Image Classification on iFood Dataset

Mustafa Soydan
Physics Department
University of Bicocca
Milan, Italy
m.soydan@campus.unimib.it

Kerem Erciyes
Physics Department
University of Milan Bicocca
Milan, Italy
k.erciyes@campus.unimib.it

***Abstract*—This project aims to image classification using feature extraction with Bag of Words (BoW), Scale-Invariant Feature Transform (SIFT) then with traditional machine learning classifier and to apply Convolutional Neural Networks (CNN).**

***Keywords*—Convolutional neural networks, classification, bag of words, SIFT**

I. Introduction

Image classification is a fundamental task in computer vision, crucial for numerous applications such as object detection, image retrieval, and automated inspection. Traditional methods for image classification typically involve manual feature extraction techniques like Bag of Words (BoW) and Scale-Invariant Feature Transform (SIFT). These methods convert images into a set of numerical features that machine learning classifiers can process. While effective, these traditional methods rely heavily on domain expertise and the quality of the extracted features. In recent years, Convolutional Neural Networks (CNNs) have emerged as a powerful alternative, capable of automatically learning hierarchical features directly from raw image data. CNNs have demonstrated superior performance in various image classification challenges due to their ability to capture spatial hierarchies and complex patterns. This project aims to evaluate and compare the performance of traditional feature extraction methods with CNN's in image classification tasks. By systematically analyzing the strengths and limitations of each approach, the study provides insights into their effectiveness and practical applications, contributing to the advancement of robust image classification systems.

II. Methodology

This study utilizes two distinct methodologies for image classification: traditional feature extraction combined with machine learning classifiers, and deep learning using Convolutional Neural Networks (CNNs). The following subsections describe the procedures employed in each method.

A. Data Preprocessing and Augmentation

- Image Resizing: All images are resized to a uniform dimension to ensure consistency in feature extraction and model input.
- Normalization: Pixel values of the images are normalized to a range of [0, 1] to improve the convergence of the learning algorithms.
- Rotation and Flipping: Images are randomly rotated and flipped to introduce variability and make the models more robust to orientation changes.
- Scaling and Translation: Random scaling and translation are applied to simulate different perspectives and improve the generalization capability of the models.
- Brightness and Contrast Adjustments: Variations in brightness and contrast are introduced to mimic different lighting conditions.

B. Traditional Feature Extraction and Classification

1) Feature Extraction

SIFT is used to detect and describe local features in images. Key points are identified based on their scale and orientation, and descriptors are generated to represent the local gradients around each key point.

The SIFT descriptors are clustered using k-means clustering to form a visual vocabulary. Each image is then represented as a histogram of visual word occurrences, effectively summarizing the local features into a global representation.

The histograms generated from the BoW model are used as input features for the SVM classifier. The SVM algorithm finds the optimal hyperplane that separates the classes in the feature space, maximizing the margin between the closest data points of different classes.

2) Deep Learning with Convolutional Neural Networks (CNN)

A Convolutional Neural Network is designed to automatically learn hierarchical features from the raw image data. The network consists of several convolutional layers, pooling layers, and fully connected layers. Convolutional layers apply filters to the input image to detect various features, while pooling layers reduce the dimensionality, and fully connected layers perform the classification based on the extracted features.

The CNN is trained on a labeled dataset of images. During training, the network adjusts its weights through backpropagation to minimize the classification error. Data augmentation techniques are applied on the fly during training to improve generalization. The performance of the CNN is evaluated on a separate test set to ensure its generalizability to unseen data.

C. Comparative Analysis

Both methods are evaluated based on their classification accuracy, computational complexity, and robustness to variations in the image data. The traditional approach using SIFT and BoW provides insights into the effectiveness of manually engineered features, while the CNN demonstrates the capability of deep learning models to automatically learn and classify complex patterns in images. The comparative analysis offers a comprehensive understanding of the advantages and limitations of each approach, contributing to the development of more robust image classification systems

III. Data Preprocessing and Augmentation.

First, since the dataset is large according to our computational resources the data is limited. Random 200 images are selected for training and validation 80% of it used for training while 20% of it used for validation. Images are used as given for testing. For data augmentation, images randomly resized to [224, 224] pixels and cropped, flipped randomly, brightness, contrast, saturation and hue values are changed, randomly cropped, normalized and transformed to tensor format. Data loaders are created for training, testing and validating steps. The data is shuffled only for training and batch size is determined as 32.

