2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N) | 978-1-6654-7436-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAC3N56670.2022.10074297

Food Detection and Recognition Using Deep Learning – A Review

Anushrie Banerjee
Department. of Data Science
Christ University
Pune, Lavasa, India
anushriebaner@gmail.com

Palak Bansal
Department. of Data Science
Christ University
Pune, Lavasa, India
palakbansal2401@gmail.com

K.T. Thomas
Department. of Data Science
Christ University
Pune, Lavasa, India
thomas.kt@christuniversity.in

***Abstract*— Studies show poor lifestyle choices and unhealthy eating patterns cause issues like obesity and other ongoing illnesses that raise the risk of heart attacks, such as hypertension, abnormal blood sugar levels, and diabetes. To improve this situation a lot of health apps have been built which use modern dietary monitoring systems that automatically evaluate dietary intake using machine learning and deep learning techniques rather. For these reasons in-depth investigations on food detection, classification, and analysis have been conducted. Some of the top methods for automatic food recognition created have been discussed in this paper. We also propose an idea for detection of Indian food items using image classification. According to our findings of the papers we reviewed, convolutional neural networks (CNN) have been extensively been used in food detection as it has been giving better results compared to other models. We also observed that Vision transformers perform better in situations where the dataset is large and a hybrid model would give better accuracy. A review of potential applications for food image analysis, shortfalls in the area, and open issues concludes the paper.**

Keywords*— *Image classification, Food recognition, Vision transformers, Convolutional neural network, ResNet, inception v3 model, Transfer learning.

I. INTRODUCTION

Despite recent medical advances, there are still many people who suffer from chronic ailments. With the advent of technology, we have an easy access to food at our doorstep on the click of a button. Due to the increasing food intake, there is a higher chance that a person would become obese. Overweight and obesity, that was once thought to be an issue mostly in countries with high income, have been increasing in low-income and middle-income countries. Children who are overweight or obese make up the majority of the population in developing nations, where the pace of increase has been more than 30% higher than in industrialised nations [1]. People frequently make notes to track their nutritional consumption in order to manage their severe weight issues. Dieticians then need this information in order to determine a patient's nutritional intake. Numerous mobile e-health tools are available that can assist users in controlling their caloric intake and addressing the problem before it becomes out of control. Numerous apps were developed to calculate food intake in order to avoid the use of erroneous data and manual labor.

In-depth exploration on food detection, categorization, and analysis has been done for a variety of applications related to eating habits and dietary evaluation. For testing and training objectives on the specific topic of calorie assessment with single food items, the scientific community has a constant need for a dataset of visual data. The fact that there are many diverse food varieties, each with a unique hue and feel, reflects how difficult it is to recognize food images. Food products accurately are non-rigid structure and there are intra-source variances since preparation techniques and cooking styles vary from location to region. The issue is worsened by the wide range of food products with low inter- and large intra-class differences and the scant information in a single image. Additionally, many recipes conceal a number of components, which might make algorithms for classifying food items more difficult to use.

Amidst these difficulties, many photos of food have unique qualities that assist in separating one meal kind from another. Several techniques for identifying food utilize handcrafted features like shape, colour, texture, and position. Some of these techniques combine handmade and deep visual features. Suitable selection of attributes is crucial for reducing unnecessary features in order to improve classification performance and reduce computational complexity.

Due to the recent development in machine learning and deep learning areas single food item detection has become easier but still the challenge of multiple food detection exists. There are several methods that can be applied to classify images include used in [2] linear SVM classifier, using different bag-of-features 11 distinct classes were categorized [3], among others. But the most commonly used methodology is CNN as it provides better results on large datasets and achieves higher performance accuracy. The experimental analysis carried out in [4] on a number of benchmarking datasets shows that a neural network without a convolutional layer, like Vision Transformer, is capable of shaping a global descriptor and achieving competitive results. The presented methodology's simplicity fosters acceptance of the architecture as an image retrieval baseline model because fine-tuning is not necessary, displacing the conventional and widely used CNN-based approaches and ushering in a new era in image retrieval techniques.

One of the most well-liked deep neural networks is the convolutional neural network (CNN). CNN is made up of convolutional, non-linear, pooling, and full connected layer. In particular, the results from applications connected to images, natural language processing (NLP) and computer vision are simply astonishing. [5].

