Image Classification of Food: A Comparative Study

Stanford CS131

Andrew Sung
Department of Computer Science
Stanford University
drewsung@stanford.edu

Abstract

This study compares various computer vision models in classifying food items from images, exploring a potential path to revolutionize personal health and nutrition. Through data pre-processing and leveraging various machine learning models for computer vision: Support Vector Machine (SVM), Convolutional Neural Network (CNN), and MobileNetV2, along with ways to improve the model performance, we conducted a comparative analysis to optimize current approaches to this problem. The MobileNetV2 deep learning framework demonstrated the highest accuracy, reinforcing the strength of pre-trained models and the transfer learning approach when it comes to the problem of food classification through computer vision.

1 Introduction

Navigating healthy dietary choices is a prevalent challenge for college students, particularly those striving to balance their academic responsibilities with their commitments to physical fitness and health. We want to address this challenge by providing an application that leverages computer vision techniques to identify and analyze the nutritional content of food items directly from images captured by a smartphone camera. This project is inspired by the HUMBIO 132 (Sports Nutrition) course, which equips us with a deeper understanding of nutritional strategies to optimize athletic performance and overall well-being. Through this project, we aim to translate these academic insights into practical tools that help students make informed dietary choices conducive to their fitness and academic goals. This project not only serves as a practical tool for individuals seeking to maintain a healthy diet but also has potential applications in healthcare, especially in managing diet-related diseases.

2 Dataset and Features

The dataset features 646 images of various plates of food, with a focus on Italian and Asian dishes as well as soups. The images are also labeled with a list of foods in them, but the number of foods per image is not fixed. For instance, an image can contain a single label like "Miss Piggy Pizza" or an image can contain multiple labels such as "Jasmine Rice," "Panang Curry," and "Spicy String Beans." There were also calorie counts recorded in the dataset, but we decided to focus on image classification for this study, which is more closely related to computer vision. A challenge with the dataset was that the labels were imbalanced, since certain food items such as Jasmine rice are more common in dishes, whereas other food items such as "Miss Piggy Pizza" were far less common. Additionally, images were of varying dimensions (primarily through differences in being vertical or horizontal) and also seemed to have captured a lot of noise with plates, utensils, and hands in the images in addition to the food items themselves. This required a rigorous pre-processing of the data, in which we standardized the dimensions of the images to 500 x 500 pixels and decided to use Contrast Limited Adaptive Histogram Equalization (CLAHE) which is a technique used to improve contrast in an image. We originally were going to use Histogram Equalization (HE) but that overly enhanced the contrast of the images and seemed to amplify the noise too much. Since CLAHE performs histogram equalization in localized parts of the image and clips the histograms at a predefined value before calculating the cumulative distribution function (CDF), it is a lot better at reducing the harsh noise amplification. We also decided to apply CLAHE to the lightness components of the images to enhance the images

Figure 1: Example of a pre-processed image

without significantly altering their color balance. We also made every image 500x500 to standardize the dimensions, so that should help as well. Above is an example of an image, and its corresponding pre-processed image.

3 Methodology: Models used

3.1 Support Vector Machine

We first used a Support Vector Machine (SVM) model as a baseline model. SVM is a supervised machine learning model that can be used for classification or regression challenges. By finding the hyperplane that best divides a dataset into classes, the SVM model performs classification by finding the decision boundary that maximizes the margin between different classes of data. In the context of food classification from the images in the dataset, the SVM classifier extracts the features of the images to learn how to distinguish between different categories of images.

In our model, we used Histogram of Oriented Gradients (HOG) to extract features. HOG is a feature descriptor used in computer vision and image processing for the purpose of object detection. The idea behind HOG is to capture the structure of objects in an image through the distribution (histograms) of directions of gradients (oriented gradients). Gradients (edge directions) are good indicators of the shape of an object. By dividing the image into small connected regions, called cells, and compiling a histogram of gradient directions or edge orientations for the pixels within each cell, the model creates a feature vector that can be used to train a machine learning model for image classification tasks.

