IMAGE ANALYSIS OF PLANT BASED MEAT PRODUCTS

CS 499: Final Report

Authors
Badal Chowdhary - 20110034
Hitesh Jain - 20110077
Under the guidance of **Prof. Shanmuganathan Raman**

1 Introduction

In the dynamic culinary landscape and amid shifting dietary preferences, plant-based meat products have sparked a transformative wave in the food industry. These pioneering products provide a sustainable and environmentally conscious alternative to traditional meat, aligning with the discerning tastes of contemporary consumers. Conventional texture analysis methods, relying on expensive spectrometry and microscopy instruments, pose limitations for routine quality assurance. In response to this challenge, we propose the development of a classification model for plant-based meat products. This model aims to distinguish between lab-made and commercial varieties, leveraging specific classes determined by cooking methods and duration. By exploring innovative approaches to texture analysis, our goal is to offer a more accessible and cost-effective solution for assessing the quality of plant-based meat alternatives.

2 Previous Work

2.1 Dataset Creation

The previous research involved creation of a dataset of images with approximately 10,900 images distributed among 14 classes of Plant-based meat product patties, sourced from IITGN FoodLab and Tata commercial product. Two distinct cooking methods namely, air frying and deep frying were used with standardized cooking durations, including instances of overcooking. Ultimately, 14 distinct classes were established as shown below in Figure 4a.

Figure 1: Flow chart showing 14 distinct classes of Plant based meat products

3 Current Work and Results

3.1 Curation of Existing Dataset

The dataset retrieved from previous research poses significant challenges due to its noisy characteristics. The images frequently feature diverse background objects with contrasting colors. A prevalent problem observed is the misalignment of the patties, leading to tilted angles. Our initial attempts to extract the patties and eliminate the background from the images were based on employing various techniques from the cv2 library. This involved converting the original image to grayscale, applying Canny edge detection to clearly define the boundaries, and subsequently utilizing the Hough Circle Transform to identify circular shapes within the image, which correspond to the patties. Ultimately, a mask was applied to the original image in order to isolate and extract the patties.

Figure 2: Comparison of Images featuring Original Patties, Extracted Images using cv2 Techniques, and SAM Results for given prompts, highlighting various backgrounds and misalignments in the patties.

Following our initial attempts, it is evident from Figure 2 that the detection algorithm performs optimally only when the patty is correctly aligned. We have observed that the effectiveness of the current method is limited by the non-circular, elliptical shape of the patties. Consequently, we propose the adoption of a segmentation model to accurately extract images of the patties from the complex background. Implementing this approach is expected to significantly enhance the quality of the dataset and improve the precision of our research results. However, it is essential to acknowledge the labor-intensive nature of annotating images for training such a segmentation model. An estimated 10,000 images would require manual annotation, which is a time-consuming task that demands careful consideration.

Grounded SAM[1] is a powerful computer vision tool that combines Grounding Dino with Segment Anything Model to identify and segment objects based on given text prompts.

Grounding DINO[2] is a state-of-the-art AI model that seamlessly locates objects in images and matches them with corresponding textual labels. It combines the power of DINO[3] (DETR with Improved deNoising anchor bOxes) with GLIP[4] (Grounded Language-Image Pre-Training), resulting in a revolutionary approach to object detection. DETR[5] (DEtection TRansformer) is a transformer-based model that revolutionized object detection by eliminating the need for hand-engineered components

such as anchor boxes and non-maximum suppression. It introduced a novel approach that casts object detection as a direct set prediction problem. Grounding DINO further enhances its performance through grounded pre-training. This technique introduces grounding Language-Image Pairs (GLIP) during the pre-training phase, providing valuable contextual information for object detection models. By incorporating textual descriptions of images, Grounding DINO establishes a better correlation between image features and textual labels, resulting in more precise object localization and recognition.

Grounding DINO, in conjunction with the Segment Anything Model (SAM)[6], accelerates the image annotation process significantly. The Segment Anything Model (SAM) is a state-of-the-art image segmentation model developed by MetaAI's FAIR lab. SAM is based on foundation models that have had a significant impact on natural language processing (NLP) and utilizes advanced techniques to achieve highly accurate and efficient segmentation results. A key feature of the Segment Anything Model (SAM) is its ability to be prompted with various input methods, such as clicks, boxes, or text. This provides users with the flexibility to interactively guide the segmentation process and refine the results based on their specific needs.

