Human-Centric Intelligent Systems (2024) 4:171–186
https://doi.org/10.1007/s44230-023-00057-9

RESEARCH ARTICLE

Automatic Food Recognition Using Deep Convolutional Neural Networks with Self-attention Mechanism

Rahib Abiyev[1] · Joseph Adepoju[1]

Received: 3 August 2023 / Accepted: 28 November 2023 / Published online: 9 January 2024

Abstract

The significance of food in human health and well-being cannot be overemphasized. Nowadays, in our dynamic life, people are increasingly concerned about their health due to increased nutritional ailments. For this reason, mobile food-tracking applications that require a reliable and robust food classification system are gaining popularity. To address this, we propose a robust food recognition model using deep convolutional neural networks with a self-attention mechanism (FRCNNSAM). By training multiple FRCNNSAM structures with varying parameters, we combine their predictions through averaging. To prevent over-fitting and under-fitting data augmentation to generate extra training data, regularization to avoid excessive model complexity was used. The FRCNNSAM model is tested on two novel datasets: Food-101 and MA Food-121. The model achieved an impressive accuracy of 96.40% on the Food-101 dataset and 95.11% on MA Food-121. Compared to baseline transfer learning models, the FRCNNSAM model surpasses performance by 8.12%. Furthermore, the evaluation on random internet images demonstrates the model's strong generalization ability, rendering it suitable for food image recognition and classification tasks.

Keywords CNN · Food-101 · Food classification · Self-attention · MA Food-121

Abbreviations

ANN	Artificial neural network
BOW	Bag of words
CNN	Convolutional neural network
IFV	Improved fisher vector
LSTM	Long short-term memory
RELU	Rectified linear unit
RCF	Randomized clustering forest
RF	Random forest
RNN	Recurrent neural network
SVM	Support vector machine
WHO	World health organization

1 Introduction

The significance of food in human health and well-being cannot be overemphasized. It is vital for survival, furnishing the body with essential nutrients and energy necessary for optimal functioning. Inadequate food intake can have adverse effects on fundamental bodily processes, such as the maintenance of a resilient immune system and the repair of cells and tissues. In today's increasingly competitive and dynamic world, people are growing more conscious of their health due to the growing prevalence of nutrition-related health issues. Moreover, Qui et al. [32] noted an upward trend in the incidence of diet-induced diseases in various populations. As reported by the World Health Organization (WHO), the incidence of obesity worldwide more than doubled from 1980 to 2014, with 13% of individuals categorized as obese and 39% of adults classified as overweight. This trend can be partially attributed to inadequate management of individuals' daily dietary intake.

Accurately assessing one's dietary habits is paramount in mitigating the detrimental effects of unhealthy food choices and the growing prevalence of diet-related ailments. In this regard, food recognition systems that leverage image recognition technology have a pivotal role to play [36]. Such

✉ Joseph Adepoju
adepoju97@gmail.com

Rahib Abiyev
rahib.abiyev@neu.edu.tr

[1] Department of Computer Engineering, Applied Artificial Intelligence Research Centre, Near East University, Lefkosa, Northern Cyprus

systems bolster food traceability and enable precise tracking of one's dietary intake, thereby making a meaningful contribution towards our collective health objectives. This cutting-edge technology is an indispensable tool in addressing the multifaceted challenges presented by modern dietary habits and their associated health risks.

The relevance of an automated food recognition system is evident in its impact on public health and well-being. As discussed earlier, food monitoring and accurate intake assessment are essential for reducing the risk of developing chronic illnesses like obesity, diabetes, and cancer. By providing a reliable and highly accurate food classification model, individuals can acquire valuable insights into their dietary habits, enabling them to make informed decisions about their food intake, thus contributing to disease prevention and health maintenance. Moreover, the integration of such a system into mobile applications allows users to easily monitor their food intake on a daily basis. Additionally, the technology can be seamlessly integrated into social media platforms, facilitating the categorization of users based on their food preferences or shared food pictures. This opens up possibilities for targeted advertising, reducing the stress associated with generic advertising and enhancing overall efficiency. Furthermore, an automatic food image recognition system holds promise in improving supply chain efficiency and food safety. By enabling quick and accurate tracking and identification of food items, food producers and distributors can enhance overall food traceability.

In recent years, food image recognition has made impressive strides, thanks to technological advancements. Convolutional Neural Networks (CNN) have played a crucial role in improving image identification and classification [27]. As a result, they have achieved remarkable accuracy when handling large image datasets. Nevertheless, deep learning, which is the most accurate method, still lags behind human recognition due to the absence of well-developed solutions [26]. The classification of food images presents unique challenges such as the diversity of food types, variations in presentation and lighting conditions, and high-level semantics. High-level semantics entail detailed information about a food item, such as its ingredients, preparation techniques, and cultural context. Recognizing food images is a complex task, and ongoing research is crucial. Recent studies [23] have emphasized the need for continued exploration in this field.

In this paper, we tackle these challenges through the development of an advanced food recognition model, FRCNNSAM. By integrating advanced techniques like scaled dot-product attention and ensemble modelling, FRCNNSAM can accurately identify intricate patterns and relationships in food images. Its ensemble approach combines prediction probabilities through rigorous averaging methods, resulting in enhanced robustness and performance in food recognition tasks. These features make FRCNNSAM a dependable and effective tool for image classification. Furthermore, FRCNNSAM model is meticulously engineered to optimize computational resources and memory usage, incorporating techniques like weight sharing and data compression for enhanced efficiency and scalability.

Moreover, our study seeks to investigate whether CNN models, developed without the utilization of transfer learning techniques, can achieve performance levels comparable to those employing transfer learning. This inquiry is rooted in the potential of our FRCNNSAM model to contribute to disease prevention and the promotion of healthier dietary habits, addressing practical challenges in food monitoring and public health.

This paper is organized as follows: Sect. 2 details an overview of previous approaches employed in food image recognition and the corresponding outcomes attained. In Sect. 3, we present the materials, such as methods and datasets, used in developing our food recognition model. Section 4 delves into the design and architecture of the FRCNNSAM model, covering aspects such as the proposed model structure and simulation results. The study concludes with Sect. 5, which offers recommendations for future research.

2 Literature Review

In this section, we delve into the literature related to food image recognition, exploring various methods and techniques that have been employed in the past decade. The field of food recognition has witnessed significant advancements due to the integration of machine learning, computer vision, and improved processing efficiency. We will discuss the prevalent use of deep learning techniques, particularly CNN, and the application of attention mechanisms and ensemble models to image recognition tasks.

CNN have gained popularity in food image recognition due to their remarkable features like equivariant representation, sparse interaction, and, parameter sharing [27]. These advanced deep learning models have been extensively applied in recent research and publications related to image identification and classification. The application of CNN in food image recognition has significantly contributed to the field's success, enabling high accuracy in analysing large image datasets. Prior to the widespread adoption of CNN for food image identification and classification, alternative methods were commonly employed. An early method for food recognition, the Fisher vector technique, was introduced by Sanchez et al. in [34]. This technique analyses the visual characteristics of food images in certain regions using a mathematical tool called the Fisher kernel. The Fisher kernel uses a generative model (e.g., Gaussian Mixture Model) to express the divergence of a sample from the model as a unique Fisher Vector that can be applied to classification. Another method that was used was the bag of visual words (BOW) technique, which involves vector quantization of affine-invariant descriptors extracted from image patches. According to Csurka et al. [12] this approach is made to address the difficulty of recognizing objects in realistic photos while taking into account inherent variances within the object class. In addition, a method was suggested by Matsuda et al. [24] that enables the recognition and categorization of food items within an image. This comprehensive approach makes use of a variety of methodologies. The method begins by identifying potential regions of food using several techniques, including whole image analysis, object detection using the deformable part model and linear SVM as suggested by Felzenszwalb et al. [15], image segmentation using JSEG algorithm and circle detection using Canny Edge Detector and Hough Transform. Subsequently, it extracts features from these regions identified as candidate areas within the image and applies a sophisticated machine learning technique called multiple-kernel learning with non-linear kernels for image classification, allowing for the identification of multiple food items.

In their 2014 study, Bossard et al. utilized discriminative components, specifically random forest mining—a framework crafted for simultaneous extraction across all food dataset classes. They also introduced the Food-101 dataset, now widely embraced as a benchmark in the research community for multi-class food datasets. Comprising 101 unique food categories, the dataset contains a total of 101,000 images. Each food category is represented by 1000 images, with 250 images designated for testing and 750 for training purposes. Notably, the random forest technique they proposed was able to learn across multiple classes. Their experiment revealed that the model accuracy was 50.76%. It is important to highlight that, apart from CNN which achieved an accuracy of 56.70%, this model surpassed alternative methods such as IFV, and BOW, and local methods like RCF and RF [10].

Although these methods have performed well in the past for recognizing and categorizing food images, deep learning algorithms, particularly CNN, have significantly outperformed them in the last several years. The next paragraph will describe how diverse studies have utilized deep learning methods for the task of food image classification.

