PGCSAI

Capstone – Project Synopsis

1.0 Problem definition

Image101 Food Classification: Introduction & Problem Definition

One essential component of daily existence is food. There is no life without food. It is a key element of our existence. Thanks to technology, we can capture and analyze images, facilitating their identification, differentiation, and wide use for various purposes. Food image is seen in the widespread use of food photography on social media, specialized photo-sharing websites, and smartphone applications. This cultural phenomenon emphasizes how important food is in today's linked world—not merely as a source of nourishment, but also as a means of expression and social interaction. Because social media and mobile devices are used so frequently in today's digital environment, food photography is becoming more and more popular. Millions of images of food are shared daily on social media platforms like Facebook, Instagram, and Telegram by users worldwide. This rise emphasizes the need for automated systems for classifying food images, which are critical for organizing massive photo collections and enhancing the access of digital content.

It can be challenging to recognize and categorize food products from photographs due to variations in lighting, perspectives, and shooting techniques. Since traditional approaches sometimes suffer from low inter-class variance (similarities between various classes) and large intra - class variance (variations within the same class) making precise categorization is a tough problem.

Our study is centered on utilizing state-of-the-art methods for computer vision, including convolutional neural networks (CNNs), ensembling and transfer learning approaches like Inception V3. These methods are crucial for extracting relevant information from food photos, enabling the system to differentiate between various food groups with a high degree of accuracy.

The main goal is to create a reliable food picture classification model that bridges the gap between object recognition and image classification, hence improving over current approaches while also accurately identifying food items. Our goal is to attain better classification accuracy by using pre-trained models and fine-tuning them on food-specific datasets such as Food-101.

Such a system can be used for many different and significant purposes. In addition to managing social media content, our approach can help with dietary management apps by helping users track their consumption of nutrients through visual food recognition. Furthermore, with precise food image classification for focused marketing tactics, it can support food advertising efforts, which are useful in targeting the right audience for profitable sale.

.

2.0 Literature survey

Food picture classification has gained substantial traction in recent years due to its uses in commerce, nutrition, and health. Food has a significant impact on our health and wellbeing on a daily basis. Precise categorization of food images can improve food safety, ease dietary supervision, and simplify food business processes. This technology is a great tool for both consumers and professionals because it may help with meal planning, allergen detection, and calorie counting. Additionally, it enhances customer service and efficiency by supporting automated systems in supermarkets and restaurants**.**

Deep learning has completely changed how image classification tasks, especially food picture classification, are approached in the field of computer vision. At the vanguard of these developments are Convolutional Neural Networks (CNNs) and Visual Transformers, which continuously enhancing state-of-the-art (SOTA) performance. But obtaining small gains in performance frequently necessitates significant increases in model complexity, which also means a growth in the number of parameters and FLOPS (Floating Point Operations Per Second). This trend is seen by looking at how SOTA models have changed over time on the ImageNet classification challenge. For instance, AlexNet needed 0.7 billion FLOPS and had almost 60 million parameters when it attained 63.3% accuracy in 2012. Higher accuracy has been shown by later models, including InceptionV3 and ResNeXt-101, albeit at the expense of higher computing needs.

A notable advancement was made by EfficientNet, especially with the 2019 release of the EfficientNet-b0 model, which achieved the best possible balance between complexity and accuracy. With just 5.3 million parameters and 0.4 billion FLOPS, model achieved 77.1% accuracy, proving that high performance does not automatically equate to large complexity. For real-world deep learning applications in food image classification, where processing power may be scarce and this balance is essential.

2.2 Algorithms and Models:

The problem of reliably identifying a variety of food products has drawn a lot of attention in the rapidly changing field of food image categorization. Bossard et al.'s introduction of the Food-101 dataset constitutes a significant advancement in this field. The collection started off with 101,000 photos divided into 101 different cuisine groups. Although the dataset contained some noisy images to promote robustness, 750 images per class were set aside for training and 250 images for testing. This dataset has been essential in evaluating several categorization techniques, demonstrating the intricacy of food image identification as well as its room for advancement.

Accurately identifying food photographs presents a number of challenges. Traditional techniques like Random Forests (RFs) were used in the early work in this subject to mine discriminative regions inside images. Bossard et al. showed that their RF-based method outperformed several alternative approaches at the time, with an accuracy of 50.76%; nonetheless, Convolutional Neural Networks (CNNs) still reached a greater accuracy of 56.40%. These findings show how difficult it can be to tell apart comparable food items and emphasize the need for more advanced methods.