3.2 Convolutional Neural Network

We then utilized a Convolutional Neural Network (CNN) for this food image classification task, leveraging its notorious capability in handling image data, hoping to see performance gains over the SVM model. CNNs are a class of deep neural networks, highly effective in recognizing patterns and structures within images due to their hierarchical architecture—which typically comprises of convolutional layers for feature extraction, pooling layers for dimensionality reduction, and fully connected layers for classification. Each convolutional layer applies filters to the input, capturing spatial hierarchies of features.

In our study, we designed a CNN with layers sequentially arranged as follows: four convolutional layers with ReLU activation for feature extraction, each followed by a max-pooling layer for down-sampling, and a fully connected network on top, concluding with a softmax activation function to classify images into predefined categories. The model was compiled with the Adam optimizer and categorical cross-entropy loss function. The mathematical representation of the operation performed by a convolutional layer is given by:

$$f(x,y) = \sum_{i=-a}^{a} \sum_{j=-b}^{b} k(i,j) \cdot g(x-i, y-j)$$

Where:

- $f(x,y)$ is the output feature map,
- $g(x,y)$ is the input image,

- $k(i, j)$ represents the kernel or filter applied over the input image dimensions x and y, with i and j indexing the kernel dimensions.

This convolution operation extracts patterns from the input image, which are then used by the network to classify images effectively. The network's ability to automatically and hierarchically extract features relevant to the classification task makes it well-suited for the complex visual patterns inherent in food images, which is what we were working with in the food dataset in the study.

3.3 MobileNetV2 Deep Learning Framework

We finally leveraged MobileNetV2, which is a compact and computationally efficient deep learning framework, which is pre-trained on 1 million images from the ImageNet database, making it suitable for real-time applications and devices with limited computational resources. The model's pre-training on the extensive ImageNet dataset enables effective feature recognition through transfer learning, crucial for distinguishing various food items in our dataset, which contains diverse and complex visual patterns across 646 images. This is particularly beneficial given the dataset's challenges, such as label imbalance, noise from non-food items in images, and varying image dimensions. Mathematically, the operation performed by a depthwise separable convolution, which is a key component of MobileNetV2, can be represented as:

$$DSC(I, K) = \sum_{c=1}^{C} (I_c * K_c)$$

Where:

- DSC denotes the depthwise separable convolution operation.
- I is the input image or feature map with C channels.
- K represents the set of C depthwise convolutional kernels, one per channel.
- $*$ denotes the convolution operation.
- I_c and K_c are the c^{th} channel of the input and the c^{th} convolutional kernel, respectively.

This formula highlights the efficiency of MobileNetV2, where each channel is processed separately with its convolutional kernel, and then the results are aggregated. This mechanism significantly reduces the computational complexity and the number of parameters, positioning MobileNetV2 as a promising choice for the food classification task in environments with computational limitations, such as mobile devices. Given the presence of noise in the dataset, such as utensils, plates, and hands, alongside the actual food items, MobileNetV2's lightweight architecture is beneficial. Its depthwise separable convolutions can effectively learn from the relevant features by reducing the influence of noise, which is crucial for achieving high accuracy in food classification. Also, the adaptation to the dataset's varied image dimensions and quality, including the pre-processing steps like the application of CLAHE, aligns well with MobileNetV2's design. Its architecture is adaptable to changes in input size and quality, ensuring that the preprocessing steps enhance rather than detract from the model's ability to classify images accurately.

3.4 Upsampling

As mentioned earlier, a challenge that we faced in this study was the fact that the dataset was pretty heavily imbalanced in terms of its labels. After assessing the performance of the baseline SVM model, we also evaluated the performances of the CNN model as well as the MobileNetV2 deep learning framework. We specifically used upsampling on CNN and MobileNetV2 because this technique helps to mitigate the effects of label imbalance by artificially increasing the presence of underrepresented classes in the training set. Upsampling tends to be particularly effective in deep learning models like CNNs and MobileNetV2, as these models rely on having sufficient examples of each class to learn the complex patterns necessary for accurate classification.

4 Results

In our research, the primary metrics for evaluating model performance were derived from detailed classification reports, emphasizing overall accuracy, precision, recall, F1 score, and support for each

classification label. This comprehensive approach enabled us to benchmark the models' general effectiveness while offering insights into their performance nuances.