The widespread adoption of deep learning algorithms has notably enhanced the accuracy of image categorization. As a result, numerous researchers have explored the potential of deep learning in the classification of food images. In the context of categorizing Indian food images, Vijaya Kumari et al. leveraged transfer learning techniques, utilizing pre-trained models such as InceptionV3, VGG16, VGG19, and ResNet. Their findings highlight the superior performance of InceptionV3 in classifying various Indian food images. Ma et al. took a different approach by utilizing deep learning to predict food categories and nutrient content using ingredient statements. Their model, named "Ingredient2Vec," demonstrated high accuracy in this predictive task. Bishop et al. [9] introduced a deep learning model employing LSTM-RNN to classify major cuisine types of takeaway food outlets, showcasing its potential in automating this categorization, valuable for public health monitoring and decision-making. Akhi et al. [6] achieved a high success rate in recognizing and classifying fast food images. They used a multi-class linear Support Vector Machine (SVM) classifier along with a pre-trained Convolutional Neural Network (CNN) as a feature extractor. Finally, Özsert Yiğit G, & Özyildirim explored food image classification using deep convolutional neural networks, comparing models trained from scratch with pre-trained structures like Alexnet and Caffenet. Their findings underscored comparable performance between the proposed trained-from-scratch models and pre-trained models, with optimization methods improving the overall performance of the compared models. These studies collectively underscore the potential and versatility of deep learning techniques in achieving high accuracy for food image recognition tasks, attributed to deep learning's capability to capture intricate and complex patterns, thus making it well-suited for tasks involving subtle differences or intricate textures in food images. However, it's crucial to acknowledge that achieving optimal performance with deep learning models often necessitates a significant amount of labelled training data, presenting challenges when such data is scarce, especially in specialized food domains, and additionally, they may be susceptible to over-fitting, particularly when the training data is limited or when the model's architecture is overly complex.

Recent research has shown that the fusion of CNN architectures, known as ensemble modeling, has surpassed the accuracy of state-of-the-art CNN models, achieving even better results. For instance, Fakhrou et al. [13] developed a smartphone-based system to recognize food and fruits for visually impaired children. They combined two deep CNN models, InceptionV3 and DenseNet201, fine-tuned on a customized food recognition dataset. They applied average voting (also known as soft voting) for ensemble learning, achieving a 95.55% accuracy on their customized food dataset. Similarly an ensemble framework that includes several pre-trained CNN models, such as InceptionV3, DenseNet201, and ResNeXt-50, was presented by Rane et al. [33]. This ensemble model achieved a better AUC-ROC than other underlying researches. These studies exemplify the success of ensemble models in enhancing the accuracy of food image classification tasks, providing valuable insights for the development of our proposed FRCNNSAM model.

Recent studies not only showcase successful results with deep learning and ensemble models but also investigate the effectiveness of self-attention in image recognition models [39]. When combined with convolutional techniques, self-attention has proven effective in various computer vision tasks. Zhao et al.'s research emphasizes the robustness and generalization advantages offered by self-attention networks.

In summary, the literature review has showcased the progression of food image recognition methods over time. From early techniques like the Fisher vector approach to the advent of deep learning, particularly CNN, the field has witnessed significant advancements in accuracy and performance. Ensemble learning has emerged as a powerful method for further enhancing the capabilities of CNN models, demonstrating improved results in food image recognition tasks. Additionally, the exploration of self-attention mechanisms has shown promise in augmenting convolutional approaches, leading to improved robustness and generalization. Building on these foundations, the proposed FRCNNSAM model amalgamates the strengths of deep learning (CNN), ensemble learning, and self-attention to create an efficient and accurate food image recognition system.

3 Materials and Method

This chapter outlines the comprehensive process that was employed in the design and implementation of the FRCNNSAM model for precise and effective food image recognition. This chapter provides an in-depth look at the major components that contribute to the development of the FRCNNSAM model.

3.1 Dataset

The FRCNNSAM model was trained and evaluated using two unique datasets, the Food-101 dataset and the MA_Food-121 dataset. Bossard et al. [10] created the Food-101 dataset, which includes food images from foodspotting.com. Users can use this website to exchange photographs, locations, and information about the food they are eating. The dataset comprises a grand total of 101,000 images, featuring a wide array of common foods such as pancakes, French toast, apple pie, chicken wings, Greek salad, pizza, pork chops, steak, and many others. To maintain uniformity, smaller photos were carefully deleted from the collection, and all images were standardized to have a minimum dimension of 512 pixels. Consequently, they successfully curated 101,000 images representing 101 distinct food classes, with each class containing precisely 1000 images. For the purpose of model development and evaluation, 250 of the 1000 images from each class were set aside for testing and the remaining 750 for training. Figure 1 depicts a preview of the Food-101 dataset.

The MA Food-121 dataset, compiled by Aguilar et al. [4] was also used. It consists of 21,175 food images representing 121 dishes from 11 popular cuisines worldwide. The dataset includes three groups: dish, cuisine, and categories (food groups). Each food item in the dataset is affiliated with at least one of the ten food categories, including Meat, Bread, Fried food, Vegetable, Dumpling, Rice, Seafood, Egg, Soup, and Noodles/Pasta. The dataset provides single-label annotations for both dish and cuisine tasks and allows for the potential of multi-label annotations specifically for categories.

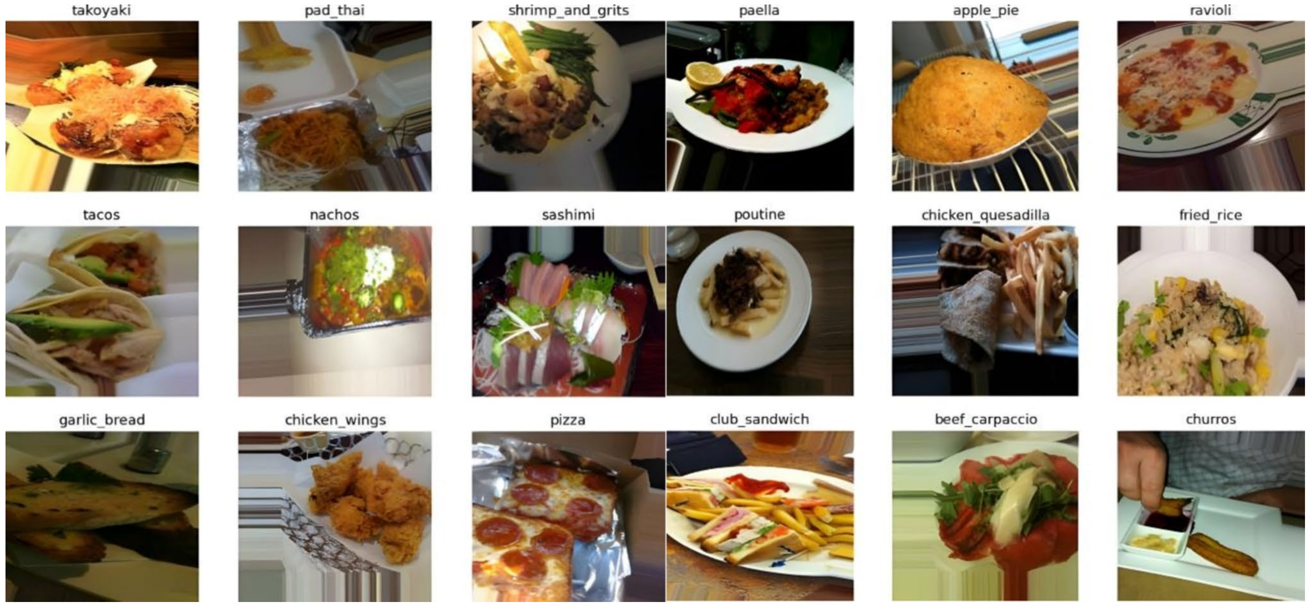

Fig. 1 Food-101 dataset preview

This dataset aids in the development and evaluation of precise food recognition models and is a useful resource for food classification research. Figure 2 depicts a preview of the MA Food-121 dataset.

3.2 CNN Architecture

CNNs, which are classified as Deep Artificial Neural Networks, are commonly used for recognizing and detecting items. They function especially well with input that has a grid structure, such as images [38]. They have similarities with Artificial Neural Networks (ANNs). CNNs organize their parts, called nodes or neurons, in layers. Each layer's output is sent to the next layer for more processing. CNNs also use a learning technique called back-propagation, much like ANNs. This involves adjusting the weights based on how far off the predictions are, similar to how ANNs use a loss function to measure the difference between what's expected and what's predicted during training. This way, back-propagation reduces the overall error in the predictions.

The three types of layers that are frequently found in CNNs are the convolution layer, pooling layer, and fully connected layer, as shown in Fig. 3 below, which depicts the overall CNN architecture. The fully connected layer handles classification, while the convolution and pooling layers extract features.