IV. Feature extraction Using Bag of Words and Scale-Invariant Feature Transform

BoW model is a popular technique for feature extraction used in natural language processing and computer vision. In the context of image classification, BoW represents an image by counting the occurrences of visual words within it, where these visual words are derived from local image features such as SIFT descriptors. This method transforms images into a histogram of visual word frequencies, which can then be used as input for machine learning classifiers. The BoW approach is valued for its simplicity and effectiveness in summarizing local features into a global representation, making it suitable for various classification tasks. However, it often requires careful preprocessing and does not inherently capture spatial relationships between features (Csurka et al., 2004).

SIFT is a robust feature extraction method used in computer vision for detecting and describing local features in images. SIFT identifies key points in an image that are invariant to scale, rotation, and partially invariant to affine transformations and illumination changes. Each key point is assigned a descriptor, a vector that encodes the local image gradient information around the key point. This descriptor is highly distinctive, allowing for reliable matching between different images of the same object or scene. SIFT has been widely adopted due to its effectiveness in various applications, including image stitching, object recognition, and 3D modeling (Lowe, 2004).

A. Machine Learning Model for Classification

Support Vector Machines (SVM) are a class of supervised learning algorithms used for classification and regression tasks. SVMs operate by finding the hyperplane that best separates the data points of different classes in a high-dimensional space. This hyperplane is determined by the support vectors, which are the data points closest to the decision boundary. The objective of SVM is to maximize the margin between the support vectors and the hyperplane, ensuring robust classification. SVMs are particularly effective in high-dimensional spaces and are versatile with different kernel functions that allow them to handle non-linear classification problems. They have been widely applied in various domains such as bioinformatics, text categorization, and image recognition (Cortes & Vapnik, 1995).

B. Application

SIFT is applied to each image to detect keypoints and compute descriptors representing local image gradients around these keypoints. Then, the collected SIFT descriptors from the training set are clustered into a number of clusters using k-means clustering, forming a visual vocabulary where each cluster center represents a visual word. Then, for each image is represented as a histogram of visual words by assigning its SIFT descriptors to the nearest visual words in the in the Bag of Words (BoW) model. These histograms, representing the BoW features, are used as feature vectors for training the Support Vector Machine (SVM) classifier. Grid Search is used to tune hyperparameters and best parameters are selected. The performance of the SVM classifier is evaluated on a validation set, with metrics such as accuracy, precision, recall, and F1-score. Finally, the trained SVM classifier is tested on a separate test set, and the test accuracy, precision, recall, and F1-score are reported to evaluate the model's performance.

V. CNN

CNNs are a class of deep learning models designed specifically for processing structured grid data, such as images. CNNs leverage a hierarchical structure, where each layer extracts increasingly complex features from the input image, allowing the network to learn spatial hierarchies and patterns effectively. This ability to automatically learn and generalize features from raw data sets CNNs apart from traditional methods, which require manual feature extraction and domain expertise. The architecture of CNNs typically includes convolutional layers, pooling layers, and fully connected layers, which work together to reduce the dimensionality of the data while preserving its essential characteristics. This structure enables CNNs to achieve state-of-the-art performance in various image classification tasks (Krizhevsky, Sutskever, & Hinton, 2012).

A. CNN Models

First, CNN is created with 2 convolutional layers and 2 linear layers. 808.431 parameters are used. Every convolutional layer followed by ReLU activation function and Max pooling layers. Dropout layers applied before the linear layers.

