Recipe Generation from Food Images

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

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

Ziqi Liu
Center for Data Science
New York University
zl5430@nyu.edu

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

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

1 Overview

Motivation. Recipe generation from images or text has advanced with the rise of transformer models, but multimodal approaches that integrate both text and images to produce coherent, accurate recipes still face significant challenges. This gap motivates us to implement, evaluate, and enhance existing models, aiming to push the boundaries of state-of-the-art performance in this domain.

Related work. The task can be divided into two parts, namely, image feature extraction and recipe generation. We briefly review works largely adopted in these fields.

Image Feature Extractor:

1. Swin Transformer[1]: A hierarchical vision transformer architecture that computes self-attention within non-overlapping windows. It significantly reduces computational complexity and is widely used in image feature extraction. However, despite its efficiency, the fixed window size can limit its ability to capture global context in images, especially in scenarios where large-scale features are critical.
2. CLIP[2]: A multimodal learning framework that pre-trains on large-scale image-text pairs using a contrastive objective, enabling it to perform zero-shot learning and demonstrate impressive generalization capabilities across a wide range of vision tasks.

Pre-trained LLM for Recipe Generation:

1. OPT (Open Pre-trained Transformer)[3]: It present an open-access suite of pre-trained transformer-based large language models with up to 175 billion parameters. The models highly emphasize transparency and efficiency in training. The authors also address ethical concerns regarding bias and toxicity inherent in large-scale models.
2. GPT-4[4]: It represents a significant advancement in language modeling, featuring improved contextual understanding and generating more coherent and accurate outputs.

Goal. This project aims to enhance recipe generation from images by improving three existing models: Inverse Cooking (Facebook), FIRE, and ChefFusion. We will leverage large language models like OPT [3] and GPT-4 [4] to better capture image embeddings and generate more accurate cooking instructions using text prompts. Our objective is to meet or exceed current baselines in key metrics (F1-score, BLEU, ROUGE), while addressing limitations in feature extraction and generalization.

If time permits, we plan to further boost performance through continued pre-training and fine-tuning of the large language models.

2 Project plan

2.1 Methods.

Inverse Cooking (Facebook)

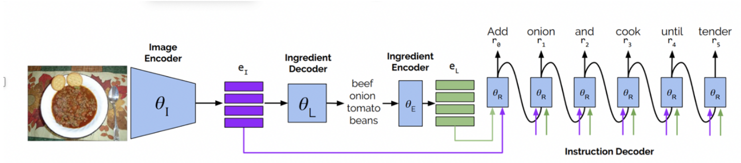
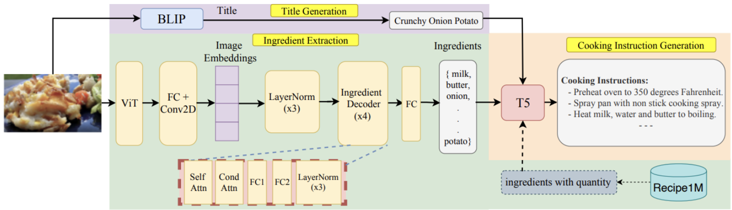


Figure 1: An example caption for Inverse Cooking (Facebook).

There are two sets of encoder-decoder pairs, as shown in 1. We extract image features $\mathbf{e_I}$ with the image encoder, parameterized by θ_I. Ingredients are predicted by θ_L, and encoded into ingredient embeddings $\mathbf{e_L}$ with θ_E. The cooking instruction decoder, parameterized by θ_R, generates a recipe title and a sequence of cooking steps by attending to image embeddings $\mathbf{e_I}$, ingredient embeddings $\mathbf{e_L}$, and previously predicted words $(r_0, r_1, \ldots, r_{t-1})$.[5]

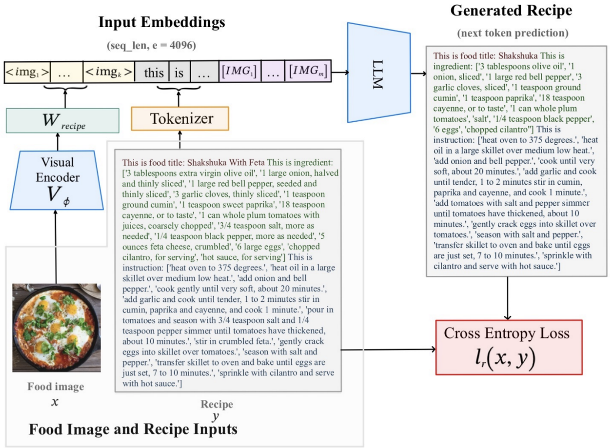
FIRE



Cooking Instructions:
- Preheat oven to 350 degrees Fahrenheit.
- Spray pan with non stick cooking spray.
- Heat milk, water and butter to boiling.
- - -

Figure 2: Architecture of FIRE method.

This system generates recipes from food images using state-of-the-art vision and language models, demonstrated in 2. First, the BLIP model is fine-tuned to generate the recipe title from the food image. Then, a Vision Transformer (ViT) extracts image embeddings passed through a decoder to predict ingredients, utilizing attention mechanisms and pooling to manage unordered ingredients. Finally, the T5 model generates the cooking instructions based on the title and ingredients, producing coherent and detailed steps for the recipe. This multimodal approach integrates advanced image processing and language generation techniques to create a complete recipe from just an image[6].