The convolution layer serves as a foundational component within the CNN architecture, facilitating a sequence of mathematical operations for feature extraction. Convolution, a linear operation, leverages a kernel to extract features. In this

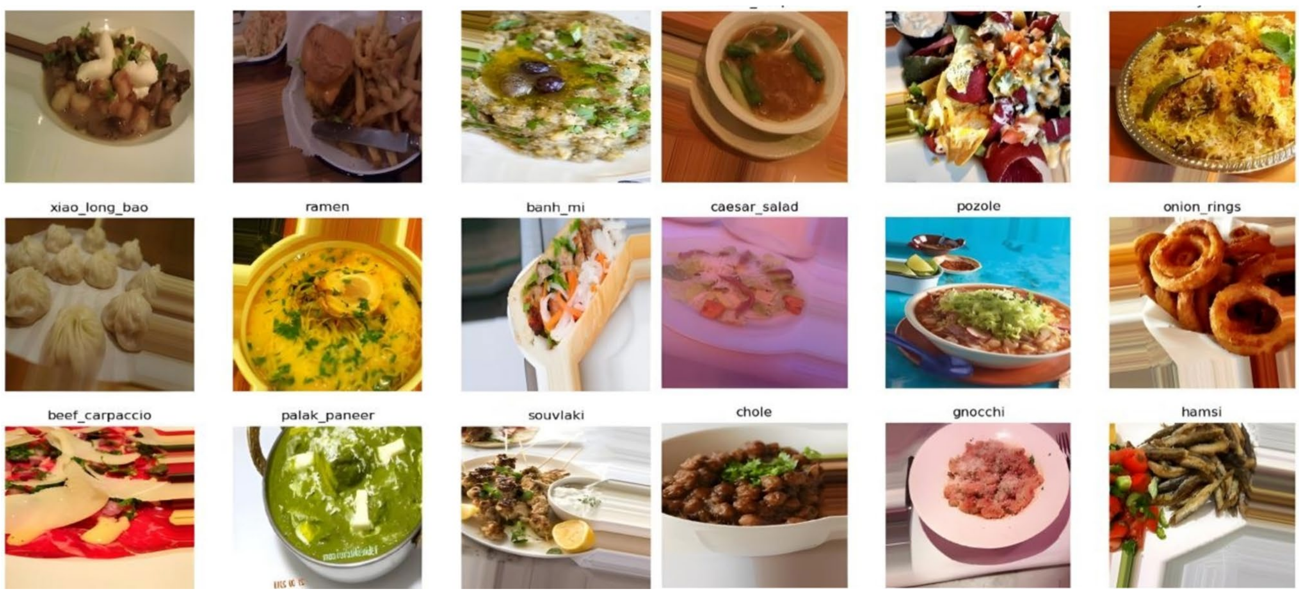

Fig. 2 MA Food-121 dataset preview

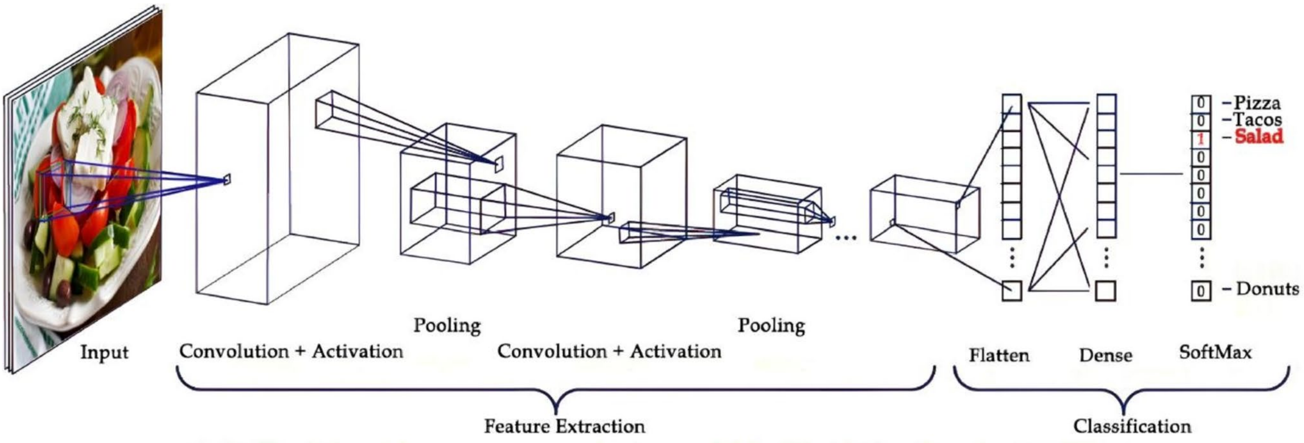

Fig. 3 CNN architecture. Source [20]

process, the kernel is applied to an input array of numbers, often referred to as a tensor, with the goal of producing a feature map. The feature map emerges through the computation of dot products between each kernel element and the input tensor, with the results summed to generate the output. The convolution operation's effectiveness is notably influenced by the depth of the output feature maps, which corresponds to the number of kernels, as well as the kernel's size, typically denoted as 7×7, 5×5, or 3×3. According to Yamashita et al. [37], these parameters are essential in determining how the convolution process is shaped. After the linear convolution process, the output is passed through activation functions such the sigmoid, rectified linear unit (ReLU), and hyperbolic tangent (tanh) functions that add non-linearity to the activation map. Specifically, the ReLU function is widely employed due to its significant contribution to the resolution of the vanishing gradient problem, which was a significant advance in the field of deep learning [22]. The RELU (Rectified Linear Unit) function introduces non-linearity to the model by evaluating the function $f(x) = max(0, x)$. This non-linear activation function ensures that the subsequent layer receives nodes with positive activation values only. When the input value, x is negative, the RELU function outputs 0, effectively preventing the neuron from being activated.

The feature map, which is the output of the convolutional layer, is the input used by the pooling layer that follows. Pooling is utilized to minimize the size of feature maps while retaining essential information, thereby conserving computational resources and expediting the training process [1]. Unlike the convolution operation, the pooling layer does not possess learnable parameters. However, it does involve hyperparameters such as the stride, filter size, and padding, as detailed by Yamashita et al. [37]. Striking a balance between efficiency and accuracy is essential, as larger strides or filter sizes can result in information loss [28]. Three prevalent types of pooling commonly used in pooling operations include Max Pooling, Sum Pooling, and Global Average Pooling, as depicted in Fig. 4. Max pooling stands out as the most popular choice for down-sampling feature maps because it maintains their depth dimension [11]. In the typical application of max pooling, patches are selected from the input feature maps, and only the maximum value within each patch is kept, while other values are discarded. It is common to use a 2×2 filter with a stride of 2, which effectively reduces the in-plane dimension of feature maps by a factor of 2. This approach is widely adopted for downsizing feature maps in practical applications [37].

Typically, the convolutional (or pooling) layer's final output feature maps are flattened and converted into a one-dimensional (1D) array. This 1D array is then used as the input for the fully connected layers, where it is processed to produce the final outputs.

The fully connected layer, often referred to as the dense layer, comprises learnable weights connecting every input node to each output node. The number of output nodes in this last dense layer in multi-class classification tasks corresponds to the total number of classes that need to be classified. To derive probabilities for each target class from the real-valued outputs of the last dense layer, the SoftMax function is commonly employed. This function normalizes the values to a range between 0 and 1, ensuring that the sum of all values equals 1. The SoftMax function possesses several key properties: firstly, it is a smooth and differentiable function, making it well-suited for utilization in optimization algorithms reliant on gradients. Secondly, it effectively maps arbitrary real-valued vectors to probability vectors, proving beneficial in generating outputs for classification tasks. Lastly, the function remains scale-invariant, consistently producing the same output irrespective of the input vector's magnitude. This property is valuable for normalizing the network's output and preventing potential dominance by large input values [25].

The SoftMax function is

$$\text{softmax}\sigma(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_i}} \tag{1}$$

Each element of the input vector is represented by x_i in the equation above, and η is the total number of elements in the vector. The softmax function modifies the output vector to ensure that the total of all elements equals one after calculating the exponentials of each element in the input vector. With this modification, the vector becomes a set of values with a range of 0 to 1, making it interpretable as probabilities (F).

The learning of network parameters (θ) commences after determining the output signals of the CNN. This learning process involves minimizing a loss function calculated at the CNN's output. To achieve this, training examples containing input–output pairs $\left\{ \left(x^{(i)}, y^{(i)} \right); i \in [1, .., N] \right\}$ are utilized. The objective is to adjust the parameters (θ) through an iterative process, using these examples, aiming

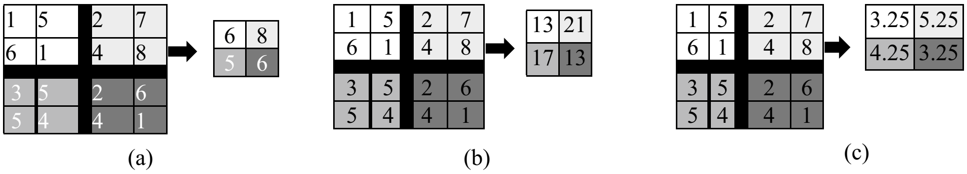

Fig. 4 Pooling operation **a** max pooling, **b** sum pooling, **c** global average pooling)

to minimize the loss function and obtain an optimal configuration for the network.

