Biomedical Testing & Sensing
2025, VOL. 4, NO. 1, 13-15
DOI: 10.47297/wspbtsWSP2755-952103.20250401

Food Recognition Simulation Based on MobileNetV2 Model

Suce Hu

School of Computer Science, University College Dublin, Dublin, County Dublin, D04 V1W8, Ireland

ABSTRACT

Food image recognition technology holds significant application value in personalized diet management, allergen early warning and other fields, but it faces challenges such as scarcity of rare samples, large intra-class differences and insufficient dynamic learning ability. Traditional machine learning methods rely on manual feature extraction, resulting in relatively low recognition accuracy. However, deep learning has significantly improved performance through models such as convolutional neural networks (CNN) and visual transformers (ViT). Recent studies have further optimized the model architecture and training strategies: EHFR-Net alleviates the problems of position information loss and background interference by integrating ViT and CNN structures; online continuous learning methods (such as power iteration clustering and knowledge distillation) solve the challenge of dynamically adding new categories and reduce the impact of intra-class differences. Future research should continue to explore lightweight and highly generalized multi-task models to meet the complex requirements of food recognition scenarios.

KEYWORDS

Food image classification; Deep learning; Intra-class variation; Continual learning; Lightweight model

1 Introduction

Food holds an important position in human life. Besides providing the essential nutrients needed by the human body, it also influences people's behaviors, moods and social activities. Gilal et al. (2024) believe that the application of deep learning technology in identifying food has significant practical significance in various fields. For instance, the system can calculate the calorie and nutritional content of food by analyzing food images, thereby assisting users in achieving personalized dietary management. This is particularly important for special groups (diabetic patients, fitness enthusiasts). In addition, the system can also alert people with allergies to the allergenic components of food, thereby reducing the risk of accidental ingestion [2]. The theoretical basis for achieving these functions is the food image classification technology. Karuppusamy et al. (2025) mentioned that the existing food image classification technology faces two major challenges. One is the scarcity of samples for rare dishes, which leads to a decline in the generalization ability of the classification model. The other is the insufficient processing of tasks, as most studies only focus on a single task (such as classification) [1]. With the development of the times, the technology for food recognition has gradually evolved from the traditional machine learning methods to the current deep learning methods. Ghosh et al. (2025) concluded that for traditional machine learning, food features are mainly extracted manually, such as Gabor filters, Local Binary Patterns (LBP), etc. The main food classification methods are Support Vector Machines, Random Forests, and KNN. These methods result in an average recognition accuracy of food models generally below 85%. Due to the limitations of traditional machine learning methods, deep learning techniques have achieved performance breakthroughs through customized convolutional neural networks and transfer learning. For example, the pre-trained EfficientNet model achieves an accuracy of 96.18% on the Food-101 dataset. Based on the emerging ViT-Base model, they significantly improved the accuracy of the new model on mainstream food datasets by injecting linear noise and adjusting the system entropy [3]. Prior to this, Sheng et al. (2024) developed an efficient lightweight hybrid model, EHFR-Net, which optimized the problem of position information loss in ViT and the susceptibility of CNN to food background interference, achieving a balance between lightweight design and high accuracy [4]. However, the challenges of deep learning technology are similar to those of traditional machine learning. Intra-class differences (such as different cooking styles of pizzas) remain the main source of errors. Moreover, traditional food classification models rely on static datasets for training and are unable to dynamically learn new food categories (such as when users upload new dish images). Based on these deficiencies, He et al.(2021) first proposed online continuous learning. They used power iteration clustering (PIC) to generate sub-clusters based on visual similarity, and then selected representative samples from the center of each sub-cluster and stored them as a sample set. This sample selection mechanism alleviates the high intra-class variance of food data. It has verified the feasibility of lifelong learning on large food databases by combining the balanced training and knowledge distillation mechanisms [5].