We started with an SVM model as a baseline model for this food image dataset, and we achieved an overall accuracy of 0.38, which we would consider to be pretty decent, considering that the task is an image classification task with a variable number of labels. We then tried a CNN model to try and achieve a better accuracy score, but we found that the CNN model initially performed worse than the SVM model at 0.31 accuracy. However, after upsampling the images with lower frequency groups, we found a substantial performance jump with the CNN model to 0.52 accuracy, suggesting that class imbalance was something that had affected the CNN model quite drastically. The MobileNetV2 model achieved the highest accuracy at 0.55, and then upsampling improved this model's performance further, pushing the model's accuracy to a score of 0.58.

The SVM model, while serving as a robust baseline, highlighted the computational cost associated with traditional machine learning methods when applied to high-dimensional data such as images. Despite achieving a respectable accuracy of 0.38, the computational inefficiency of SVMs poses significant challenges for scalability and real-time applications. This limitation accentuates the necessity for more computationally efficient models that can leverage the inherent structure of image data.

The initial performance of the CNN model, yielding a lower accuracy of 0.31 compared to the SVM, initially suggested a potential misalignment between the model architecture and the task complexity. However, the remarkable improvement in accuracy to 0.52, following the strategic upsampling of underrepresented classes, illuminated the profound influence of class imbalance on model performance. This revelation not only attests to the sensitivity of CNN architectures to the distribution of training data but also highlights the efficacy of upsampling as a method to mitigate class imbalance, thereby facilitating a more equitable learning environment for all classes.

5 Conclusion / Future Work

This study presents a comprehensive analysis of the factors influencing model performance in image classification tasks, with a particular focus on the food image dataset. Our findings advocate for the strategic combination of data balancing techniques, efficient model architectures, and the leveraging of pre-trained models to enhance accuracy and computational efficiency. Future research should explore the scalability of these approaches to larger datasets and more complex classification tasks, potentially incorporating additional methods such as data augmentation and transfer learning to further bolster model robustness and generalization.

Tables

	precision	recall	f1-score	support
bread_sticks	0.25	0.25	0.25	4
brocolli_cheddar_soup	0.00	0.00	0.00	3
brown_rice	0.51	0.68	0.58	37
cashew_chicken	0.00	0.00	0.00	7
cheese_pizza	0.22	0.33	0.27	6
chicken_coconut_curry	0.44	0.57	0.50	7
chicken_noodle_soup	0.00	0.00	0.00	1
ciabatta	0.50	0.57	0.53	7
classic_chili	0.00	0.00	0.00	1
combo_supreme	0.50	0.50	0.50	2
corn_bread	0.00	0.00	0.00	3
creamy_mushroom_soup	0.00	0.00	0.00	1
creamy_tomato_basil_soup	0.00	0.00	0.00	1
ginger_chicken	0.00	0.00	0.00	1
jasmine_rice	0.27	0.39	0.32	23
lobster_bisque_soup	0.12	0.33	0.18	3
mapo_tofu	0.00	0.00	0.00	2
meatlovers_pizza	0.00	0.00	0.00	4
pepperoni_pizza	0.00	0.00	0.00	7
pineapple_pizza	0.67	0.33	0.44	6
spicy_string_beans	0.00	0.00	0.00	1
stirfry_beef	0.00	0.00	0.00	1
vegetarian_lentils_soup	1.00	0.50	0.67	2
accuracy			0.38	130
macro avg	0.20	0.19	0.18	130
weighted avg	0.32	0.38	0.34	130