The loss function is given by:

$$L = \frac{1}{N}\sum_{i=1}^{N} l(\theta; y^{(i)}, o^{(i)}) \tag{2}$$

where the signals denoted by o^{i} and y_i are, respectively, the current output and target output. The loss function is used to determine the unknown parameters θ.

Section 4.1 provides a comprehensive overview of the CNN architecture used in the FRCNNSAM model, including its components such as convolutional layers, max pooling operations, and the SoftMax activation function. A visual representation an high-level overview of the model is presented in Fig. 6 for better clarity.

3.3 Ensemble Learning

Ensemble learning stands as a powerful technique in machine learning, centered around combining multiple models. The core idea driving ensemble learning is that the ensemble, through the integration of predictions from various diverse models, often outperforms the capabilities of any single model [5]. Ensemble learning can take various forms, including Bagging, Boosting, and Stacking [16].

Bagging is the process of training several instances of the same model using randomly selected subsets of the training data with replacement. The final prediction is typically obtained by averaging or voting on the predictions of these individual models. In contrast, Boosting is another widely embraced ensemble method. In Boosting, base models are trained consecutively, with each model focusing on correcting errors made by its predecessor. The final prediction is a weighted combination of all models' predictions. Additionally, Stacking entails training multiple base models and using their predictions as input for a higher-level model, often referred to as the meta-model or aggregator. The role of the meta-model is to learn how to combine the outputs of the base models to generate the ultimate prediction.

Ensemble learning is a versatile technique extensively used across domains, including recommendation systems, computer vision, and, natural language processing. In the context of food image recognition, ensemble learning has demonstrated significant potential in enhancing classification accuracy and robustness. Several studies have underscored the significance of ensemble models in achieving improved recognition outcomes [13, 30]. In a food recognition-based task, for example, Fakhrou et al. [13] discovered that their ensemble model performed better than the most advanced CNN models, obtaining high accuracy.

Motivated by the success of ensemble learning in food image recognition, the FRCNNSAM model adopts a similar strategy to that applied by Fakhrou et al. [13]. However, as one of the primary aims of this research is to explore the performance of CNN models built from scratch without utilizing transfer learning techniques, we devised three distinct models by fine-tuning various parameters of the Image data generator. These parameters include RotationRange, WidthShiftRange, HeightShiftRange, ShearRange, ZoomRange, HorizontalFlip, and FillMode. Additionally, we adjusted the architecture of the FRCNNSAM model by exploring diverse layer and dense layer setups. To build the FRCNNSAM ensemble model, we first obtained predicted class probabilities from each individual model. Then, using an averaging technique, we combined these probabilities to create the ensemble model. This method allows us to capitalize on the capabilities of each unique model while capturing a variety of features and patterns from the input data. By combining multiple models, the FRCNNSAM ensemble model exhibits improved accuracy and robustness in food image recognition tasks.

The mathematical formular for the ensemble model by averaging used for combining the prediction probability of the FRCNNSAM model is given below:

$$\frac{1}{n}\sum_{i=1}^{n} p_i \tag{3}$$

where n represents the total number of models in the ensemble, i represents the index of an individual model in the ensemble, ranging from 1 to n and p_i signifies the prediction made by the i-th model for a specific input.

3.4 Self-Attention Mechanism

Self-attention mechanism, also known as scaled dot-product attention, is a mechanism used in deep learning models to capture long-range dependencies and relationships within input sequences, such as sentences or images. It was originally introduced in the context of natural language processing but has since been successfully applied to various other tasks, including computer vision.

Self-attention has been effectively used in computer vision applications for tasks like segmentation and image recognition. By incorporating self-attention into CNN, the models can effectively capture global dependencies in an image, making them more robust to variations in scale, rotation, and other transformations (Fig. 5).

The FRCNNSAM model takes advantage of the self-attention mechanism to enhance its feature representation capabilities. This mechanism allows the model to focus on important relationships within the input data, capturing fine-grained details and contextual information. In the FRCNNSAM model, the self-attention layer was strategically integrated before the flattening step. This placement enabled

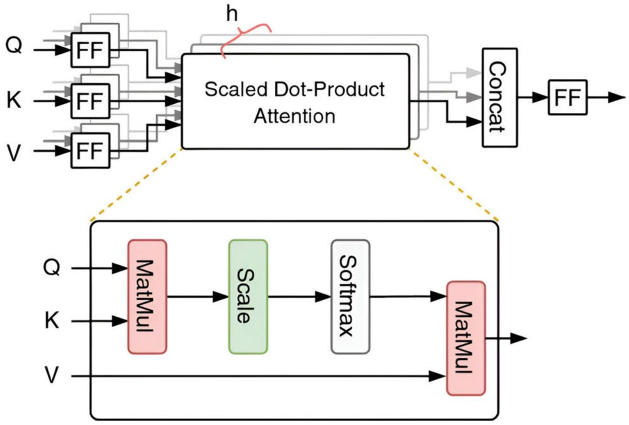

Fig. 5 Scaled dot-product attention mechanism. Source [7]

the model to assign varying weights to different regions of food images, emphasizing their relevance to the classification task. By attending to crucial regions, the model gained the ability to effectively extract discriminative features.

The mechanism involves queries (Q), keys (K), and values (V), which are transformed through feed-forward layers (FF) to obtain representations. The mechanism computes the dot product between queries and keys, scaled by the square root of the key dimension (d_k), to stabilize gradients during training. The resulting scores pass through a SoftMax function to generate attention weights. These weights are applied to the values to compute a weighted sum, representing the attended output. The mechanism enabled the model to focus on relevant parts of the input sequence, capturing dependencies and extracting meaningful representations. Overall, it is a critical element influencing the model's learning and predictive accuracy.

4 Model Design, Simulation and Results

4.1 Model Architecture

The FRCNNSAM design follows the standard CNN architecture as discussed in Sect. 3.2. A layer in the FRCNNSAM model comprises a sequence of components, which include a convolution layer, batch normalization, another convolution layer, batch normalization, and a max-pooling layer. After the final max-pooling layer, the subsequent components involve a self-attention mechanism, flattening, a dense layer, batch normalization, a dropout layer, fully-connected layers, and the final classification layer. The self-attention mechanism used for the development of the FRCNNSAM model was discussed in Sect. 3.4 of this paper. Figure 6 depicts the high-level overview of the FRCNNSAM model.

The FRCNNSAM model was developed by adapting the general structure depicted in Fig. 6. To construct the ensemble model, as described in Sect. 3.3, we made specific adjustments to the layers. Drawing inspiration from the experiment conducted by Josephine et al. [19], where they explored various configurations of CNN models with different numbers of dense layers, we developed three distinct models with varying architectures. Our first model

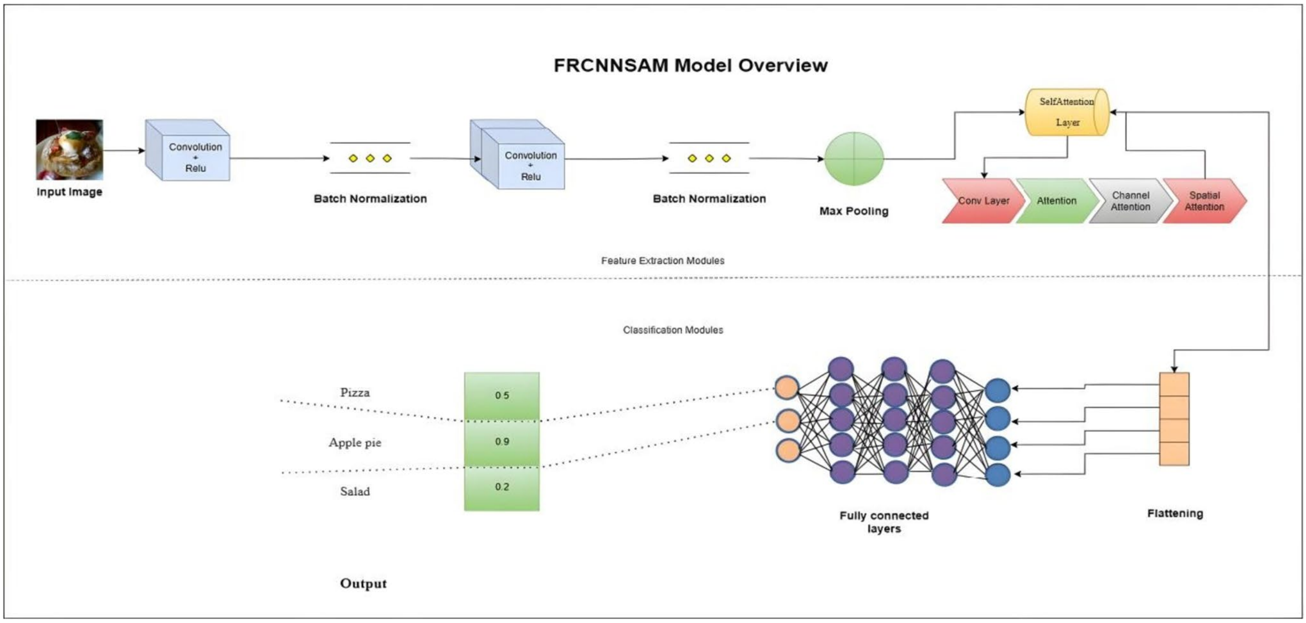

Fig. 6 The FRCNNSAM high-level model overview

comprised six dense layers with values of 1024, 512, 256, 128, 101, and 101 neurons, respectively, the second had four dense layers with values of 512, 256, 128, and 101 neurons, and the third contained five dense layers with values of 512, 256, 128, 101, and 101 neurons. Although the overall architecture of these models resembled the structure shown in Fig. 6, we introduced variations in the number of layers. Consequently, the first model consisted of five layers, the second had four layers, and the third model utilized six layers. The results, as presented in Sect. 4.5, revealed that the first model outperformed the other two, which aligned with the findings of Josephine et al. [19] who also observed better performance from models with six dense layers.