We used Grounded SAM i.e. combination of Grounding Dino and SAM in order to detect and extract the patty from the image to mitigate the noise in the dataset. Our Grounded SAM pipeline involves the utilization of Grounding Dino, which takes an image and a corresponding text prompt ("Patty") as input. This component identifies objects in the image aligning with the provided text prompt and generates a bounding box around the object exhibiting the highest confidence level in matching the input text. Subsequently, this bounding box is fed into SAM, which performs segmentation to create a precise mask for the patty. By leveraging this mask, we effectively extract the patty from the image. Figure 4 provides a visual representation of the entire pipeline.

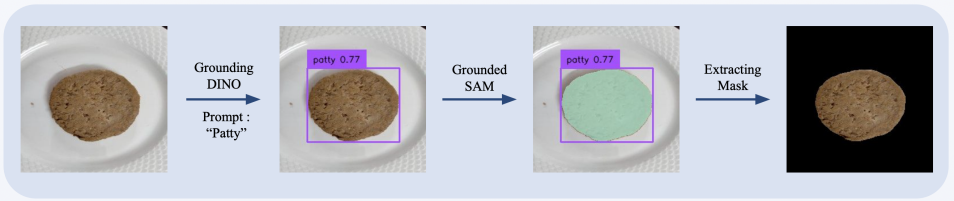


Figure 3: Figure showing dataset curation pipeline using Grounded SAM

This methodology was systematically applied across the entire dataset to curate and refine it. The curated dataset, thus obtained, serves as a crucial resource for our study, allowing for a more efficient analysis of the patty's texture and other pertinent properties while eliminating any interference from background noise. This meticulous curation process enhances the quality of our dataset, contributing to the reliability and accuracy of our investigation into the characteristics of the patty in focus.

3.2 Zero Shot inference on Foundational Model CLIP

The CLIP[7] model, short for "Contrastive Language-Image Pre-training," is a state-of-the-art deep learning model created by OpenAI. It is specifically designed to comprehend and establish connections between images and text, enabling it to effectively accomplish diverse tasks. As a result, CLIP proves to be a versatile tool applicable in numerous domains and scenarios. One notable advantage of the CLIP model is its ability to perform zero-shot image classification effectively. Zero-shot learning refers

to the capability of a model to classify images into classes that were not present during training. CLIP achieves this by leveraging the pre-training on a large dataset containing image-text pairs collected from the internet. This pre-training enables CLIP to learn the correspondence between natural language descriptions and images, allowing it to generalize to unseen classes during inference.

In this study, we employed zero-shot inference with CLIP to assess the accuracy of the 14 distinct classes. The code for this procedure can be found in the GitHub repository [8]. We established three metrics, namely R1, R3, and R5, as measurements of recall. The recall score, denoted as R@K, signifies the proportion of the top K retrieved captions that are relevant to the input image. For R1, the model associates one caption with a given image. Likewise, for R3 and R5, the model associates three and five captions, respectively, with a given image. An image-text pair is deemed correct if there is at least one predicted label that matches an actual label.

Table1 presents the inferences of each class on the R1, R3, and R5 metrics.

Class Name	**R@1**	**R@3**	**R@5**
commercial_air_normal	0.0%	0.5%	16.5%
commercial_air_over	0.0%	2.0%	24.0%
commercial_deep_normal	0.0%	0.5%	56.5%
commercial_deep_over	0.0%	63.0%	76.5%
commercial_unbaked	98.0%	100.0%	100.0%
inhouse_air_normal	0.0%	0.0%	3.0%
inhouse_air_over	0.0%	4.0%	24.0%
inhouse_deep_normal	0.0%	1.0%	11.5%
inhouse_deep_over	0.0%	10.5%	74.5%
inhouse_old_air_normal	0.0%	0.0%	3.5%
inhouse_old_air_over	0.0%	4.5%	15.5%
inhouse_old_deep_normal	0.0%	1.0%	10.5%
inhouse_old_deep_over	0.0%	12.0%	44.0%
inhouse_unbaked	44.5%	98.0%	99.5%

Table 1: Zero shot inference for 14 Classes

In the current dataset, two different cooking methods, namely Deep Frying and Air Frying, were used for every three types of products. However, research presented in [9] suggests that the cooking method employed does not significantly impact critical parameters such as appearance, color, and texture of the plant-based meat products. To streamline the dataset, we have therefore reduced it to only one cooking method for all three types of products, resulting in a reduction of the number of classes from fourteen to eight. The inferences of the reduced eight classes on the R@1, R@3, and R@5 metrics are presented in Table3.

