Data augmentation was done to provide a training dataset more diversity.   
Training Process: The model was trained using the improved training set. The learning rate, batch size, and number of epochs were among the hyperparameters that were adjusted to optimize the model's performance.   
Evaluation: Using the test set, the model's performance and accuracy were evaluated. Metrics including accuracy and loss were monitored throughout training.

### (5) Results

The final method using a pretrained VGG16 model achieved a test accuracy of 60.6%, marking a significant improvement over previous methods.



**Training Vs Validation**

When easy CNN Ensembling Method was used then 27.7% accuracy was obtained. The lack of unique features and the simplicity of the three similar CNN models with feature concatenation limited the performance of this technique. When ensembling, pre-trained VGG16 models yielded 45.6% accuracy. Combining predictions from many pretrained VGG16 models increased accuracy over the usual CNN ensemble, however this strategy did not achieve its full potential. With only pre-trained VGG16, accuracy was 60.6%. This approach made use of a single pretrained VGG16 model that was adjusted and had additional data. It made great use of deep feature extraction and optimization approaches, outperforming both previous methods.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Methodology** |  | | --- | --- | | Simple CNN Ensemble |  | | Ensemble of Pretrained VGG16 Models |  | | Single Pretrained VGG16 Model |  | | **Accuracy (%)**  27.7  45.6    60.6 |

**(6) Discussion and Conclusions**

The outcomes show that the pretrained VGG16 model performs better than the ensemble of pretrained VGG16 models as well as the simple CNN ensemble. The single pretrained VGG16 model's outstanding accuracy demonstrates the advantages of transfer learning and the stability of the VGG16 architecture. Due to its restricted feature extraction capabilities, the simpler CNN ensemble performed poorly; in contrast, the ensemble of pretrained VGG16 models outperformed the single pretrained model, but not by much.

In order to improve classification accuracy, future work should concentrate on further optimizing the pretrained models and investigating alternative sophisticated architectures. Furthermore, utilizing data augmentation methods could enhance the generalization of the model. The study emphasizes how crucial it is to use sophisticated architectures and pretrained models to achieve better results in food image categorization tasks.

**(7) References**

[1] B. Antonio, D. Moroni, and M. Martinelli, “Efficient adaptive ensembling for image classification,” *Expert Syst.*, 2023.

[2] M. Jiang, *Food Image Classification with Convolutional Neural Networks*. .

[3] L. Bossard, M. Guillaumin, and L. Van Gool, “Food-101 – mining discriminative components with random forests,” in *Computer Vision – ECCV 2014*, Cham: Springer International Publishing, 2014, pp. 446–461.

[4] C. Liu, Y. Cao, Y. Luo, G. Chen, V. Vokkarane, and Y. Ma, “DeepFood: Deep learning-based food image recognition for computer-aided dietary assessment,” *arXiv [cs.CV]*, 2016.

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[6] J. Hu, L. Shen, and G. Sun, “Squeeze-and-Excitation Networks,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[7] Y. Zhang and L. Xu, “A Comprehensive Review of Ensemble Deep Learning: Opportunities and Challenges,” *International Journal of Computer Vision*.