Transformers rely on a straightforward yet effective process called self-attention, which concentrates on particular input components to produce more effective results. In particular, natural language processing (NLP) techniques like machine translation, language modelling, and speech recognition are currently regarded as state-of-the-art models for sequential data [6]. In order to create long-range contextual relationships between pixels in images, they instead use multiheaded attention mechanisms as the primary building component. The pictures are first divided into patches, which are then flattened and embedded to create sequences. These patches have embedding position added in order to preserve positional information. We get the final output by feeding the resulting sequence into a number of multiheaded attention layers.

This work reviews the research to assess the utility of several models for categorizing food photos, and it discusses potential future experiments in section III that could improve the accuracy of food item recognition.

II. Related works

Numerous Deep Learning techniques have been utilised extensively in the past decade to recognise food. This has shown to be quite effective in several circumstances and has produced incredibly precise outcomes. For identifying the food dishes two types of methodologies employed by classification systems were classical techniques and deep learning strategies. Using a feature extraction methodology, the visual picture features of food plate images are retrieved and expressed as feature vectors in the standard way.

Methods for food recognition using statistics that we could find is Yang et al. [7], (2010) who use pairwise statistics between local features. In the paper, a method for representing food items that computes pairwise statistics between local characteristics after softly segmenting the image into eight different sort of ingredients at the pixel level. They have then compiled the statistics into a multidimensional histogram, and a discriminative classifier used the histogram as a feature vector.

A similar approach was applied by Kawano et al. [2], (2014). In their investigation, they suggested a real-time remedy for food photo detection on mobile devices. GrabCut was used to segment the image's region, and two different types of information were used to extract picture characteristics: one method combined the standard bag-of-features and colour histograms with two kernel feature maps, and the other method used a HOG patch descriptor and a colour patch descriptor with Fisher Vector representation. Finally, they classified it using a linear SVM. In order to encourage a user to move a smartphone camera, the system predicts the direction of food sections where the higher SVM output score is projected to be made. In their trials, they were able to get a classification rate of 79.2 percent for the top 5 category choices from a dataset of 100 distinct food categories using colour patches, HOG with the Fisher Vector coding as picture features.

As mentioned before a lot of work is done is classifying food images using deep learning. Yu et al. [8], (2016) on the food detection task suggested a CNN-based food detection algorithm. Based on the Inception-ResNet and Inception V3 models, the complete architecture was transferred learned and fine-tuned. They used the Food-101 dataset. They discovered that Inception-ResNet achieves best classification accuracy of 72.55 percent and top-5 accuracy of 91.31 percent, converging far more quickly.

The success of a solution for automatic food recognition utilizing a smartphone camera in the actual world is dependent on a number of variables. In the paper "Smartphone-based food recognition system using multiple deep CNN model" [9], Fakhrou et al. (2021) has demonstrated a smartphone-based method that trains a convolutional Network to identify fruit and food on platters for children who are blind or visually impaired. The deep CNN model was developed using the ensemble learning technique. Additionally, they assessed the effectiveness of many convolutional Network models for recognizing food images with the help of the transfer learning technique. They discovered that the ensemble model beat state-of-the-art Convolutional Neural Networks algorithms in the customized food dataset and achieved a 95.55% accuracy rate for identifying foods. To demonstrate the effectiveness of the suggested Deep Convolutional Neural model for food recognition tasks, two publicly accessible food datasets are used in further evaluation.

Zahisham, Lee et al. [10], (2020) presented a paper based on ResNet 50 architecture a Deep Convolutional Neural Network (DCNN). The ResNet model was replicated and the pre-trained weights were imported. The model's final layers were trained using three online-obtained datasets. This process is referred to as fine-tuning a trained model. The datasets ETHZ-FOOD101, UECFOOD100, and UECFOOD256 were used to assess the model's performance. The parameter settings and outcomes of the suggested approach were also included in this study. Their experimental results showed that their suggested technique outperformed the other methods in comparison, and it attained accuracy values of 41.08%, 39.75%, and 35.32% for the ETHZ FOOD-101, UECFOOD100, and UECFOOD256 datasets, respectively.

"Indian Food Image Classification with Transfer Learning" [11], (2019) is a further paper that applied the transfer learning methodology. Rajayogi et al. have carried out image classification using a variety of transfer learning techniques. Pre-trained models were employed in this research, which improves results while reducing costs and improves computational efficiency. For training and

verification, the dataset of Indian cuisine was utilized, which consists of 20 different classes with 500 photos each. Used models included ResNet, VGG19, IncceptionV3, and VGG16. Test results revealed that Google InceptionV3 which had a loss rate of 0.5893 and an accuracy of 87.9%, outperformed other models.