Class Name	R@1	R@3	R@5
commercial_deep_normal	0.0%	3.0%	89.5%
commercial_deep_over	0.0%	71.0%	88.5%
commercial_unbaked	98.5%	100.0%	100.0%
inhouse_deep_normal	0.0%	0.5%	15.0%
inhouse_deep_over	0.0%	18.0%	86.0%
inhouse_old_deep_normal	0.0%	2.0%	18.5%
inhouse_old_deep_over	0.0%	14.5%	53.0%
inhouse_unbaked	40.0%	100.0%	100.0%

Table 2: Zero shot inference for 8 Classes

Reducing the number of classes from 14 to 8 has contributed to minimizing noise and redundancy within the dataset, consequently enhancing the zero-shot inference accuracy of these 8 classes. Specifically, the zero-shot inference accuracy has experienced a notable increase of approximately 3% in the R@2 metric and 9% in the R@5 metric. Furthermore, the overall accuracy of these 8 classes has shown an approximate improvement of 4%.

After curating the dataset, we utilized CLIP[7] to examine the classification accuracy of the eight classes using the newly curated images. The inferences of each class are presented in Table 4.

Class Name	R@1	R@3	R@5
commercial_deep_normal	0.0%	4.0%	92.5%
commercial_deep_over	0.0%	83.5%	92.0%
commercial_unbaked	89.5%	100.0%	100.0%
inhouse_deep_normal	0.0%	0.0%	5.0%
inhouse_deep_over	0.0%	3.5%	51.5%
inhouse_old_deep_normal	1.5%	29.0%	72.5%
inhouse_old_deep_over	0.0%	23.0%	65.0%
inhouse_unbaked	81.5%	98.0%	100.0%

Table 3: Zero shot inference for the curated dataset

When the classes were reduced from 14 to 8, the classification accuracy using CLIP showed significant improvement with an average R@5 accuracy of 39.96%(for 14 classes) to 68.81%(for 8 classes). Moreover, after curating the dataset with Grounded SAM, the average R@5 accuracy was 72.31%.

The classification labels used in CLIP were the original class names in the dataset, which were imprecise and impeded the model's performance. Improved and more meaningful labels could lead to better results.

3.3 Quantifying Texture as a way to classify Image

Texture serves as a critical factor in understanding the visual and tactile properties of plant-based meat products. In our approach, we employ the cv2 RGB2LAB function for converting the RGB color space to the CIE Lab color space, facilitating a more comprehensive analysis of texture nuances. This

conversion is particularly useful when attempting to capture subtle variations that may not be adequately represented in the RGB color space. Additionally, we incorporated the Histogram of Gradients (HoG) algorithm [10], which effectively quantifies textures by counting gradient orientation occurrences in localized image areas. The resulting HoG vector served as a representative feature for each image, enabling subsequent classification through traditional machine learning classification algorithms.

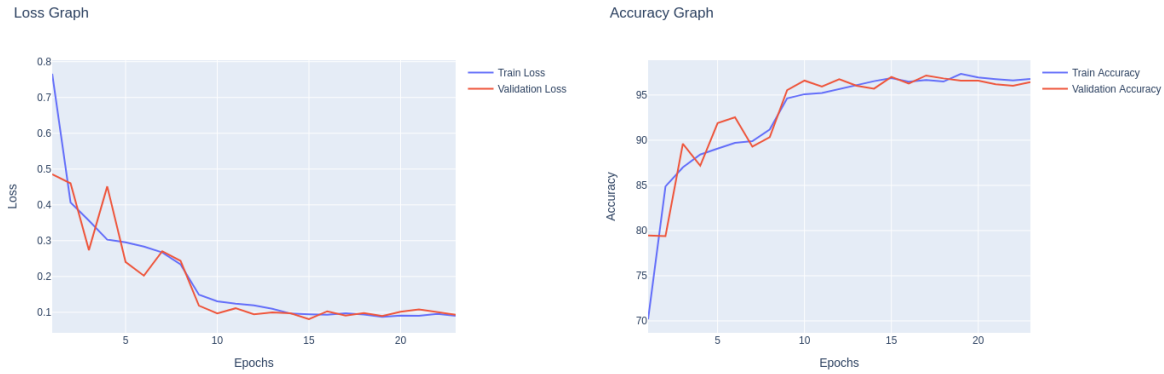
Despite our efforts to address challenges by employing masks from Grounded-SAM to isolate the patty region and create a curated dataset, the application of traditional machine learning classification algorithms yielded lower-than-expected accuracy. The detailed results and parameters used for each model are presented in the attached table 4. This outcome suggests that while the isolation of the patty region helped in minimizing some irregularities, there may still be inherent complexities or nuances in the texture and visual characteristics of the plant-based meat products that traditional machine learning algorithms find challenging to capture. These irregularities may stem from factors such as inconsistent lighting conditions or variations in measurements during the patty-making process in different batches.