Second, new model is created with 6 convolutional layers and 1 average pooling layer. After each convolutional layers 2D batch normalization and ReLU activation function and 2D max pooling are applied. For classifier dropout is performed just before the last convolutional layer and at the end everything is flattened. The model has 961.659 parameters. More convolutional layers are used in this model comparing to the other in order to perform better feature extraction.

B. Training

The first model trained for 200 epochs, the second one trained for 80 epochs. For each training Adam Optimizer is used with 0.001 initial learning rate. Best results according to the best validation loss is saved and used for testing.

VI. Experimental Results

A. Feature Extraction and SVM

1) Experiment

Hyperparameters are found by grid search.

a) Hyperparameters

- Kernel: RBF, linear,
- C: 1, 10 ,100
- Gamma: Scale, 0.01, 0.1, 1

Best Hyperparameters are:

- Kernel: RBF
- C: 1
- Gamma: Scale

Scale of results are given below:

b) Results:

- Validation Accuracy: 0.010 - 0.060
- Validation Precision: 0.020 - 0.059
- Validation Recall: 0.010 - 0.060
- Validation F1-score: 0.010 – 0.050
- Test Accuracy: 0.015 - 0.06
- Test Precision: 0.012 - 0.06
- Test Recall: 0.013 - 0.06
- Test F1-score: 0.012 - 0.05

With the selected hyperparameters, best results are:

- Validation Accuracy: 0.060
- Validation Precision: 0.053
- Validation Recall: 0.060
- Validation F1-score: 0.048
- Test Accuracy: 0.060
- Test Precision: 0.061
- Test Recall: 0.060
- Test F1-score: 0.056

