**Capstone Project – Interim Report**

**Food 101 Object Classification**

**1**. **Introduction**

**Food Image Classification**

In recent years, food image classification has emerged as a pivotal area of research with significant implications for commerce, nutrition, and health. The accurate classification of food images can enhance food safety, simplify dietary supervision, and streamline business processes within the food industry. This technology supports a range of applications from meal planning and allergen detection to calorie counting, benefiting both consumers and professionals. In retail and hospitality settings, effective food image classification systems can boost customer service and operational efficiency by automating various processes.

The advent of deep learning has transformed image classification, including food image categorization, by introducing advanced models such as Convolutional Neural Networks (CNNs) and Visual Transformers. These models have set new benchmarks in classification performance but often at the expense of increased computational complexity. For instance, early models like AlexNet achieved notable accuracy with 0.7 billion FLOPS and 60 million parameters, but newer models such as InceptionV3 and ResNeXt-101, while delivering higher accuracy, also demand significantly more computational resources.

A key advancement in this field is the EfficientNet model, which achieved a remarkable balance between model complexity and accuracy with just 5.3 million parameters and 0.4 billion FLOPS, reaching 77.1% accuracy. This demonstrates that high performance does not necessarily require excessive complexity, making it particularly suitable for real-world applications where computational resources may be limited.

Despite these advances, food image classification remains challenging due to the variety and similarity of food items, variations in image quality, and potential mislabelling. The Food-101 dataset, introduced by Bossard et al., represents a significant resource for addressing these challenges, containing 101,000 images across 101 food categories. This dataset has been instrumental in evaluating various classification techniques and highlights the need for sophisticated methods to manage the complexity of food image identification.

Traditional methods such as Random Forests (RFs) initially showed promise but were eventually surpassed by deep learning techniques, particularly CNNs. CNNs demonstrated superior performance in feature extraction and classification, achieving significant accuracy improvements through techniques like data augmentation. The introduction of models like DeepFood, which leverages Inception modules and bounding boxes, further enhanced classification accuracy by addressing issues related to background noise and intra-class variability.

Recent advancements in ensembling techniques, including bagging, boosting, and stacking, have provided new avenues for improving classification performance. Ensembling methods combine multiple models to reduce variance and bias, enhancing robustness and generalization. These techniques have been effectively integrated with deep learning models, demonstrating improved performance in various computer vision tasks.

This paper aims to push the boundaries of food image classification by proposing a novel ensembling approach that combines CNN-based feature extraction with advanced ensembling techniques. By utilizing bagging to balance model complexity and computational efficiency, we seek to enhance classification accuracy and robustness while accommodating resource constraints. Our approach will be evaluated using a subset of the Food-101 dataset, providing a practical testbed for assessing the effectiveness of the proposed method.

Through this work, we aim to contribute to the development of more efficient and accurate food image classification systems, offering insights into how ensembling techniques can be leveraged to overcome the inherent challenges of food image categorization.

**2. Exploratory Data Analysis**

**Summary**

This project's dataset comes from the Food-101 dataset, a popular benchmark for classifying food images. Luc Van Gool, Matthieu Guillaumin, and Lukas Bossard presented the Food-101 dataset, which consists of 16,256 photos from 17 different food classes. These classes encompass a diverse range of foods, with each class representing a distinct culinary category, such as apple pie.

We carefully chosen and separated the 8,125 retrieved photos into training and testing sets for this research. The images were organized into 17 folders under both the training and testing directories, with each folder corresponding to a different food class. The dataset was divided in a 70:30 ratio to form the training and testing sets.

About 5,686 photos were placed in the training set and 2,437 images in the testing set during this process. Because both sets are guaranteed to be representative of the entire dataset by this stratified technique, robust model training and evaluation are made possible.

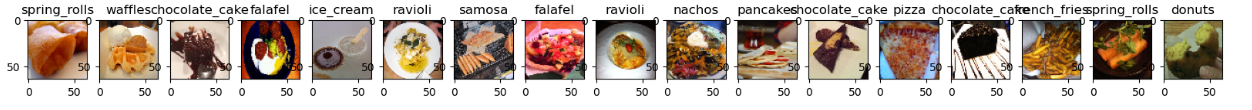
**Sample Dataset**

This is how original pictures look in the food image 101 dataset folders.



**Visualization**

After loading and exploring dataset, we visualized a sample of training data:



Techniques for augmenting data were used to improve the model's capacity for generalization. These methods, which include rotation, shifting, and horizontal flipping, aid in simulating different lighting situations that could be present when taking pictures of food.

**3. Base Model and Architecture**

The architecture of the model is built to manage image data efficiently. Several layers and components specifically designed for feature extraction and classification make up the architecture. Images scaled to 64x64 pixels with three color channels (RGB) are accepted by the input layer. Three identical CNNs make up the model, which is assembled in an ensembling manner. The following is each CNN model's specific structure:

**1. Input Layer:** Supports 64x64x3 input pictures.

**2. Convolutional Layers:** There are three convolutional layers in every CNN model:

- The first convolutional layer employs the ReLU activation function and features 32 filters with a 3x3 kernel size. A max-pooling layer with a 2x2 pool size comes after it.