In the convolution layer, we utilized 32 filters for all convolutions in the first layer, with subsequent layers having initial filters of 64 and 128, respectively where the filter size doubled in each layer. The filter size was set to 3 × 3, and the padding was specified as 'same.' ReLU activation function was applied, with the kernel initializer 'he_normal.' The pooling layer employed a pooling size of 2 × 2. In the dense layer, the number of neurons was specified as stated above and ReLU activation was applied for most layers, except for the final layer where softmax activation was used. We also incorporated l_2 regularization in the last two dense layers to prevent overfitting. For optimization, we employed the Adam optimizer, which dynamically adjusted the learning rate during training. The batch size was set to 32, and to enhance the model's generalization capabilities, dropout with a rate of 0.4 was also applied.

The FRCNNSAM model was meticulously designed, drawing insights from related research and experiments, to create an ensemble model with optimized architectural configurations that prioritize higher accuracy and robustness in food image recognition. In contrast to the structure proposed by Özsert Yiğit & Özyildirim [29], the FRCNNSAM model introduced several key enhancements. It incorporated a BatchNormalization layer after each convolutional layer, strategically placed a self-attention mechanism, and included additional dense layers. The FRCNNSAM model also employed stacked convolutional layers, with the convolution filter size doubling at each subsequent layer. The empirical evaluation results, as presented in Sect. 4.5, provide compelling evidence of the exceptional performance of the FRCNNSAM model compared to the previously proposed structure by Özsert Yiğit & Özyildirim [29]. The FRCNNSAM's ability to achieve higher accuracy and robustness in food image recognition can be attributed to the thoughtful integration of advanced techniques and architectural adjustments.

4.2 Data Preprocessing

Before the training of the FRCNNSAM model, images coming from the datasets were pre-processed by resizing and augmenting them. Resizing images is a crucial step in deep learning as it enables the model to learn more effectively from small-sized images. Additionally, data augmentation was implemented on the dataset, employing techniques such as shearing, rotation, or flipping to generate new iterations of the images. This process is done to create a more diverse and robust dataset that can mitigate overfitting and enhance the accuracy and robustness of the model [21].

To streamline the handling of image datasets and create a diverse training dataset, we utilized the Keras ImageDataGenerator class. This powerful tool from the Keras library facilitates the generation of image data batches and offers a wide range of options for pre-processing and data augmentation, making it particularly effective for training deep learning models, including CNN (R. H. [3]. A summary of the data augmentation parameters used with the image data generator can be found in Table 1. While constructing the three different models, we fine-tuned these parameters to achieve improved accuracy. Although some values were altered to optimize performance, the values presented in Table 1 were ultimately found to yield superior results.

4.3 Overfitting and Underfitting

In deep learning, two common challenges are overfitting and underfitting. Overfitting arises when a model is excessively fine-tuned to a limited dataset, causing it to struggle with generalizing to new data. In these situations, the model performs well on training data but poorly on unknown data. On the other hand, underfitting happens when a model performs poorly on both the training data and unseen data due to insufficient training on the available data. To mitigate the risks of underfitting and overfitting, it is crucial to utilize a sufficiently large training dataset and employ strategies like early stopping and regularization to prevent the model from becoming overly complex. Additionally, evaluating the model on a held-out dataset is imperative to ascertain its capacity to generalize to new data [2, 31]).

Table 1 ImageDataGenerator parameters

Image data generator table	
Parameters	Values
Rescale	1/0.255
Rotation range	20
Height shift Range	0.2
Width shift range	0.2
Zoom range	0.3
Shear range	0.2

To address overfitting and underfitting in the FRCNNSAM model, several methods were employed:

Data augmentation was applied to introduce random transformations to the existing training data, effectively generating additional training samples. This strategy aids in preventing overfitting by providing the model with a more extensive and varied dataset to learn from.

Regularization, a strategy that enforces a penalty on the model's objective function during training, is renowned for curbing model complexity and mitigating overfitting to the training data [14]. Within the context of the FRCNNSAM model, regularization was implemented in two key steps. Firstly, weight initialization was introduced, carefully selecting initial values for the neural network's layer weights to maintain activation variation across layers. Secondly, L2 regularization, also referred to as weight decay, was employed. This approach introduces constraints on the weights of the network through an additional term in the loss function, penalizing weights that are too large. By encouraging the network to learn fewer impactful weights, the risk of overfitting is mitigated, consequently enhancing the model's overall performance. The L2 regularization term is included in the loss function to achieve this goal.

$$loss = original_{loss} + \lambda * sum(w^2) \tag{4}$$

The term 'original_loss' in Eq. (4) above denotes the loss that the network would experience in the absence of regularization. 'w' refers to a vector that contains the weights in the network, and 'λ' is a tuning parameter that regulates the regularization strength. A higher value for 'λ' intensifies regularization, whereas a lower value reduces its impact. While L2 regularization is effective at countering overfitting, it may also introduce challenges during training. This is because the regularization term can add complexity to the loss function, making it more intricate to optimize. Hence, the careful selection of 'λ' is imperative, and experimentation with different values is advisable to determine the most suitable choice for a given scenario.

4.4 Model Evaluation Metrics

The process of model validation involves dividing the available data into two separate sets, namely the test set and training set. The training set is used to train the model, and the test set is used to assess its performance. In this case, a 75%–25% split was utilized to train and test the FRCNNSAM model respectively. To evaluate the model's performance, various metrics such as accuracy, precision, recall, and f1-score were utilized.

The formulas presented below define accuracy, predictions, recall, and F1 score:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$f1 = \frac{2(precision * recall)}{Precision + recall} \tag{8}$$

TN (True Negatives): Denotes the count of negative samples correctly classified by the model.

TP (True Positives): Represents the count of positive samples accurately classified by the model.

FN (False Negatives): Indicates the count of negative samples inaccurately classified by the model.

FP (False Positives): Signifies the count of positive samples inaccurately classified by the model.

4.5 Simulation & Result

The FRCNNSAM model was trained on a high-performance system, equipped with an Intel® Core™ i9-9900 k processor running at 3.60 GHz, 32 GB of RAM, and operated on a 64-bit operating system. This system was optimized for machine learning tasks and featured Nvidia compatibility, including the powerful Nvidia GeForce RTX 2080 Ti graphics card. Renowned for its robust performance within NVIDIA's lineup, the graphics card boasted 11 GB of GDDR6 memory and 4352 CUDA cores, making it exceptionally well-suited for handling complex graphical and computational operations, particularly in machine learning applications. The development environment for the FRCNNSAM model was established using Anaconda, providing a streamlined and efficient platform for model development. During the training phase, the FRCNNSAM model demonstrated impressive efficiency, with training times ranging from a minimum of 942 s to a maximum of 966 s. This optimized training environment significantly accelerated the learning process for the FRCNNSAM model, ensuring effective and timely model development.

In this study, we evaluated the performance of the FRCNNSAM ensemble model by training three distinct models with varying configurations of dense layers and layer parameters, as detailed in Sect. 4.1. For the training process, we utilized two datasets: the Food-101 dataset and the MA_Food-121 dataset, as described in Sect. 3.1.

Each model underwent training for 200 epochs, and the results are as follows:

- The first model achieved an accuracy of 94.11% on the Food-101 dataset.
- The second model attained an accuracy of 93.85% on the Food-101 dataset.
- The third model showcased a notable accuracy of 91.48% on the Food-101 dataset.

Although, the base model exhibited commendable performance across various metrics. However, with the strategic integration of a self-attention mechanism, notable enhancements were observed across all metrics, particularly in model generalization capabilities. This augmentation underscored the discernible impact of leveraging self-attention mechanisms to further refine and optimize the model's predictive capacity.