Ramesh et al. [12], (2021) developed a program that can be used as a solo or connected programming interface that can automatically locate food products in immediate circumstances. A Single Shot Detector (SSD) setup was trained using a dataset assembled from several web sources. The most effective method was found to be the combination of a Single Shot Detector with the Google InceptionV2 and CNN architecture. When the predictions in blind pictures are known, the localization of the frame on the image by the SSD is far more accurate than the metrics would suggest. For 1900 of the anticipated 2000 images, the class classification confidence was greater than 80% and above 95% for 1750 of the tested 2000 images. For every incorrect or erroneous prediction made during testing, there were 250 right guesses. Knowing that, the model's accuracy may be calculated to be 97.6%.

The new Vision Transformer approach, which in some situations is thought to be superior to the most widely used CNN technology was employed by Nijhawan et el. in "Food classification of Indian cuisines using handcrafted features and vision transformer network" [13], (2021). He employed a hybrid vision Transformer food categorization system that combines hand-crafted features with the Vision Transformer ensemble architecture. This method classifies images using patches of the picture and a Transformer-like design. The methodology involves combining two classifiers. The first one uses the vision transformer approach to identify the most important properties required for the training and testing of the picture dataset. The second classifier takes the information from the image itself using manually crafted features that are then applied to our dataset. These manually created characteristics include Local Binary Pattern (LBP), GIST, and HoG (Histogram of Oriented Gradients). The hybrid system was ultimately shown to be more successful than any of the separate models at improving the system's overall efficiency and accuracy. On CNN, a number of tests were carried out using different classifiers, including Random Forest and SVM. They also contrasted various approaches against a number of CNN architectural ensembles utilising various feature extraction techniques. The paper compared several algorithms on the same dataset and table below shows the results:

TABLE I. EVALUATION OF NUMEROUS EXPERIMENTAL SITUATIONS FOR ACCURACY KAPPA COEFFICIENT, SENSITIVITY, SPECIFICITY, AND ACCURACY (AC, SEN, SPE, K.C) [13]

Classifier	AC	Sen	Spe	K.C
RF	62.14%	65.43%	68.99%	0.62
SVM	70.76%	65.54%	71.32%	0.67
KNN	51.31%	50.62%	54.66%	0.51
CNN + SVM	74.21%	73.32%	75.42%	0.74
S1 (CNN-2L + RF)	75.54%	72.65%	75.43%	0.77
S2 (CNN-3L + SVM)	78.95%	77.93%	83.76%	0.79
S3 (CNN-4L + SVM)	79.21%	82.98%	84.21%	0.79
S4 (CNN-5L +RF)	83.11%	81.45%	87.87%	0.82
S5 (CNN-(4L + LBP, GIST and HoG)) + SVM	91.21%	82.22%	93.21%	0.89
PA (Hybrid Vision Transformer + LBP, GIST and HoG)	94.63%	84.42%	95.23%	0.93

The tests' findings demonstrated that our recommended method outperformed the most sophisticated ensemble CNN architectures for classifying culinary cultures. With accuracy of 94.63%, specificity of 95.23%, sensitivity of 84.42%, and kappa value of 0.93, the recommended hybrid approach surpassed all others.

III. Discussion / Future Work

Our future goal is to investigate and identify Indian cuisine items utilizing vision transformers as the base model. We want to evaluate several CNN models and test hybrid CNN and Vision Transformer models, which might provide more accurate results. As far as we can see, there have been no published large-scale attempts to recognize Indian food items using Vision transformers. The paucity of datasets with annotations for Indian cuisine, the absence of clear distinctions between the meals, and the substantial intra-class variance are some of the key reasons why there hasn't been much work done on identifying Indian food items. [14]. Such scenarios make it extremely challenging to recognizing items specially mixed food items.