Accuracy: 0.38461538461538464

Table 1: SVM Classification Report

	precision	recall	f1-score	support
bread_sticks	0.00	0.00	0.00	4
brocolli_cheddar_soup	0.00	0.00	0.00	3
brown_rice	0.46	0.57	0.51	37
cashew_chicken	0.00	0.00	0.00	7
cheese_pizza	0.40	0.33	0.36	6
chicken_coconut_curry	0.50	0.29	0.36	7
chicken_noodle_soup	0.00	0.00	0.00	1
ciabatta	0.17	0.14	0.15	7
classic_chili	0.00	0.00	0.00	1
combo_supreme	1.00	0.50	0.67	2
corn_bread	0.00	0.00	0.00	3
cream_of_chicken_soup	0.00	0.00	0.00	0
creamy_mushroom_soup	0.00	0.00	0.00	1
creamy_tomato_basil_soup	0.00	0.00	0.00	1
ginger_chicken	0.00	0.00	0.00	1
jasmine_rice	0.30	0.39	0.34	23
lobster_bisque_soup	0.00	0.00	0.00	3
mapo_tofu	0.00	0.00	0.00	2
meatlovers_pizza	0.00	0.00	0.00	4
pepperoni_pizza	0.11	0.14	0.12	7
pineapple_pizza	0.67	0.33	0.44	6
spicy_string_beans	0.00	0.00	0.00	1
stirfry_beef	0.00	0.00	0.00	1
vegetarian_lentils_soup	0.50	0.50	0.50	2
accuracy			0.31	130
macro avg	0.17	0.13	0.14	130
weighted avg	0.30	0.31	0.29	130

Table 2: CNN Classification Report

	precision	recall	f1-score	support
bread_sticks	0.50	0.50	0.50	4
brocolli_cheddar_soup	0.00	0.00	0.00	3
brown_rice	0.50	0.59	0.54	37
cashew_chicken	0.20	0.14	0.17	7
cheese_pizza	0.75	0.50	0.60	6
chicken_coconut_curry	0.60	0.86	0.71	7
chicken_noodle_soup	1.00	1.00	1.00	1
ciabatta	1.00	0.71	0.83	7
classic_chili	0.33	1.00	0.50	1
combo_supreme	0.50	0.50	0.50	2
corn_bread	0.00	0.00	0.00	3
cream_of_chicken_soup	0.00	0.00	0.00	0
creamy_mushroom_soup	0.00	0.00	0.00	1
creamy_tomato_basil_soup	0.50	1.00	0.67	1
ginger_chicken	0.00	0.00	0.00	1
jasmine_rice	0.38	0.43	0.41	23
lobster_bisque_soup	0.67	0.67	0.67	3
mapo_tofu	0.00	0.00	0.00	2
meatlovers_pizza	0.75	0.75	0.75	4
pepperoni_pizza	0.57	0.57	0.57	7
pineapple_pizza	0.75	0.50	0.60	6
spicy_string_beans	0.00	0.00	0.00	1
stirfry_beef	0.00	0.00	0.00	1
vegetarian_lentils_soup	1.00	1.00	1.00	2
vegetarian_pizza	0.00	0.00	0.00	0
accuracy			0.52	130
macro avg	0.40	0.43	0.40	130
weighted avg	0.50	0.52	0.50	130

Table 3: CNN Classification Report (upsampled)

	precision	recall	f1-score	support
bread_sticks	0.50	0.25	0.33	4
brocolli_cheddar_soup	0.00	0.00	0.00	3
brown_rice	0.58	0.84	0.69	37
cashew_chicken	0.00	0.00	0.00	7
cheese_pizza	0.50	1.00	0.67	6
chicken_alfredo	0.00	0.00	0.00	0
chicken_coconut_curry	0.50	0.71	0.59	7
chicken_noodle_soup	0.00	0.00	0.00	1
ciabatta	0.62	0.71	0.67	7
classic_chili	0.00	0.00	0.00	1
combo_supreme	1.00	0.50	0.67	2
corn_bread	0.00	0.00	0.00	3
cream_of_chicken_soup	0.00	0.00	0.00	0
creamy_mushroom_soup	0.00	0.00	0.00	1
creamy_tomato_basil_soup	0.00	0.00	0.00	1
ginger_chicken	0.00	0.00	0.00	1
jasmine_rice	0.55	0.52	0.53	23
lobster_bisque_soup	0.17	0.33	0.22	3
mapo_tofu	0.00	0.00	0.00	2
market_vegetables_soup	0.00	0.00	0.00	0
meat_lasagna	0.00	0.00	0.00	0
meatlovers_pizza	0.00	0.00	0.00	4
orange_chicken	0.00	0.00	0.00	0
panang_curry_chicken	0.00	0.00	0.00	0
pepperoni_pizza	0.75	0.43	0.55	7
pineapple_pizza	0.80	0.67	0.73	6
side_salad	0.00	0.00	0.00	0
spicy_string_beans	0.00	0.00	0.00	1
spinach_red_curry_with_tofu	0.00	0.00	0.00	0
stir-fry_garlic_soba_noodles	0.00	0.00	0.00	0
stirfry_beef	0.00	0.00	0.00	1
vegetarian_lasagna	0.00	0.00	0.00	0
vegetarian_lentils_soup	1.00	1.00	1.00	2
vegetarian_pizza	0.00	0.00	0.00	0
whole_wheat_bread	0.00	0.00	0.00	0
micro avg	0.55	0.55	0.55	130
macro avg	0.20	0.20	0.19	130
weighted avg	0.47	0.55	0.49	130