To develop the FRCNNSAM ensemble model, we employed the methodology described in Sect. 3.3, which involved obtaining the predicted class probabilities from each individual model. By averaging these probabilities, we constructed the ensemble model, which demonstrated exceptional performance on the Food-101 dataset, achieving an accuracy of 96.40%.

Additionally, we applied the same approach to build an ensemble model using the MA Food-121 dataset, resulting in a commendable accuracy of 95.11%. In order to guarantee a thorough assessment, both datasets were split into subsets for training and validation, with 75% of the images going to training and the remaining 25% going to validation. This approach allowed us to assess the models' generalization capabilities on previously unseen data, providing reliable and robust results.

The outcomes of our study highlight the effectiveness of the FRCNNSAM ensemble model in food image recognition, with strong performance on multiple datasets. These results underscore the potential of ensemble learning to enhance the recognition capabilities of CNN-based models and demonstrate the significance of our proposed model to the food image classification task.

The visual representation of our results is presented in the figures below, which include the loss and validation plots, as well as the model test results plot (Figs. 7, 8, 9).

To evaluate the FRCNNSAM ensemble model's capacity to generalize to unfamiliar data, a selection of random food images from both the Food-101 and MA Food-121 datasets was sourced from the internet. The test results are presented below.

The simulation results we obtained indicate that the FRCNNSAM model excelled in the classification and identification of food images, demonstrating its suitability for various food image recognition tasks.

4.6 Model Comparison

As discussed in Sect. 2, various CNN-based architectures have been employed, and Table 2 above showcases the results of some of these studies that utilized the same dataset. Notably, Aguilar et al. [5] acknowledged that using single CNN models and architectures yielded limited results, leading to the adoption of ensemble modeling as a common approach to improve accuracy and robustness.

Ensemble modeling has shown great promise in food image recognition, as demonstrated by Aguilar et al. [5] and Fakhrou et al. [13] with their utilization of transfer learning techniques, employing CNN models like ResNet50, InceptionV3, DenseNet201, and InceptionV3. Additionally, Qui et al. [32] introduced the PAR-NET, a combination of three sub-networks classifying original full input images, images with discriminative regions erased,

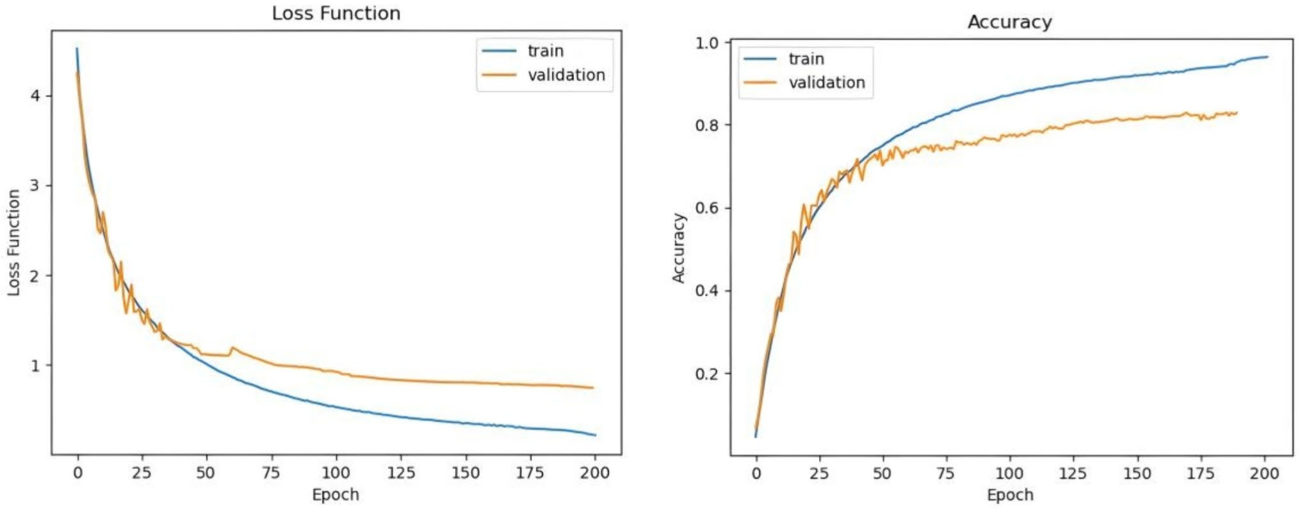

Fig. 7 Food-101 training, validation and loss function plot

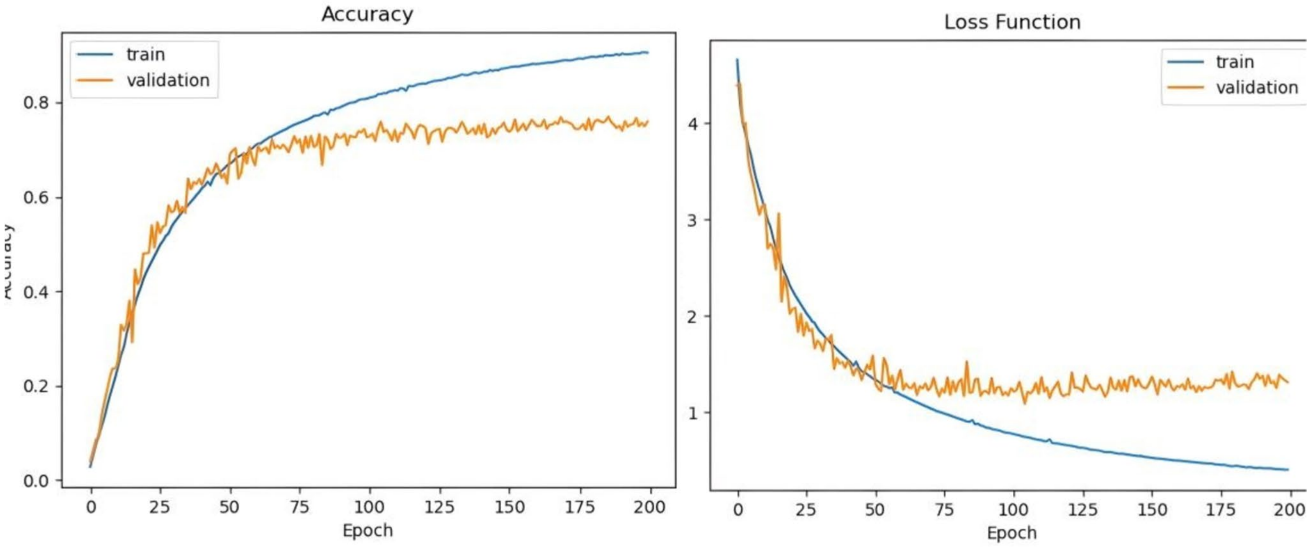

Fig. 8 MA Food-121 training, validation and loss function plot

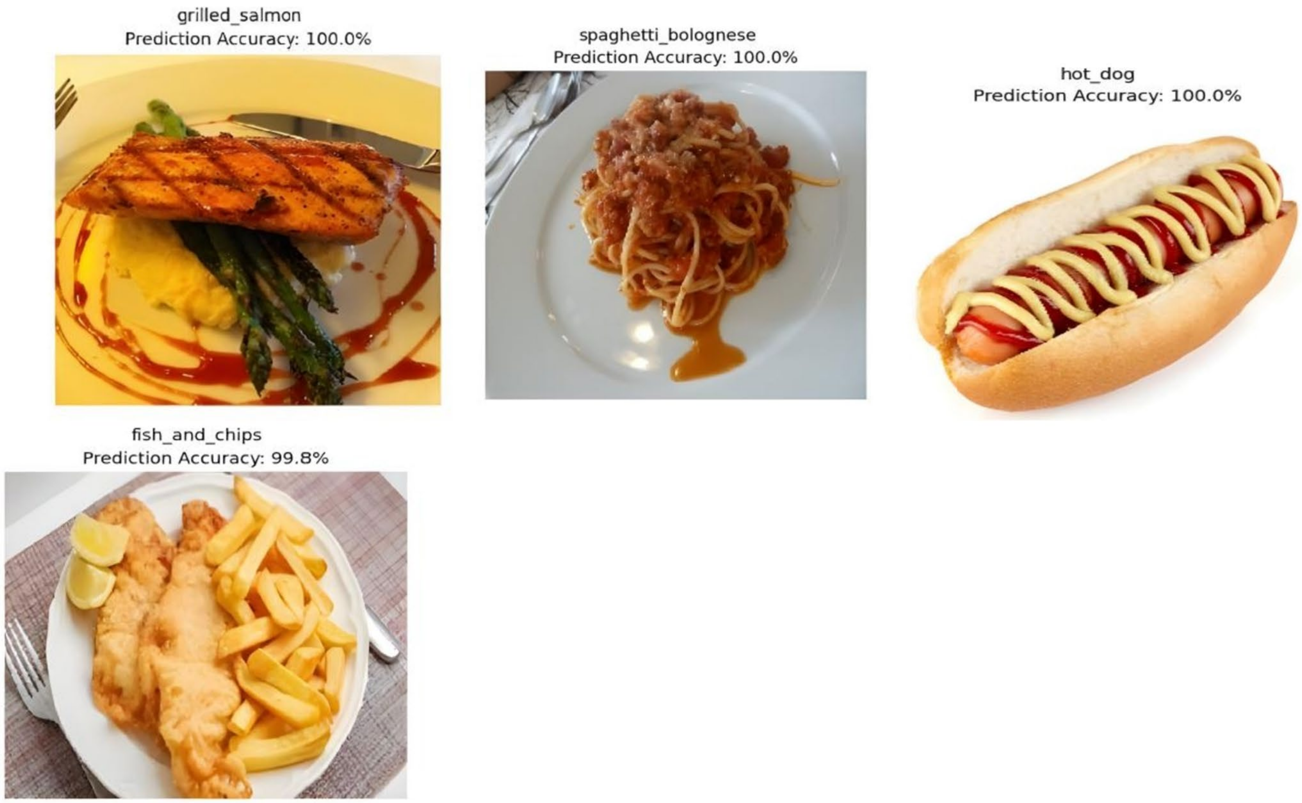
grilled_salmon
Prediction Accuracy: 100.0%
spaghetti_bolognese
Prediction Accuracy: 100.0%
hot_dog
Prediction Accuracy: 100.0%
fish_and_chips
Prediction Accuracy: 99.8%