B. CNN

1) First Experiment

- Number of Epochs: 200
- Batch Size: 32

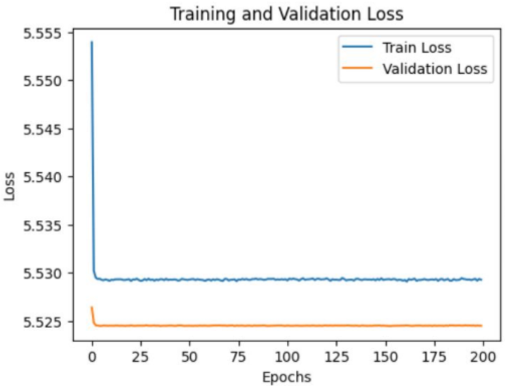

Figure 1: Loss Epoch Graph of 1st Experiment

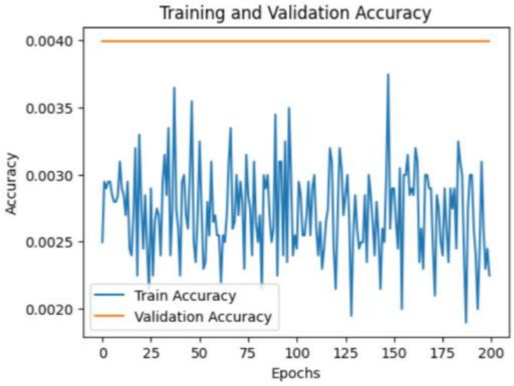

Figure 2: Accuracy Epoch Graph of 1st Experiment

2) Second Experiment

- Number of Epochs: 80
- Batch Size: 32

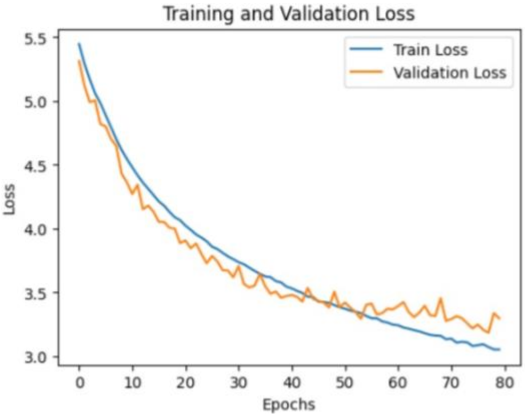

Figure 3: Epoch Loss Graph of 2nd Experiment

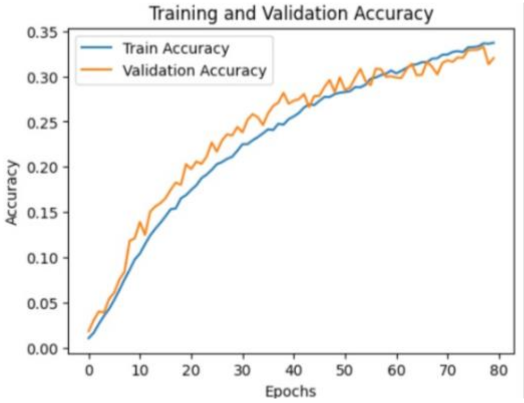

Figure 4: Accuracy Epoch Graph of 2nd Experiment

a) Results:

- Test Accuracy: 0.39
- Test Precision: 0.43
- Test Recall: 0.39
- Test F1: 0.39
- MCE: 0.60
- Throughput: 894.75 images/sec

VII. DISCUSSION

At the CNN part, using Max Pooling layer instead of Linear Layers allow to both stick to the requirement of keep the number of parameters lower than 1M and adding more convolutional layers. The accuracy increased significantly. The structural difference shown in Figure 5. Using more convolutional layers leads to perform better feature extraction.

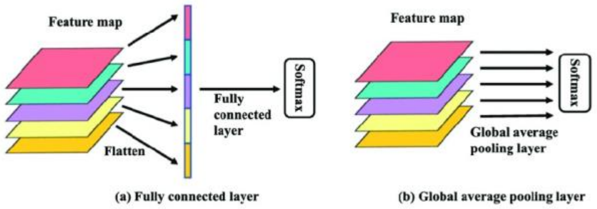
Feature map
Flatten
Fully connected layer
Softmax
(a) Fully connected layer
Feature map
Softmax
Global average pooling layer
(b) Global average pooling layer

Figure 5: The Structure of Different Classifier Layers

From the computational complexity view, while the SIFT and BoW approach involves significant computation during feature extraction, the CNN requires substantial computational resources during training. However, the CNN's training process benefits from GPU acceleration, whereas SIFT and BoW primarily rely on CPU processing.

The accuracy achieved with the best CNN model is 39%. CNN is not demonstrated the superior ability to learn and classify complex patterns because, the model is limited to have 1M parameters at most which is not complex enough for this classification problem.

From the computational complexity view, while the SIFT and BoW approach involves significant computation during feature extraction, the CNN requires substantial computational resources during training. However, the CNN's training process benefits from GPU acceleration, whereas SIFT and BoW primarily rely on CPU processing.

The CNN demonstrates greater robustness to a wide range of image variations compared to the traditional approach, which is primarily robust to scale and rotation but less so to other transformations.

The CNN outperforms the traditional SIFT and BoW approach with SVM in terms of classification accuracy, achieving higher accuracy on the test dataset.

The CNN automatically learns relevant features from the raw data with different convolutional layers, whereas the traditional method depends on manually engineered features, which not able to capture all the relevant information in this case.

Food images vary within the same class due to differences in presentation, lighting, angle, and background. SIFT descriptors are failed to capture the complex features ,as like challenging texture and patterns, required to distinguish between such variations. Clustering local features into a histogram, may not capture the rich, detailed information needed for accurate classification.

The results clearly show that the limitations of the traditional SIFT-BoW-SVM approach in handling the complexity and variability in the food image dataset. In contrast, CNNs show superior performance due to their ability to learn and generalize features directly from the data.

ACKNOWLEDGMENT

The authors would like to thank Prof. Simone Bianco for his guidance and support throughout this project. Additionally, we extend our gratitude to his assistant, Mirko Agarla, for his assistance and contributions.

REFERENCES

[1] Csurka, G., Dance, C. R., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. *In Workshop on Statistical Learning in Computer Vision, ECCV* (pp. 1-22).

[2] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91-110.

[3] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297.

[4] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1097-1105.