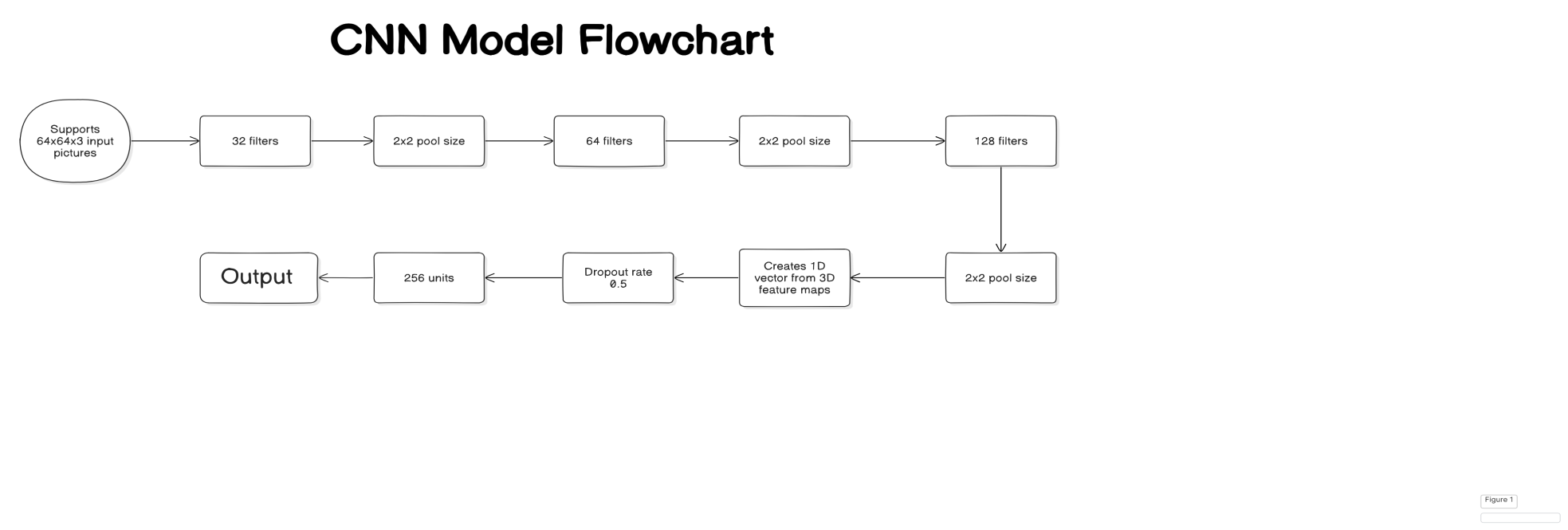
- ReLU activation is used in the second convolutional layer, which has 64 filters and a 3x3 kernel size.

There is one more max-pooling layer after this.   
- ReLU activation and 128 filters with a 3x3 kernel size make up the third convolutional layer. A final max-pooling layer follows it.

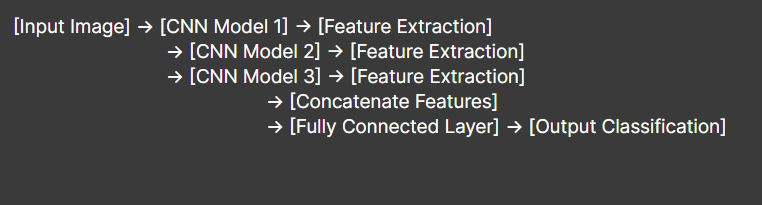
**3. Flatten Layer:** Creates a 1D vector from the 3D feature maps.

**4. Dense Layers:** Every CNN model consists of a dropout layer with a dropout rate of 0.5 that is followed by a fully connected layer with 256 units and ReLU activation.

**Structure of CNN Model**

  
  
The outputs from the three CNN models are concatenated and sent to a final fully connected layer for classification after feature extraction. The following succinctly describes the final model architecture:

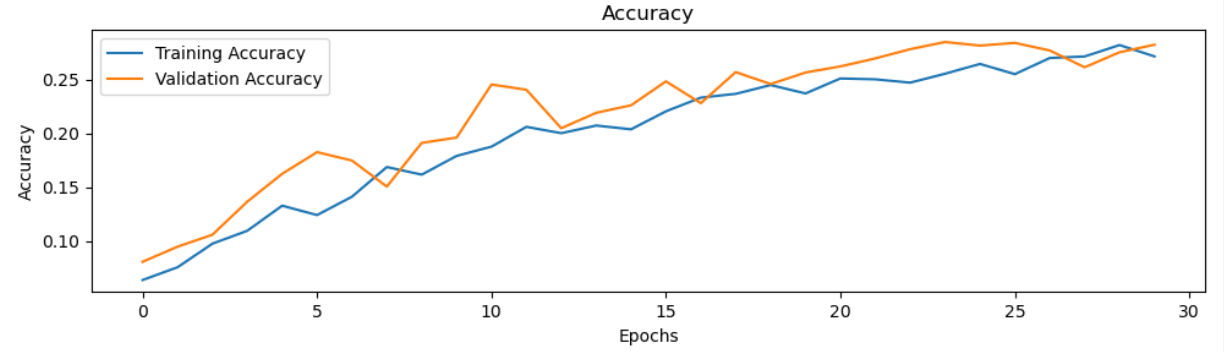
**Model Diagram**



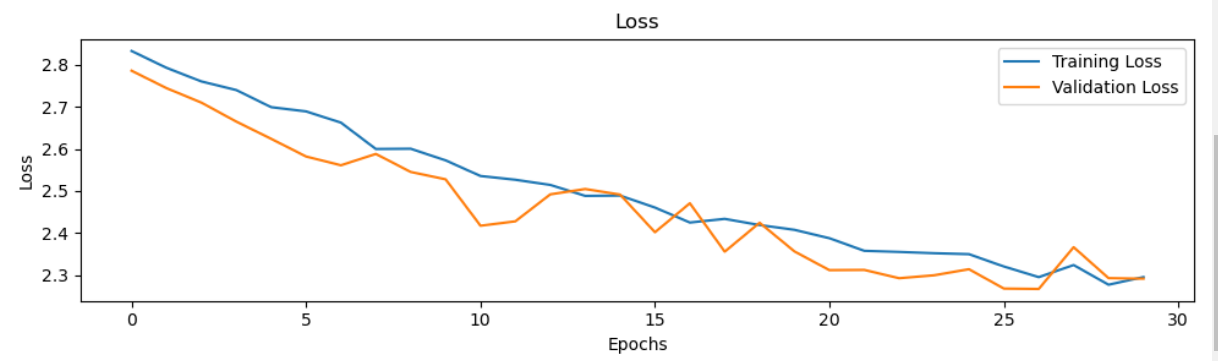
**1. Feature Concatenation:** The three CNN models' extracted features are combined.   
**2. Fully Connected Layer:** Concatenated features are combined with a thick layer that has 512 units and ReLU activation.

**3. Output Layer:** Classification is done using a dense layer that uses softmax activation and a number of units equal to the number of food groups (17).   
  
To improve the model's resilience and avoid overfitting, data augmentation techniques are employed during training. Rotation, shifting, shearing, zooming, and horizontal flipping are all included in the data augmentation.

**4. Early Results**

****

**Training Accuracy Vs Validation Accuracy**



**Training Loss Vs Validation Loss**

Thirty epochs were used to train the model. Thus far, the following are the outcomes:   
  
- Test Accuracy: 27.70% - Test Loss: 2.27

These data show that early development is still ongoing for the model. Although there has been considerable progress, further fine tuning is needed to get greater precision. The findings imply that in order to improve the model's performance, more tuning and optimization are required.

**5. Tentative Algorithms**

Several techniques are employed to enhance the model's performance:

**1. Convolutional Neural Networks (CNNs):** CNNs are used for their effectiveness in feature extraction from image data. They have shown to be highly effective in handling image classification tasks.

**2. Ensembling Techniques:** By combining the outputs from multiple CNN models, we aim to improve **classification accuracy. The ensembling approach leverages the strengths of each individual model.**

**3. Data Augmentation:** Techniques such as rotation, shifting, and horizontal flipping are applied to the training data to prevent overfitting and improve model robustness.

**4. Early Stopping:** The early stopping callback monitors the validation loss and restores the best weights when necessary, helping to manage training time and avoid overfitting.

Future work will focus on optimizing hyperparameters, refining the ensembling strategy, and exploring advanced techniques such as boosting and stacking to further enhance the model's performance.

**Conclusion and Future Work**

The current performance of the model indicates that while progress has been made, further optimization and refinement are needed. The preliminary results demonstrate that the model's accuracy is still below the desired level. Future work will involve optimizing hyperparameters, refining the ensembling strategy,transfer learning from pre-trained models and exploring advanced techniques to improve accuracy. The project will continue to build on the current results and implement additional improvements. By leveraging advanced techniques and exploring innovative approaches, we aim to develop a robust classification system capable of accurately handling various food image scenarios.

**6. 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 Zurich, Switzerland, ESAT, PSI-VISICS, K.U. Leuven, Belgium.

[4] Liu, L., et al. (2016). "DeepFood: Deep Learning for Food Recognition." European Conference on Computer Vision (ECCV).

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). "Deep Residual Learning for Image Recognition." IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[6] Hu, J., Shen, L., & Sun, G. (2018). "Squeeze-and-Excitation Networks." IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[7] Zhang, Y., & Xu, L. (2021). "A Comprehensive Review of Ensemble Deep Learning: Opportunities and Challenges." International Journal of Computer Vision (IJCV).