Table 4: MobileNetV2 Classification Report

	precision	recall	f1-score	support
bread_sticks	0.67	0.50	0.57	4
brocolli_cheddar_soup	0.00	0.00	0.00	3
brown_rice	0.61	0.73	0.67	37
cashew_chicken	0.00	0.00	0.00	7
cheese_pizza	1.00	0.50	0.67	6
chicken_alfredo	0.00	0.00	0.00	0
chicken_coconut_curry	0.55	0.86	0.67	7
chicken_noodle_soup	1.00	1.00	1.00	1
ciabatta	0.83	0.71	0.77	7
classic_chili	0.00	0.00	0.00	1
combo_supreme	0.33	0.50	0.40	2
corn_bread	0.67	0.67	0.67	3
cream_of_chicken_soup	0.00	0.00	0.00	0
creamy_mushroom_soup	0.00	0.00	0.00	1
creamy_tomato_basil_soup	0.50	1.00	0.67	1
ginger_chicken	0.00	0.00	0.00	1
jasmine_rice	0.45	0.57	0.50	23
lobster_bisque_soup	0.50	0.33	0.40	3
mapo_tofu	0.00	0.00	0.00	2
market_vegetables_soup	0.00	0.00	0.00	0
meat_lasagna	0.00	0.00	0.00	0
meatlovers_pizza	1.00	0.50	0.67	4
orange_chicken	0.00	0.00	0.00	0
panang_curry_chicken	0.00	0.00	0.00	0
pepperoni_pizza	0.60	0.86	0.71	7
pineapple_pizza	0.80	0.67	0.73	6
side_salad	0.00	0.00	0.00	0
spicy_string_beans	0.00	0.00	0.00	1
inach_red_curry_with_tofu	0.00	0.00	0.00	0
r-fry_garlic_soba_noodles	0.00	0.00	0.00	0
stirfry_beef	0.00	0.00	0.00	1
vegetarian_lasagna	0.00	0.00	0.00	0
vegetarian_lentils_soup	1.00	1.00	1.00	2
vegetarian_pizza	0.00	0.00	0.00	0
whole_wheat_bread	0.00	0.00	0.00	0
micro avg	0.58	0.58	0.58	130
macro avg	0.30	0.30	0.29	130
weighted avg	0.55	0.58	0.55	130

Table 5: MobileNetV2 Classification Report (upsampled)

References

Beijbom, Oscar, et al. "Menu-Match Dataset." Menu-Match: Restaurant-Specific Food Logging from Images, Institute of Electrical and Electronics Engineers (IEEE), 23 Feb. 2015, neelj.com/projects/menumatch/data/.

Sharma, Nitika. "What Is MobileNetV2? Features, Architecture, Application and More." Analytics Vidhya, 31 Dec. 2023, www.analyticsvidhya.com/blog/2023/12/what-is-mobilenetv2/.

"Adaptive Histogram Equalization." Wikipedia, 21 Jan. 2020, en.wikipedia.org/wiki/Adaptive$_h istogram_e qualization$.

Juan Carlos Niebles, Adrien Gaidon "Computer Vision: Foundations and Applications" CS 131 at Stanford, Win 2024