The focus of later research switched to deep learning, specifically CNNs, which provided a notable improvement in performance. For instance, Lu used a five-layer CNN to process ten food classifications, which is a subset of ImageNet data. By using CNN, accuracy was significantly increased, reaching 74% as opposed to a bag-of-features model's 56%. The accuracy was improved even further using data augmentation approaches, reaching a remarkable 90%. However, because this study only included a small number of classes, its conclusions do not apply to the larger and more complicated Food-101 dataset.

With the introduction of the DeepFood model by Liu et al., food image categorization has seen a more recent advancement. This model, which was influenced by architectures like LeNet-5, AlexNet, and GoogleNet, used Inception modules to improve network depth and finished first on the Food-101 dataset with a top-1 accuracy of 77.4%. By isolating the dish from background noise with bounding boxes, the performance was substantially enhanced. This development shows how CNNs are becoming more and more capable of managing challenging picture categorization jobs.

By contrasting these approaches, it can be seen that although CNNs have continuously outperformed more conventional approaches like as RFs, improving accuracy has mostly depended on incorporating sophisticated algorithms and data augmentation. Out of all the techniques, ensembling—specifically, bagging—stands out as a potentially effective strategy. Multiple poor models are combined by ensembling to produce a more reliable classifier. By combining predictions from several models that were each trained on a distinct portion of the data, this method can lower variance and enhance model performance. In contrast to individual models, ensembling makes better use of the advantages of each component model, hence resolving bias and variance concerns.

To create a more powerful prediction model, ensembling entails merging several models, sometimes referred to as weak learners. This strategy usually entails averaging or voting among the weak learners' outputs to aggregate them. Although ensembling can greatly enhance performance, it also raises the computational and overall model complexity. Our work investigates how to achieve a competitive performance-to-complexity ratio in food image classification tasks by applying a well-defined ensembling technique that makes use of an effective base model.

Ensembling involves combining several weak learners to produce a powerful learner. These weak learners may be heterogeneous, combining several machine learning paradigms, or homogenous, belonging to the same model family. The principal objective of ensembling is to minimize the volatility and bias that impact individual models. The intricate patterns in the training data are frequently missed by low-complexity models due to their high bias. On the other hand, more complicated models frequently overfit the training set, which results in excessive variance and inadequate generalization to new data. By striking a balance between the trade-offs between bias and variance, ensembling resolves these problems.

In ensembling, three main methods are frequently employed: bagging, boosting, and stacking. By training the weak learners on various subsets of the training data, bagging lowers variance among them. With this method, the dataset is divided into discrete subsets, a weak learner is trained on each subset, and the outputs are aggregated via weighted voting or averaging. Boosting, on the other hand, trains weak learners in a sequential fashion, with each learner fixing the mistakes of its predecessor, with the goal of minimizing both bias and variation. In stacking, many weak learners are trained in simultaneously, and their outputs are combined using a meta-model.

To improve their performance, deep learning models have been more and more integrated with these ensembling approaches in recent years. To increase the robustness and generalization capacities of numerous deep neural networks, bagging has been used, for example, during training on various data subsets. It has been demonstrated that by iteratively improving their predictions, boosting increases the accuracy of object identification algorithms. The outputs of different neural network architectures have been combined via stacking, utilizing their complimentary qualities to provide better results.

Recent literature has provided ample evidence of the deep learning field's use of ensembling techniques. For instance, research has shown that by decreasing overfitting and increasing generalization, bagging can dramatically boost CNN performance on image classification tasks. It has been demonstrated that by iteratively improving their predictions, boosting increases the accuracy of object identification algorithms. Combining the outputs of many deep learning architectures through stacking has enhanced performance on a range of computer vision applications.

Our goal is to push the boundaries of food image classification by combining effective deep learning models with ensembling approaches. This method assures the models' cost-effectiveness and sustainability while also improving the models' accuracy and resilience. We show the effectiveness of our suggested approach in accomplishing these goals through thorough testing and strict validation.

3.0 Data

The Food-101 dataset, originally utilized in the paper "Food-101 – Mining Discriminative Components with Random Forests" by Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool, consists of 101,000 images across 101 food categories, with each category containing 1,000 images. However, for this project, we are utilizing a sampled subset of the dataset to implement and evaluate ensembling techniques. Specifically, we will use 500 images from each of 16 randomly selected food categories and 125 images from the 'apple pie' category, resulting in a small, yet diverse dataset. This approach allows for efficient experimentation while maintaining a representative sample of the overall dataset.

The sampled dataset comprises 8,125 images, carefully selected from the Food-101 dataset for classification. These images are organized into folders, with each folder representing a different food class. Each folder contains a set of sample images relevant to its class. The dataset will be further divided into training and testing sets to evaluate model performance. This structured organization facilitates effective data management and training. For reference, sample images from various classes are provided below to illustrate the dataset's diversity.