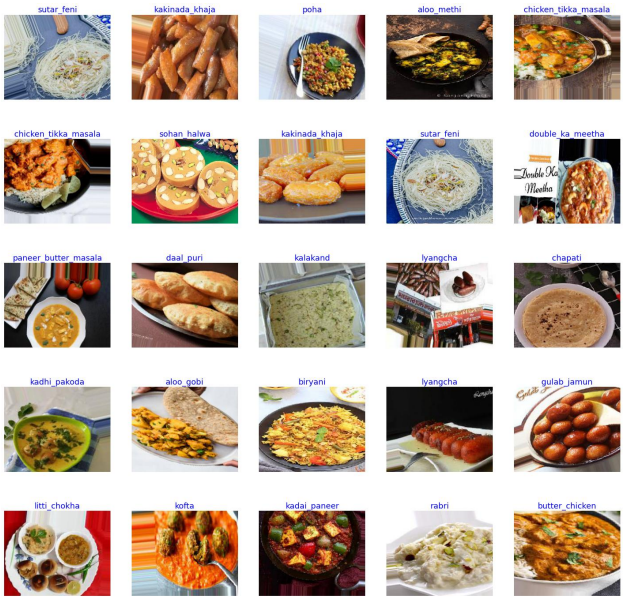

Fig. 1. Indian Food Images

Convolutional neural networks, the state-of-the-art in computer vision at the moment and widely employed in a variety of picture recognition applications, have been proposed as a competitive replacement., the Vision Transformer first appeared in 2022. In terms of accuracy and processing demands, vision transformers perform better than the present state-of-the-art (CNN) by practically a factor of four. The norm in Natural Language Processing is transformer models. Vision Transformers and Multilayer Perceptrons have seen a recent increase in attention in computer vision research (MLPs) [15].

Embedding the patch, Extraction of features using stacked transformer encoders, and classification head make up the architecture of vision transformers. Following a series of transformations, an image of shape is embedded into a feature vector in the first stage.

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \cdots; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, \qquad \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}, \mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D}$$

Using a stack of L transformer encoders, the network learns additional hidden features from the patches that are embedded in the second step. This matches the paper's equations below:

$$\mathbf{z}'_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \qquad \ell = 1 \ldots L$$
$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \qquad \ell = 1 \ldots L$$

The multi-head attention step present in each of the stacked transformers is represented by the equations below.

$$[\mathbf{q}, \mathbf{k}, \mathbf{v}] = \mathbf{z}\mathbf{U}_{qkv} \qquad \mathbf{U}_{qkv} \in \mathbb{R}^{D \times 3D_h},$$
$$A = \text{softmax}\left(\mathbf{q}\mathbf{k}^\top / \sqrt{D_h}\right) \qquad A \in \mathbb{R}^{N \times N},$$
$$\text{SA}(\mathbf{z}) = A\mathbf{v}.$$
$$\text{MSA}(\mathbf{z}) = [\text{SA}_1(z); \text{SA}_2(z); \cdots; \text{SA}_k(z)]\,\mathbf{U}_{msa} \qquad \mathbf{U}_{msa} \in \mathbb{R}^{k \cdot D_h \times D}$$

According to the multi-head attention intuition, having several heads enables the system to learn from numerous aspects of the abstract representation.

There are two weight matrices because pre-training is carried out using a 2-layer MLP. A single linear layer is utilised for fine-tuning, hence there is only one tensor. In each scenario, the network's final output is a vector that contains the probabilities related to each of the ***n_class*** classes. [16]

For our discussion, there are two main implications from this research. First, it illustrates the benefits of utilizing deep learning to recognize food images. It also shows how the vision transformer method might even outperform deep learning because it can discover global features early on, requiring less processing effort. Widely used method is CNN, and in particular VGG16, VGG19, ResNet, Inception V2 and Inception V3. Second, because the vision transformer relies on a self-attention mechanism that takes into account both coarse- and fine-grained global interactions [17], with a big library of food item photos, it might only continue to perform better.

IV. Conclusion

In this study, we investigated a number of vision based techniques for detecting food images, including SVM, CNN, Vision Transformer, etc. Although modern methods are very effective, there are still certain constraints and difficulties. One of the major challenges is adding sizeable food image datasets as it improves the overall performance because of which there is a great requirement for whole datasets benchmarking and performance evaluation of these algorithms. Since it is challenging to distinguish between combined food items, we also need to work on proper identification.

From the papers we reviewed, convolutional neural networks (CNN) have been extensively been used in food detection as it has been giving better results compared to other models. Additionally, we found that Vision transformers work better when the dataset is huge and a hybrid model might provide more accurate results. In [18] they have attempted to investigate Vision Transformers for image recognition. When pre-trained on enormous amounts of data and applied to a number of small or medium-sized image recognition benchmarks (CIFAR-100, VTAB, ImageNet, etc.), Vision Transformer (ViT) produces excellent results while requiring substantially less CPU resources during training.

We also provided a suggestion for future work that will use hybrid models based on CNN and Vision Transformers to better categorise photographs of Indian food. In order to further train the models and increase accuracy, we also intend to collect more data.