This is food title: Shakshuka With Feta This is ingredient: ['3 tablespoons extra virgin olive oil', '1 large onion, halved and thinly sliced', '1 large red bell pepper, seeded and thinly sliced', '3 garlic cloves, thinly sliced', '1 teaspoon ground cumin', '1 teaspoon sweet paprika', '18 teaspoon cayenne, or to taste', '1 can whole plum tomatoes with juices, coarsely chopped', '3/4 teaspoon salt, more as needed', '1/4 teaspoon black pepper, more as needed', '5 ounces feta cheese, crumbled', '6 large eggs', 'chopped cilantro, for serving', 'hot sauce, for serving'] This is instruction: ['heat oven to 375 degrees.', 'heat oil in a large skillet over medium low heat.', 'add onion and bell pepper.', 'cook gently until very soft, about 20 minutes.', 'add garlic and cook until tender, 1 to 2 minutes stir in cumin, paprika and cayenne, and cook 1 minute.', 'pour in tomatoes and season with 3/4 teaspoon salt and 1/4 teaspoon pepper simmer until tomatoes have thickened, about 10 minutes.', 'stir in crumbled feta.', 'gently crack eggs into skillet over tomatoes.', 'season with salt and pepper.', 'transfer skillet to oven and bake until eggs are just set, 7 to 10 minutes.', 'sprinkle with cilantro and serve with hot sauce.']

This is food title: Shakshuka This is ingredient: ['3 tablespoons olive oil', '1 onion, sliced', '1 large red bell pepper', '3 garlic cloves, sliced', '1 teaspoon ground cumin', '1 teaspoon paprika', '18 teaspoon cayenne, or to taste', '1 can whole plum tomatoes', 'salt', '1/4 teaspoon black pepper', '6 eggs', 'chopped cilantro"] This is instruction: ['heat oven to 375 degrees.', 'heat oil in a large skillet over medium low heat.', 'add onion and bell pepper.', 'cook until very soft, about 20 minutes.', 'add garlic and cook until tender, 1 to 2 minutes stir in cumin, paprika and cayenne, and cook 1 minute.', 'add tomatoes with salt and pepper simmer until tomatoes have thickened, about 10 minutes.', 'gently crack eggs into skillet over tomatoes.', 'season with salt and pepper.', 'transfer skillet to oven and bake until eggs are just set, 7 to 10 minutes.', 'sprinkle with cilantro and serve with hot sauce.']

Figure 3: Architecture of ChefFusion method.

ChefFusion

ChefFusion is a multitask model that addresses t2t, t2i, i2t, it2t, and t2it, released in [7]. We take advantage of its structure to extract information from image and generate recipe, the model mainly has four steps.

First, The paper employs the CLIP model's visual encoder (ViT-L model) to extract features from the input image x. The CLIP visual encoder encodes the input image as a visual feature vector $v_\phi(x) \in \mathbb{R}^d$, where d is the dimension of the visual feature vector.

Second, Transforming Visual Embeddings into Text Inputs. For this, a trainable linear mapping matrix $W_{recipe} \in \mathbb{R}^{d \times k_e}$ is defined, which transforms the visual feature $v_\phi(x)$ into a sequence of vectors matching the embedding dimension of the LLM.

Third, Generating Recipes with the Language Model. After the transformation, we fed the embedding sequence into a pre-trained language model (LLM). The LLM used in original ChefFusion is **OPT-6.7B**.

Finally, Train the Model. During training, the model receives an input image paired with its ground-truth recipe. By adjusting the parameters of the linear mapping matrix, the model learns to generate accurate recipes from the image.

2.2 Baselines.

We plan to use the Inverse Cooking (Facebook) method[5] as our baseline to improve upon it.

2.3 Data.

We plan to train and evaluate our models on the Recipe1M dataset, composed of 1,029,720 recipes scraped from cooking websites[8]. The dataset contains 720,639 training, 155 036 validation, and 154,045 test recipes, containing a title, a list of ingredients, a list of cooking instructions, and (optionally) an image.

2.4 Evaluation.

Inverse Cooking (Facebook)

The authors evaluate a recipe generation system using an instruction decoder compared with a retrieval-based method. They compute recall and precision based on ground truth ingredients.[5]

FIRE

The paper used set-level and document-level metrics for different stages of recipe generation. For ingredient extraction, F1-score and IoU were used. It also utilized SacreBLEU and Rouge-L to evaluate cooking instruction generation.

ChefFusion

In the original paper, the team leverages SacreBLEU and ROUGE-2 as evaluation metrics. SacreBLEU measures the n-gram overlap between generated and reference recipes, focusing on translation quality and fluency, while ROUGE-2 evaluates bigram overlap to assess how well key phrases and concepts are captured. ChefFusion outperforms baseline models in both metrics, demonstrating higher accuracy and relevance in recipe generation from images.

2.5 Compute.

Our model utilizes large language models with integrated vision support, and the cost for using the ChatGPT-4o API is estimated at $0.000638 per image (150×150 pixels) and 70 text tokens. To manage expenses, we will limit the use of the Recipe1M dataset for retrieval-augmented generation (RAG) and fine-tuning, working with a subset of 10,000 data points, which is expected to cost approximately $6. Additionally, we are considering AWS Bedrock, which offers access to a variety of models such as Claude and Llama. By enrolling in the AWS Educate program, we anticipate receiving credits sufficient to cover the project's needs.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

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

[7] 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.

[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.