Fig. 9 FRCNNSAM model test results

and cropped and upsampled discriminative regions. However, the PAR NET's performance limitations stem from the necessity of maintaining independent sub-networks for different tasks.

The results presented in Table 2 showcase the commendable performance of the FRCNNSAM model, surpassing other underlying models that utilized the same dataset. The FRCNNSAM model incorporates advanced techniques, such

Table 2 FRCNNSAM model comparison with other models

Model comparison					
Datasets	Research works	Method	Accuracy (%)	F1 Score	Precision
Food-101	Bossard et al. [10]	Random forest	50.76	NA	NA
	Özsert Yiğit & Özyildirim [29]	Convolution neural network (CNN)	73.80	NA	NA
	Liu et al. [22]	GoogleNet	77.40	NA	NA
	VijayaKumari et al. [35]	EfficientNetB0	80.16	81%	83%
	Attokaren et al. [8]	Inception V3	86.97	NA	NA
	Hassannejad et al. [18]	Inception V3	88.28	NA	NA
	Qui et al. [32]	PAR-NET	90.4	NA	NA
	Current paper	FRCNNSAM	96.40	97.0%	97.55%
MAFood-121	Aguilar et al. [4]	RUMTL	83.82	85.02%	86.40%
	Fakhrou et al. [13]	Ensemble	84.95	NA	NA
	Aguilar et al. [5]	ResNet50	83.16	NA	NA
	Aguilar et al. [5]	Inception V3	86.94	NA	NA
	Aguilar et al. [5]	FS_UAMS	88.95	NA	NA
	Current paper	FRCNNSAM	95.11	95.60%	96.04%

as scaled dot-product attention and ensemble modeling, contributing to its high accuracy and efficiency in food recognition. In comparison to previous research, Özsert Yiğit and Özyildirim [29] also developed a CNN model without utilizing transfer learning technique, however the FRCNNSAM model outperformed their proposed model by integrating diverse cutting-edge techniques. The high accuracy of the FRCNNSAM model can be attributed to several factors, including strategic placement of a Batch-normalization layer after the convolutional layer provided stability during training and accelerated convergence, the inclusion of extra dense layers to capture intricate patterns, and the incorporation of a self-attention mechanism to identify meaningful patterns. Over-fitting was prevented and generalization was improved by employing regularization techniques such as weight initialization and l2 regularization.

The integration of the scaled dot product attention mechanism into the model architecture significantly bolstered robustness and generalization capabilities. This observation aligns well with the findings elucidated by Zhao et al. [39], emphasizing the efficacy of attention mechanisms in enhancing model performance and adaptability.

Furthermore, our research recognizes the importance of sophisticated imaging systems, as highlighted by García-Armenta and Gutiérrez-López [17], in understanding food micro-structure. In our research, advanced imaging techniques have been instrumental in comprehending the features and patterns within food images, leading to improved accuracy in the food recognition system. The potential for interdisciplinary collaborations between computer vision researchers and food scientists is evident, where insights from food micro-structure analysis can inform and enhance the development of accurate and robust food recognition systems, paving the way for future research and exploration.

Overall, our research demonstrates that CNN models built without transfer learning can achieve comparable performance to those using transfer learning techniques. Proper fine-tuning, parameter adjustments, and integration of advanced techniques play crucial roles in maximizing the model's accuracy. These results support the notion that ensemble modelling and advanced CNN architectures significantly contribute to enhancing food image recognition systems, enabling accurate dietary assessment and ultimately improving public health.

5 Conclusions & Recommendation

5.1 Conclusion

In this study, we developed an automated food recognition model named FRCNNSAM, utilizing a CNN architecture for its development. The task involved training the model to categorize and identify distinct food categories in images, with potential applications in various domains. The FRCNNSAM model was carefully designed, incorporating insights from related research and experiments to create an effective ensemble model with optimized architectural configurations. Drawing inspiration from the experiment conducted by Josephine et al. [19], we developed three different models, each taking advantage of different CNN architectures. The experiment results aligned with Josephine et al.'s findings, as models with six dense layers demonstrated better performance. Two novel datasets, the food-101 dataset with

101 food classes and 101,000 images and the MA Food-121 dataset with 121 food classes and 21,175 food images, were used to train the FRCNNSAM model. With the Food-101 dataset, the FRCNNSAM ensemble model demonstrated a remarkable accuracy of 96.40%, while with the MA Food-121 dataset, the FRCNNSAM ensemble model demonstrated a respectable accuracy of 95.11%. These results further demonstrate that CNN models built without the use of transfer learning techniques can achieve comparable performance to those utilizing transfer learning, especially when advanced state-of-the-art techniques are integrated. The high accuracy of the proposed FRCNNSAM model can be attributed to the incorporation of diverse advanced cutting-edge techniques. Notably, the model benefited from the inclusion of a self-attention mechanism and ensemble modelling. Data pre-processing using the ImageDataGenerator function was employed to generate additional training data, preventing over-fitting and enhancing generalization capabilities. Additionally, to control model complexity and penalize large weights, regularization strategies including weight initialization and l2 regularization were used, which enhanced performance. These techniques contributed to the model's ability to effectively recognize and categorize different food items.

In conclusion, our study presents an effective FRCNNSAM model for automated food recognition, achieving state-of-the-art accuracy without relying on transfer learning techniques. The integration of advanced techniques, data pre-processing, and regularization played key roles in maximizing the model's accuracy. The results highlight the potential of building CNN models without utilizing transfer learning techniques and highlight the importance of employing advanced methods in the field of food image recognition.

5.2 Recommendation

Based on the research's findings, a number of recommendations can be made for future endeavours to improve the FRCNNSAM model's applicability and performance in automated food recognition tasks. Firstly, it is critical to investigate the use of cutting-edge methodologies beyond scaled dot-product attention and ensemble modelling for heightened model accuracy and robustness. Additionally, considering the use of diverse datasets, a potential future exploration could involve combining the separate datasets used to develop a novel dataset, allowing for a comprehensive evaluation of the ensemble model's performance on a wider range of food classes and variations. Furthermore, to assess the model's generalization capabilities, it is highly recommended to evaluate the FRCNNSAM model on larger and more diverse datasets, providing valuable insights into its real-world performance and its ability to handle a broader set of input variations. Moreover, in practical applications such as dietary assessment tools or food recognition mobile apps, deploying the FRCNNSAM model will allow for a better understanding of its usability and effectiveness in promoting public health through improved nutrition monitoring. Additionally, focusing on the interpretability and explainability of the model is essential. Employing visualization techniques like feature visualization, saliency maps, or model explainability methods such as Grad-CAM will contribute valuable information about the model's decision strategy and enhance the trust and interpretability of its predictions. Incorporating these recommendations into future research endeavours can significantly contribute to advancing the understanding and performance of the FRCNNSAM model, ultimately leading to more robust and accurate results in food image recognition. Furthermore, the potential for interdisciplinary collaborations between computer vision researchers and food scientists is evident. Leveraging insights from food micro-structure analysis to enhance the development of accurate and robust food recognition systems offers intriguing avenues for future research and exploration, further contributing to improved public health and dietary assessment practices.

Acknowledgements We extend our gratitude to the Near East University for providing access to the research facilities at the University, which played a crucial role in facilitating the successful completion of this study.

Author contributions Both authors made equal and significant contributions throughout the research process, from conducting the simulations to writing and presenting the results.

Funding This research was conducted without any external funding or financial support.

Data availability The dataset used for this research is openly available and can be found via the links below: https://datasets.activeloop.ai/docs/ml/datasets/food-101-dataset/, https://www.kaggle.com/datasets/dansbecker/f.ood-101 and also http://www.ub.edu/cvub/mafood121/.