2 Literature review

The technological evolution of deep learning in food recognition mainly consists of two stages: the CNN dominance period and the Transformer breakthrough period. From 2010 to 2019, CNN's early research utilized models such as VGG and ResNet to achieve food classification through local feature extraction. For instance, VGG30 achieved an accuracy rate of 71% on the Food-101 dataset [1]. However, CNN has difficulty in capturing the overall features of food (such as the spatial relationships of scattered ingredients), and the need for deep networks leads to a large number of parameters [4]. Enguang et al.(2025) claimed that most of the current datasets only provide category labels and lack detailed information such as ingredients and cooking methods [6]. The period from 2020 to 2023 was a breakthrough period for Transformers. Ghosh et al., was the first to introduce Vision Transformer into food recognition, using the self-attention mechanism to model pixel-level long-range dependencies. The top-1 accuracy rate on Food-101 reached 99.5%. ViT also has two efficient variants. NoisyViT adjusts the entropy value by injecting noise to enhance the model's robustness [3]. LP-ViT retains the spatial position information and solves the problem of spatial information loss caused by the embedding layer in traditional ViT [4]. From 2023 to the present, Sheng et al., proposed EHFR-Net, which combines the local feature extraction capability of CNN with the global modeling ability of LP-ViT, and achieves an accuracy rate of 90.7% (Food-101) through the lightweight HBlock unit, with only 2.8M parameters. When the number of parameters is comparable, the hybrid model has a 3.5% higher accuracy than MobileViTv2.

In the deep learning of food recognition, another important concept is continuous learning. The first challenge of continuous learning in food recognition is the problem of catastrophic forgetting. The traditional static training model is unable to adapt to new food categories and requires repeated training, resulting in waste of resources [5]. Furthermore, the high intra-class variation in food data (such as different presentations of the same dish) increases the risk of forgetting [2]. Another challenge is the imbalance and dynamic distribution of data. In real scenarios, the frequency of food consumption follows a long-tail distribution, and the number of future categories is unpredictable. Online learning requires processing data streams in a single instance, making it difficult to balance samples of new and old categories [5].

Based on the challenges mentioned above, He et al., proposed an example selection method based on Power Iteration Clustering, generating dynamic clustering based on visual similarity, and selecting samples from the cluster centers to cover a more diverse range of food forms. Compared with the Herding algorithm (which only approximates the class mean), this solution achieves a 13.9% improvement in accuracy on the Food-1K dataset. In addition, the document also proposed balancing training and knowledge distillation, allowing each new data to be paired with random old examples of the same class for training, thereby alleviating the imbalance of categories. Apply operations such as flipping and color distortion to the examples to enhance the contrast of the generated batches. Constraint the outputs of the teacher-student model through the KL divergence loss. In terms of model selection, this scheme combines the hybrid architecture (such as EHFR-Net) with dynamic example storage, achieving an average accuracy rate of 61.2% on the Food-1K dataset, outperforming baseline models such as GDUMB.

3 Experiment design

This experiment, in combination with the above introduction of continuous learning, conducted a continuous learning simulation experiment for food recognition on the Food-11 dataset. The core objective of the experiment is to achieve end-to-end classification for 11 types of foods (such as bread, dairy products, desserts, etc.) and to build a continuous learning framework to support the dynamic update of the model. The experiment will be carried out in two stages. Firstly, the standard supervised learning process will be optimized, and the basic classification performance will be improved through training techniques. Then, the focus will be on studying the continuous learning scenario, simulating the situation where new food categories (such as the first 7 basic food categories + the last 4 newly added food categories) are introduced in the real world. A continuous learning strategy will be adopted to avoid catastrophic forgetting. Finally, by comparing the performance differences between single training and staged continuous learning, an adaptive optimization scheme for lightweight models in dynamic data environments will be explored to provide technical support for model iteration and upgrade in practical applications.