Food sample images from the dataset

The categories include a variety of food items such as apple pie, pizza, waffles, tacos, hot dogs, French fries, donuts, and chocolate cakes. This sampling strategy ensures a mix of visually similar and distinct categories, providing a robust testbed for evaluating the performance of ensembling methods.

This dataset poses unique challenges compared to other food image datasets, such as the 10-class food image dataset from ImageNet. The ImageNet dataset features relatively distinct and fewer food categories, including apple, banana, broccoli, burger, egg, French fries, hot dog, pizza, rice, and strawberry. In contrast, the Food-101 dataset includes food items that are similar in both content and presentation, such as pho and ramen. Furthermore, the training images in the Food-101 dataset vary significantly in lighting, coloring, and size, and contain mislabeled images. These intentional anomalies encourage models to develop robustness against labeling inaccuracies.

Images from the Food-101 dataset were normalized and resized appropriately, either to 128x128 or 256x256 pixels in the initial model implementations, or to specifications required by the ensembling models. Image data augmentation techniques, including rotation, shifting, and horizontal flipping, were employed to prevent overfitting. During the ensembling process, images were preprocessed using custom preprocessing functions, reflecting the original methods detailed in the model papers. Additionally, the robustness of ensembling techniques allows for improved performance, leveraging the combined strengths of multiple models to handle the inherent variability and noise within the dataset.

In summary, this approach uses CNNs to extract features from food images, combines these features into a single vector, and then uses a fully connected layer for classification. It leverages CNNs and feature fusion, making it a sophisticated yet distinct method from traditional ensemble techniques.

4.0 Tentative list of Algorithms

Several techniques have been investigated in previous studies on food picture classification to solve the difficulties of reliably classifying a variety of food items from photos. At first, Bossard et al. mined discriminative regions in images using Random Forests, and they were able to achieve 50.76% accuracy. This study showed how useful ensemble approaches are for managing noisy data and enhancing the robustness of models. Later research on Convolutional Neural Networks (CNNs) showed how well these networks performed in extracting high-level characteristics from photos. Lu's study on CNNs, for example, used data augmentation approaches to attain 90% accuracy, demonstrating the CNN's versatility in processing a variety of picture attributes.

But even with these improvements, single-model techniques might still have drawbacks like bias or high variance. For example, even if CNNs are good at extracting features, overfitting or improper treatment of intra- and inter-class variances may make generalization difficult for them. In a similar vein, approaches that only use Random Forests or Gradient Boosting may perform poorly at collecting intricate details or have a high variance rate.

Our method combines CNN-based feature extraction with the ensembling model to overcome these drawbacks. By using the CNN to extract hierarchical and detailed information from food photos, the model is better able to accurately distinguish between different types of food. Through the extraction of strong characteristics from the images, we establish a solid basis for further classification.

Instead of following the traditional ensemble approach of combining and voting on the output of weak learners. We extract and combine the features that the CNNs extract from the input. After that the features are concatenated in the concatenation layer and then a final fully connected layer acts on the combined feature vector.

Our implementation entails a number of crucial actions. Initially, using data augmentation approaches to increase robustness, the CNN is trained to extract high-level features from the food photos. This strategy is intended to overcome the shortcomings of earlier approaches by utilizing cutting-edge ensembling techniques to produce better categorization results for food images.

5.0 References

[1] Antonio Bruno, Davide Moroni, Massimo Martinelli, Efficient Adaptive Ensembling for Image Classification, Institute of Information Science and Technologies (ISTI) Institute of Information Science and Technologies (ISTI) National Research Council of Italy (CNR) Via Moruzzi 1, Pisa, Italy.

[2] Malina Jiang, Food Image Classification with Convolutional Neural Networks, Department of Computer Science Stanford University.

[3] Lukas Bossard1, Matthieu Guillaumin1, and Luc Van Gool1, Food-101 – Mining Discriminative Components with Random Forests, Computer Vision Lab, ETH Z¨urich, Switzerland, ESAT, PSI-VISICS, K.U. Leuven, Belgium.

[4] Alex M. Goh and Xiaoyu L. Yann, (2021), “FOOD-IMAGE CLASSIFICATION USING NEURALNETWORK MODEL” Int. J. of Electronics Engineering and Applications, Vol. 9, No. 3, pp. 12-22,DOI 10.30696/IJEEA.IX.III.2021.12-22.

[5] M. Tan, Q. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, Vol. 97 of Proceedings of Machine Learning Research, PMLR, 2019, pp. 6105–6114.