Model	**Parameters**	**Accuracy**
Logistic Regression	$\max_i ter = 1000$	36.22
SVM	kernel = 'rbf', C= 12.0	35.73
Random Forest	n_estimators = 200, random_state = 42	37.92
KNN	$\mathrm{n}_n eighbours = 32$	31.68

Table 4: Accuracy on different classification algorithms using curated dataset

3.4 Finetuning ResNet-50 model using curated dataset

We partitioned the curated dataset into an 80:20 ratio for training and validation sets. The ResNet-50 model, initially trained on the Food-101 dataset[11], underwent fine-tuning via a comprehensive training pipeline. This pipeline integrated a robust loss function, specifically cross-entropy loss, to quantify dissimilarity between predicted and actual values during training. Iterative updating of model weights in the last layer was carried out using the Adam optimizer, a widely employed algorithm known for its efficiency in promoting convergence. To optimize training efficiency and counteract overfitting, strategic techniques were incorporated. Learning rate scheduling dynamically adjusted the learning rate during training to optimize model convergence. Additionally, early stopping, a preventive measure against overfitting, halted training when model performance on a validation set plateaued or degraded. Throughout the 25-epoch training process, the model utilized a learning rate (lr) of 1.00E-03 and a weight decay of 0.001. The fine-tuned ResNet-50 model, applied to our curated dataset, showcased exceptional performance metrics. The validation accuracy, representing the model's effectiveness on unseen data, reached an impressive 96.43%. Simultaneously, the training accuracy, reflecting the model's performance on the training data, achieved a high accuracy of 96.77%. The loss graph and accuracy graph for training and validation are shown below in Figure 4



(a) Loss plot for training and validation sets

(b) Accuracy plot for training and validation sets

Figure 4: Loss and accuracy plots of fine-tuned ResNet-50 model

4 Application and Future Aspects

Creating a comprehensive dataset that includes both plant-based and real meat counterparts can provide valuable insights into the nuanced differences between these products. To improve the quality of this dataset, it is important to implement standardized conditions during data collection. This includes ensuring consistency in factors such as lighting, background, and the way patties are prepared, as these factors can significantly affect subsequent analyses.

We can explore the potential of Vision Transformers (ViT) as a sophisticated approach to image classification. ViT represents a state-of-the-art paradigm in deep learning, utilizing transformer architectures to analyze images in a holistic manner. By investigating the capabilities of ViT, we can aim to classify images based on the rich texture information extracted through HoG, thereby improving our understanding and accuracy in classifying diverse textures in both plant-based and real meat products.

References

[1] Grounded-SAM Contributors, "Grounded-Segment-Anything," Apr. 2023. [Online]. Available: https://github.com/IDEA-Research/Grounded-Segment-Anything

[2] Anonymous, "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," 2023, under review. [Online]. Available: https://openreview.net/forum?id=DS5qRs0tQz

[3] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," 2022.

[4] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, K.-W. Chang, and J. Gao, "Grounded language-image pre-training," 2022.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," 2020.

[6] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," *arXiv:2304.02643*, 2023.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[8] "Image analysis of plant based meat products." [Online]. Available: https://github.com/badalchowdhary/Food_Recognition

[9] G. Vu, H. Zhou, and D. J. McClements, "Impact of cooking method on properties of beef and plant-based burgers: Appearance, texture, thermal properties, and shrinkage," *Journal of Agriculture and Food Research*, vol. 9, p. 100355, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666154322000886

[10] H. Demir, "Classification of texture images based on the histogram of oriented gradients using support vector machines," *Electrica*, vol. 18, pp. 90–94, 2018. [Online]. Available: https://electricajournal.org/Content/files/sayilar/28/90-94.pdf

[11] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – mining discriminative components with random forests," in *European Conference on Computer Vision*, 2014.