The experimental content covers four aspects: data, model, training strategy and evaluation. The first step is data collection and preprocessing. Firstly, a custom food image dataset (with 11 categories) is constructed. The data is stored in the folders: ./training, ./validation and ./evaluation. Then, image enhancement techniques (such as random cropping and horizontal flipping) and data normalization (based on the mean and standard deviation of ImageNet) are used to improve the generalization ability of the model. Finally, in the continuous learning training, an additional data partitioning strategy is introduced: the training set is divided into 7 categories of food and 4 categories of food to simulate the continuous learning scenario. The second step is model architecture and initialization. The experiment first uses the pre-trained MobileNetV2 as the base model and employs transfer learning technology. Then, the parameters of the convolutional layers (model.features.parameters()) are frozen to reduce computational costs and accelerate convergence. Finally, in the continuous learning training, the model is initialized in stages: stage 1 (training with old data) and stage 2 (training with new data), and the learning rate is adjusted to control the update intensity. The third step is training strategy and optimization. The experiment initially adopts the standard supervised training strategy, using a standard end-to-end training approach (20 epochs), with the Adam optimizer and cross-entropy loss. To prevent overfitting, early stopping based on the validation set (saving the best model) is employed. Then comes the continual learning strategy, where in stage 1, the model is trained only with the old data (7 types of food), simulating the acquisition of initial knowledge. In stage 2, the model from stage 1 is loaded and trained with new data (4 types of food), simulating continuous learning. The optimizer is uniformly set to Adam, and the loss function is the cross-entropy loss (nn.CrossEntropyLoss()) as well as the knowledge distillation loss function. The fourth step is evaluation and validation. The experiment first monitors the training loss and accuracy in real time during the training process. Secondly, the validation set (./validation) is used for model selection and early stopping, and the test set (./evaluation) is used for final performance evaluation. Then, in the continuous learning training, the validation set and test set are independently evaluated after each stage to quantify the effect of continuous learning. The core evaluation metric is the classification accuracy (in percentage form), which is output to the console for analysis.

4 Experiment result

In the end-to-end experiments of the first part, the MobileNetV2 model achieved a test accuracy of 78.73% on all 11 types of food. In the second part of the experiment, the teacher model achieved a test accuracy rate of 81.06% for the first seven types of food, 0% for the last four types of food, and 53.69% for all types of food. The student model that was trained on the teacher model achieved an accuracy rate of 74.52% for the first seven categories of food, 79.82% for the last four categories, and 76.31% for all food items. These accuracy rates indicate that the model did not catastrophically forget old knowledge while learning new knowledge.

5 Conclusion

In this experiment, the continuous learning method demonstrated significant advantages in the food classification task. Firstly, the end-to-end trained MobileNetV2 model achieved a test accuracy of 78.73% across all 11 food categories. In the continuous learning model using the teacher-student framework, the student model maintained an accuracy of 74.52% for the old categories (the first 7 categories), while achieving an accuracy of 79.82% for the new categories (the last 4 categories), resulting in an overall accuracy of 76.31%. This is significantly better than the 0% performance of the teacher model in the new categories. This indicates that the model successfully learns new knowledge effectively while avoiding the catastrophic forgetting problem. Future improvement directions include further enhancing the model's ability to retain knowledge of old categories, optimizing the knowledge distillation strategy to improve the integration effect of new and old knowledge, and exploring more efficient continuous learning frameworks to adapt to larger-scale category increment scenarios.

References

[1] Karuppusamy, S., Murugan, P. S. S., Srinesh, M., Surya, S., & Ahamed, S. T. (2025). Food recognition system using deep learning. In Challenges in Information, Communication and Computing Technology (pp. 687-692). CRC Press.

[2] Gilal, N. U., Al-Thelaya, K., Al-Saeed, J. K., Abdallah, M., Schneider, J., She, J., ... & Agus, M. (2024). Evaluating machine learning technologies for food computing from a data set perspective. Multimedia Tools and Applications, 83(11), 32041-32068.

[3] Ghosh, T., & Sazonov, E. (2025). Improving Food Image Recognition with Noisy Vision Transformer. arXiv preprint arXiv:2503.18997.

[4] Sheng, G., Min, W., Zhu, X., Xu, L., Sun, Q., Yang, Y., ... & Jiang, S. (2024). A lightweight hybrid model with location-preserving ViT for efficient food recognition. Nutrients, 16(2), 200.

[5] He, J., & Zhu, F. (2021). Online continual learning for visual food classification. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 2337-2346).

[6] Liu, D., Zuo, E., Wang, D., He, L., Dong, L., & Lu, X. (2025). Deep Learning in Food Image Recognition: A Comprehensive Review. Applied Sciences, 15(14), 7626.