References

[1] World Health Organization. (n.d.). *Obesity*. World Health Organization. Retrieved August 29, 2022, from https://www.who.int/health-topics/obesity#tab=tab_1

[2] Kawano, Y., & Yanai, K. (2014). FoodCam: A real-time food recognition system on a smartphone. *Multimedia Tools and Applications*, *74*(14), 5263–5287. https://doi.org/10.1007/s11042-014-2000-8

[3] Anthimopoulos, M. M., Gianola, L., Scarnato, L., Diem, P., & Mougiakakou, S. G. (2014). A food recognition system for diabetic patients based on an optimized bag-of-features model. *IEEE Journal of Biomedical and Health Informatics*, *18*(4), 1261–1271. https://doi.org/10.1109/jbhi.2014.2308928

[4] Gkelios, S., Boutalis, Y., & Chatzichristofis, S. A. (2021). Investigating the vision transformer model for image retrieval tasks. *2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. https://doi.org/10.1109/dcoss52077.2021.00065

[5] *Understanding of a convolutional neural network*. IEEE Xplore. (n.d.). Retrieved August 29, 2022, from https://ieeexplore.ieee.org/abstract/document/8308186

[6] Bazi, Y., Bashmal, L., Rahhal, M. M. A., Dayil, R. A., & Ajlan, N. A. (2021, February 1). *Vision Transformers for Remote Sensing Image Classification*. MDPI. Retrieved August 29, 2022, from https://www.mdpi.com/2072-4292/13/3/516

[7] *Food recognition using statistics of pairwise local features*. IEEE Xplore. (n.d.). Retrieved August 29, 2022, from https://ieeexplore.ieee.org/abstract/document/5539907

[8] Yu, Q., Wang, J., & Mao, D. (n.d.). *Deep Learning Based Food Recognition*. Retrieved August 29, 2022, from http://cs229.stanford.edu/proj2016/report/YuMaoWang-Deep%20Learning%20Based%20Food%20Recognition-report.pdf

[9] Fakhrou, A., Kunhoth, J., & Al Maadeed, S. (2021). Smartphone-based food recognition system using multiple deep CNN Models.

Multimedia Tools and Applications, *80*(21-23), 33011–33032. https://doi.org/10.1007/s11042-021-11329-6

[10] Zahisham, Z., Lee, C. P., & Lim, K. M. (2020). Food recognition with resnet-50. *2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*. https://doi.org/10.1109/iicaiet49801.2020.9257825

[11] R, R. J., G, M., & G, S. (2020, March 12). *Indian food image classification with transfer learning*. IEEE Xplore. Retrieved August 29, 2022, from https://ieeexplore.ieee.org/abstract/document/9031051

[12] Ramesh, A., Sivakumar, A., & Sherly Angel, S. (2020). Real-time food-object detection and localization for Indian cuisines using Deep Neural Networks. *2020 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT)*. https://doi.org/10.1109/icmlant50963.2020.9355987

[13] Nijhawan, R., Batra, A., Loyola-Gonz´alez, O., Kumar, M., & Jain, D. K. (2022). Food classification of Indian cuisines using handcrafted features and vision transformer network. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4014907

[14] Pandey, D., Parmar, P., Toshniwal, G., Goel, M., Agrawal, V., Dhiman, S., Gupta, L., & Bagler, G. (2022, May 10). *Object detection in Indian food platters using transfer learning with Yolov4*. arXiv.org. Retrieved August 29, 2022, from https://arxiv.org/abs/2205.04841v1

[15] Boesch, G. (2022, August 23). *Vision transformers (VIT) in Image Recognition - 2022 guide*. viso.ai. Retrieved August 29, 2022, from https://viso.ai/deep-learning/vision-transformer-vit/

[16] Rad, A.-C. (2021, January 12). *Understanding the vision transformer and counting its parameters*. Medium. Retrieved August 29, 2022, from https://medium.com/analytics-vidhya/understanding-the-vision-transformer-and-counting-its-parameters-988a4ea2b8f3

[17] Yang, J., Li, C., Zhang, P., Dai, X., Xiao, B., Yuan, L., & Gao, J. (2021, July 1). *Focal self-attention for local-global interactions in Vision Transformers. arXiv.org. Retrieved August 29, 2022, from https://arxiv.org/abs/2107.00641*

[18] *Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021, June 3). An image is worth 16x16 words: Transformers for image recognition at scale.*