COSE474-2024F: Final Project Proposal "Integrating Food Images and Ingredients Representation in Visual-Language Models for Food Recognition"

Wan Amir Zarif

1. Introduction

Vision-language models such as CLIP have great potential in connecting images with their descriptions in text. However, the process of food recognition using these models remains tricky, especially when including extra details, such as an ingredient listing. Food image classification finds great use in health and diet tracking through analysis of what we eat from photos (Yang et al., 2024). Many modern systems of food recognition are based on datasets which, by default, include a set number of food categories and hence do not work well when the addition of new foods takes place. For this issue, my project would make use of the pre-trained models for deeper understanding of food images with added dish names and their ingredients in text as inputs.

2. Problem definition & challenges

The project tries to develop a vision-language model which connects images of food items with their names and ingredient lists, and help the model also identify and differentiate between similar-looking foods. The key challenges would be to train the model which identifies the ingredients from pictures, handle diverse food presentations, and to join text with images in the embedded space when both of these inputs might not be perfect. It would also be difficult for the model to be trained on real-world food pictures compared to the neat and organized, readily available datasets (Liu et al., 2024).

3. Related Works

Other researchers have pointed out that one of the major deficiencies in food recognition models lies in their inability to distinguish between similar foods (Zhuang et al., 2024) and they often do not work well when the test data differs from what they were trained on (Liu et al., 2024). Some new methods will employ attention mechanisms, focus on better datasets, and improve the results. For example, multimodal models such as FoodLMM try to handle food and ingredient recognition, recipe creation, and nutritional values estimation (Yin et al., 2023).

4. Datasets

The plan is to use publicly available food datasets such as Food-101, Recipe1M, and potentially new datasets like DailyFood-172 and DailyFood-16 (Liu et al., 2024), which would be more suitable for the comparison of real-life meal images. These datasets will be split into training, validation, and test sets, a standard approach to ensure that the evaluation will not be biased. This will also enable the project to handle problems such as data drift and other issues which may rise with food recognition due to the use of different datasets (Zhuang et al., 2024).

5. State-of-the-art methods and baselines

To see how well my model works, comparison against other food recognition models using ResNet or EfficientNet will be conducted to gauge its benchmark performance. Other models, such as FoodLMM-which does more than image classification, including segmenting the food items or estimating nutrition values could also be considered (Yin et al., 2023). This comparison will consider accuracy and other common performance measures that indicate if the incorporation of ingredient text prompts really matters.

6. Schedule

Week 1-2: Literature review, dataset collection, and preprocessing.
Week 3-4: Model adaptation and initial experiments with prompt integration.
Week 5-6: Fine-tuning and optimization of the vision-language model, possibly integrating techniques like Gaussian and causal-attention for fine-grained food recognition (Zhuang et al., 2024).
Week 7: Evaluation against baseline and SOTA models, incorporating benchmarks like ETH-FOOD101, UEC-FOOD256, and Vireo-FOOD172 (Zhuang et al., 2024).
Week 8: Final adjustments, documentation, and report preparation.

References

Liu, G., Jiao, Y., Chen, J., Zhu, B., and Jiang, Y.-G. From canteen food to daily meals: Generalizing food recognition to more practical scenarios. *IEEE Transactions on Multimedia*, 2024.

Yang, J., Duan, Z., He, J., and Zhu, F. Learning to classify new foods incrementally via compressed exemplars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3695–3704, 2024.

Yin, Y., Qi, H., Zhu, B., Chen, J., Jiang, Y.-G., and Ngo, C.-W. Foodlmm: A versatile food assistant using large multi-modal model. *arXiv preprint arXiv:2312.14991*, 2023.

Zhuang, G., Hu, Y., Yan, T., and Gao, J. Gcam: Gaussian and causal-attention model of food fine-grained recognition. *Signal, Image and Video Processing*, 18(10):7171–7182, 2024.