Declarations

Conflict of interest The authors declare that there are no competing interests for this research.

References

1. Abiyev RH, Arslan M. Head mouse control system for people with disabilities. Expert Syst. 2019. https://doi.org/10.1111/exsy.12398.
2. Abiyev RH, Abdullahi I. COVID-19 and pneumonia diagnosis in X-ray images using convolutional neural networks. Math Probl Eng. 2021;2021(1–14):3281135. https://doi.org/10.1155/2021/3281135.
3. Abiyev RH, Adepoju JA. Deep convolutional network for food image identification. Stud Comput Intell. 2023. https://doi.org/10.1007/978-3-031-42924-8_2.
4. Aguilar E, Bolaños M, Radeva P. Regularized uncertainty-based multi-task learning model for food analysis. J Vis Commun Image Represent. 2019;60:360–70. https://doi.org/10.1016/j.jvcir.2019.03.011.
5. Aguilar E, Nagarajan B, Radeva P. Uncertainty-aware selecting for an ensemble of deep food recognition models. Comput Biol Med. 2022;146: 105645. https://doi.org/10.1016/j.compbiomed.2022.105645.
6. Akhi AB, Akter F, Khatun T, Uddin MS. Recognition and classification of fast food images. Global J Comput Sci Technol. 2018;18(1):7–13.
7. Asgari-Chenaghlu M, Feizi-Derakhshi M, Farzinvash L, Balafar MA, Motamed C. CWI: a multimodal deep learning approach for named entity recognition from social media using character, word and image features. Neural Comput Appl. 2021;34(3):1905–22. https://doi.org/10.1007/s00521-021-06488-4.
8. Attokaren DJ, Fernandes IG, Sriram A, Murthy YVS, Koolagudi SG (2017) Food classification from images using convolutional neural networks. TENCON 2017 - 2017 IEEE Region 10 Conference. 2801–2806, https://doi.org/10.1109/tencon.2017.8228338
9. Bishop TR, von Hinke S, Hollingsworth B, Lake AA, Brown H, Burgoine T. Automatic classification of takeaway food outlet cuisine type using machine (deep) learning. Mach Learn Appl. 2021;6: 100106. https://doi.org/10.1016/j.mlwa.2021.100106.
10. Bossard L, Guillaumin M, Van Gool L. Food-101 – mining discriminative components with random forests. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. Computer vision – ECCV 2014. Lecture notes in computer science, vol. 8694. Cham: Springer; 2014. https://doi.org/10.1007/978-3-319-10599-4_29.
11. Bush IJ, Abiyev R, Arslan M. Impact of machine learning techniques on hand gesture recognition. J Intell Fuzzy Syst. 2019;37(3):4241–52. https://doi.org/10.3233/jifs-190353.
12. Csurka G, Dance C, Fan L, Willamowski J, Bray C (2014) Visual categorization with bags of keypoints. In Proc ECCV Workshop on statistical learning in computer vision, 1:59–74, Prague
13. Fakhrou A, Kunhoth J, Al Maadeed S. Smartphone-based food recognition system using multiple deep cnn models. Multimed Tool Appl. 2021;80(21):33011–32.
14. Feinman R, Lake BM (2019) Learning a smooth kernel regularizer for convolutional neural networks. arXiv preprint arXiv:1903.01882. Accessed 10 Nov 2022.
15. Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D. Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell. 2010;32(9):1627–45. https://doi.org/10.1109/tpami.2009.167.
16. Ganaie MA, Hu M, Malik AK, Tanveer M, Suganthan PN. Ensemble deep learning: a review. Eng Appl Artif Intell. 2022;115: 105151.
17. García-Armenta E, Gutiérrez-López GF. Fractal micro-structure of foods. Food Eng Rev. 2022;14(1):1–19. https://doi.org/10.1007/s12393-021-09302-y.
18. Hassannejad H, Matrella G, Ciampolini P, De Munari I, Mordonini M, Cagnoni S (2016) Food image recognition using very deep convolutional networks. Proc. of the 2nd Int. Workshop on Multimedia Assisted Dietary Management, pp 41–49. https://doi.org/10.1145/2986035.2986042
19. Helen Josephine VL, Nirmala A, Alluri VL. Impact of hidden dense layers in convolutional neural network to enhance performance of classification model. IOP Conf Ser: Mater Sci Eng. 2021;1131(1): 012007. https://doi.org/10.1088/1757-899x/1131/1/012007.
20. Kiourt C, Pavlidis G, Markantonatou S. Deep learning approaches in food recognition. machine learning paradigms. Learn Anal Intell Syst. 2020;18:83–108. https://doi.org/10.1007/978-3-030-49724-84.
21. Lashgari E, Liang D, Maoz U. Data augmentation for deep-learning-based electroencephalography. J Neurosci Methods. 2020;346: 108885. https://doi.org/10.1016/j.jneumeth.2020.108885.
22. Liu C, Cao Y, Luo Y, Chen G, Vokkarane V, Ma Y. DeepFood: Deep Learning-Based Food Image Recognition for Computer-Aided Dietary Assessment. In: Chang C, Chiari L, Cao Y, Jin H, Mokhtari M, Aloulou H, editors. Inclusive Smart Cities and Digital Health. ICOST 2016. Lecture notes in computer science, vol. 9677. Cham: Springer; 2016. https://doi.org/10.1007/978-3-319-39601-9_4.
23. Ma P, Zhang Z, Li Y, Yu N, Sheng J, Küçük McGinty H, Wang Q, Ahuja JK. Deep learning accurately predicts food categories and nutrients based on ingredient statements. Food Chem. 2022;391: 133243. https://doi.org/10.1016/j.foodchem.2022.133243.
24. Matsuda Y, Hoashi H, Yanai K (2012) Recognition of multiple-food images by detecting candidate regions. 2012 IEEE International Conference on Multimedia and Expo pp 25–30. https://doi.org/10.1109/icme.2012.157
25. Mikulski B (2019) Understanding the softmax activation function | Bartosz Mikulski. Mikulskibartosz. https://www.mikulskibartosz.name/understanding-the-softmax-activation-function/. Accessed 1 Dec 2022.
26. Mezgec S. The state of the art of automated food recognition. Alternator. 2021. https://doi.org/10.3986/alternator.2021.25.
27. Mishra M (2020) Convolutional neural networks, explained. Towards Data Science. Retrieved November 9, 2022, from https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939. Accessed 9 Nov 2022.
28. Naseri H, Mehrdad V. Novel CNN with investigation on accuracy by modifying stride, padding, kernel size and filter numbers. Multimed Tools Appl. 2023. https://doi.org/10.1007/s11042-023-14603-x.
29. Özsert Yiğit G, Özyildirim BM. Comparison of convolutional neural network models for food image classification. J Inf Telecommun. 2018;2(3):347–57. https://doi.org/10.1080/24751839.2018.1446236.
30. Pandey P, Deepthi A, Mandal B, Puhan NB. FoodNet: recognizing foods using ensemble of deep networks. IEEE Signal Process Lett. 2017;24(12):1758–62. https://doi.org/10.1109/lsp.2017.2758862.
31. Perez L, Wang J (2017) The effectiveness of data augmentation in image classification using deep learning. Computer Vision and Pattern Recognition. arXiv:1712.04621v1, https://doi.org/10.48550/arXiv.1712.04621. Accessed 2 Nov 2022.
32. Qiu J, Lo FPW, Sun Y, Wang S, Lo B (2022) Mining discriminative food regions for accurate food recognition. arXiv preprint arXiv:2207.03692. Accessed 26 Apr 2023.
33. Rane C, Mehrotra R, Bhattacharyya S, Sharma M, Bhattacharya M. A novel attention fusion network-based framework to ensemble the predictions of CNNs for lymph node metastasis detection. J Supercomput. 2020;77(4):4201–20. https://doi.org/10.1007/s11227-020-03432-6.
34. Sánchez J, Perronnin F, Mensink T, Verbeek J. Image classification with the fisher vector: theory and practice. Int J

Comput Vision. 2013;105(3):222–45. https://doi.org/10.1007/s11263-013-0636-x.

35. VijayaKumari G, Priyanka V, Vishwanath P. Food classification using transfer learning technique. Global Trans Proc. 2022;3(1):225–9.
36. Yadav S, Alpana, Chand S (2021) Automated food image classification using deep learning approach. In: 2021 7th international conference on advanced computing and communication systems (ICACCS), 19–20 March 2021, Coimbatore, India. IEEE. https://doi.org/10.1109/icaccs51430.2021.9441889
37. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging. 2018;9(4):611–29. https://doi.org/10.1007/s13244-018-0639-9.
38. Zhao Z, Zheng P, Xu S, Wu X. Object detection with deep learning: a review. IEEE Trans Neural Netw Learn Syst. 2019;30(11):3212–32. https://doi.org/10.1109/tnnls.2018.2876865.
39. Zhao H, Jia J, Koltun V (2020) Exploring self-attention for image recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp 10076–10085)