Recipe Generation from Food Images

Ziqi Liu
Center for Data Science
New York University
ziqiliu@nyu.edu

Chenxin Gu
Center for Data Science
New York University
cg3423@nyu.edu

Will Hu
Center for Data Science
New York University
kh3492@nyu.edu

Zihang Xia
Center for Data Science
New York University
zihang.xia@nyu.edu

Zheng Tong
Center for Data Science
New York University
zt672@nyu.edu

Abstract

Our project addresses the challenge of generating recipes from food images by extending two advanced models, ChefFusion[1] and FIRE[2], using Inverse Cooking[3] as a baseline. We refine ChefFusion's embedding transformations and integrate advanced language models to improve instruction generation. For FIRE, we enhance attention mechanisms to produce coherent, multi-step instructions. Our approach aims to boost ingredient prediction accuracy and instruction fluency, evaluated with F1-score, SacreBLEU, and ROUGE metrics. Preliminary results show promising advancements in recipe quality and coherence.

1 Introduction

Our project centers on enhancing the Inverse Cooking model, which employs a dual-encoder to extract ingredients and generate cooking instructions from food images. We have focused on improving ingredient prediction accuracy and producing contextually relevant, coherent recipes. This work sets the foundation for expanding into more advanced models like ChefFusion, which leverages multimodal embeddings to connect vision and language, and FIRE, which integrates Vision Transformers with language models for richer output. These future steps aim to push recipe generation capabilities even further.

2 Related Work

Our exploration builds upon advancements in both image feature extraction and text generation. Given the multimodal nature of our problem, we provide a comprehensive review of existing methods.

2.1 Image Feature Extraction

Swin Transformer[4]: A hierarchical vision transformer that computes self-attention within non-overlapping windows, reducing computational complexity. Despite its efficiency, the fixed window size may limit its ability to capture global context, especially for large-scale features.

CLIP[5]: A multimodal framework pre-trained on large-scale image-text pairs using a contrastive objective. It enables zero-shot learning and demonstrates strong generalization across various vision tasks.

2.2 Recipe Generation Models

OPT[6]: The Open Pre-trained Transformer (OPT) suite provides a range of pre-trained models up to 175 billion parameters, with an emphasis on efficiency and transparent development practices. While OPT models perform well on general natural language tasks, domain adaptation remains a hurdle. Specifically, generating coherent, structured instructions from visual cues involves significant challenges. The model's pre-training data and architecture must be carefully aligned with the nuances of recipe generation, which we aim to refine further.

GPT-4[7]: GPT-4 represents one of the most advanced language models, with marked improvements in contextual understanding and output quality. For our project, GPT-4's ability to generate detailed, coherent text is highly beneficial, especially in the context of sequential instructions like recipes. Its capacity to handle long-range dependencies and nuanced language makes it a suitable candidate for enhancing our models. However, challenges like model fine-tuning and cost management are crucial considerations.

3 Approach

Our methodology involves a comprehensive analysis and extension of two key models, building on the Inverse Cooking framework as our baseline.

Inverse Cooking (Baseline) Developed by Facebook AI, Inverse Cooking uses a dual-encoder approach. An image encoder predicts ingredients and encodes them, while a cooking instruction decoder generates detailed steps using attention mechanisms. We leverage this model as our baseline, analyzing its strengths and weaknesses to guide our extensions.

ChefFusion ChefFusion is a multitask model that integrates image and text modalities. It uses CLIP's visual encoder to transform image features into text inputs, which are then processed by a large pre-trained language model. The original model employs a linear mapping to align visual embeddings with text features, feeding them into a language model for recipe generation. We refine this embedding transformation and experiment with more advanced language models to improve output quality and instruction coherence.

FIRE FIRE employs a Vision Transformer to extract visual features, followed by the T5 model for sequential recipe generation. This model uses innovative attention mechanisms to manage unordered ingredients and generate coherent instructions. Our improvements focus on optimizing the vision-to-language transfer process and refining attention layers to enhance ingredient accuracy and instruction relevance.

4 Experiments

4.1 Data

We utilize the Recipe1M dataset[8], it provides images, titles, ingredients, and step-by-step instructions. Due to resource limitation, we sampled a subset of 1% samples from the Recipe1M dataset called Recipe1M_1.

4.2 Evaluation method

We employ the following metrics:

Accuracy Accuracy is used to quantify the percentage of ingredients in our recipe compared with the ground truth ingredients.

F1-score To assess the accuracy of ingredient prediction and recipe generation, particularly the model's ability to recall relevant components and avoid false positives.

Jaccard Similarity Jaccard Similarity reflects the model's ability to measure the percentages overlapping between true ingredients and generated recipe ingredients.

4.3 Experimental details

In this section, we will present the experiment setup for evaluating the baseline choice model of ours, Inverse Cooking, on Recipe1M_1 dataset. The model utilizes ResNet50 as the image encoder, with a learning rate of 0.001 and a decay rate of 0.99 applied every epoch. The embedding size was set to 512, and the model's attention mechanisms included 8 layers for recipe generation and 4 layers specifically for ingredient prediction. Training was conducted over 400 epochs, with a batch size of 128 and dropout rates of 0.3 across encoder and decoder layers. Early stopping was implemented with a patience of 50 to prevent overfitting.

4.4 Results

As for now, we have only processed to set up baseline results for further comparison. After training for 74 epochs, the model achieved an accuracy of 0.9878, indicating robust performance in ingredient recognition. As shown in table 1, the F1 score for recipe generation reached 0.3754, and Jaccard similarity is 0.231, which are close to the results training with the original Recipe1M.

Model	**F1-score**	**Jaccard Similarity**
Inverse Cooking (Recipe1M)	0.4826	0.3810
Inverse Cooking (Recipe1M_1)	0.3754	0.2310

Table 1: Evaluation Metrics for Inverse Cooking Baseline

5 Future Work

In future work, we will conduct experiments to replicate and compare the performance of ChefFusion and FIRE models with the Inverse Cooking baseline. Our initial results from the baseline indicated challenges in generating diverse and coherent instructions, with 231 samples containing repetitive steps. By running these additional models, we aim to observe any improvements in recipe coherence and ingredient alignment. This comparison will allow us to assess each model's capabilities and limitations, guiding future exploration in recipe generation from images.

References

[1] Peiyu Li, Xiaobao Huang, Yijun Tian, and Nitesh V Chawla. Cheffusion: Multimodal foundation model integrating recipe and food image generation. *arXiv preprint arXiv:2409.12010*, 2024.

[2] Prateek Chhikara, Dhiraj Chaurasia, Yifan Jiang, Omkar Masur, and Filip Ilievski. Fire: Food image to recipe generation, 2024.

[3] Amaia Salvador, Michal Drozdzal, Xavier Giro-i Nieto, and Adriana Romero. Inverse cooking: Recipe generation from food images. *arXiv preprint arXiv:1812.06164*, 2019.

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[6] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[7] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[8] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3020–3028, 2017.