Understanding and Interpreting K-Food Semantics

Changwoo Yoo
Department of Computer Science
Korea University
cwyoo01@korea.ac.kr

Jaeseung Lee
Department of Computer Science
Korea University
jseuyi@gmail.com

Jiyoon Lee
Department of Data Science
Korea University
1001baam@korea.ac.kr

1 Problem Identification

1.1 Background

The global popularity of Korean media has created a surge in tourism interest, yet many visitors face a "culinary barrier" with Korean food (Hansik). This hesitation stems from:

- **Lack of Information:** Unfamiliarity with dishes, ingredients, and flavor profiles.
- **Cultural Hesitancy:** A general reluctance to try unknown foods.
- **Language Barrier:** Difficulty navigating menus without a visual guide.

This barrier prevents tourists from fully experiencing a vital part of Korean culture. Our project aims to solve this by providing clear, accessible information from just an image.

1.2 Significance and Importance of the Project

This project is compelling due to its specialized focus on Korean cuisine:

- **Bridging Cultural and Economic Gaps:** The tool acts as a digital ambassador for Hansik, making it more approachable. This empowers tourists to explore with confidence, enhancing their travel experience and boosting spending at local restaurants.
- **Innovation in Specialized AI:** General food classification models lack the nuance for Korean food, where many dishes like stews (jjigae) or side dishes (banchan) appear visually similar. Our project's key innovation is a highly specialized model trained exclusively on Korean food. This focus will provide a level of accuracy and practical detail that generic models cannot, offering unique value to users.

2 Methodology

The primary objective of this research is to develop a model that achieves the highest possible accuracy and description quality while remaining compact enough to operate within a resource-constrained environment like Google Colab. To achieve this, this study evaluates two distinct methodologies. The first is a pragmatic, retrieval-augmented approach designed to empower smaller models, while the second explores the feasibility of creating a compact and flexible end-to-end system. Our final goal is to compare these two models to determine the optimal architecture that best satisfies the project's objectives.

2.1 Hybrid Classification-Retrieval-Generation (CRG)

The CRG methodology was devised as a practical solution for leveraging smaller Language Models (LLMs) that may lack extensive prior knowledge. This RAG-inspired pipeline circumvents the need for a massive, all-knowing LLM by externalizing the knowledge base. A lightweight CNN first classifies the image to retrieve factual data from a database. This data is then passed to a small LLM, whose task is simplified from generating knowledge to fluently articulating the provided facts. This approach is designed to achieve high factual accuracy and reliability in a resource-efficient manner, making it an ideal strategy for our target environment.

2.2 Encoder-based Generative Modeling

This methodology was developed to explore the potential of creating a fully end-to-end model that is both compact and flexible. The goal is to determine if an integrated system can be trained to produce high-quality descriptions without relying on an external database. This approach involves connecting a visual **Encoder** to a generative **Language Model** via a lightweight **Mapping Network**. The core experiment is to identify the most efficient encoder (e.g., CLIP, a Custom VLM, or ViT) that provides rich visual context without excessive computational overhead. We expect that this approach will allow us to create a small but highly adaptable end-to-end model.

3 Expected Results

We will go through a two-phase evaluation. The initial phase will focus on the **Encoder-based Modeling** approach to identify the optimal visual encoder. This will be achieved by a comparative analysis of candidates like CLIP and a Custom VLM, using standard image captioning metrics such as **CIDEr** and **SPICE**, culminating in a single, optimized end-to-end model.

In the subsequent phase, this optimized end-to-end model will be benchmarked against the **Hybrid CRG model**. Given its RAG-based architecture, the CRG model may exhibit superior performance in terms of factual accuracy and reliability, establishing a strong baseline with a near-zero hallucination rate. However, the definitive evaluation of which model better balances performance and efficiency is an empirical question that can only be answered through direct comparison. The final results will clarify whether the end-to-end model's capacity for richer, more nuanced descriptions can outweigh the inherent stability of the CRG approach. This will lead to a data-driven recommendation for the optimal architecture that satisfies the project's constraints and objectives.

4 Research Plan

4.1 Project Timeline

- **10/1–10/7:** Data collection and preprocessing of Korean food images (labeling, cleaning, augmentation).
- **10/8–10/15:** Baseline model implementation (CNN, ViT) and initial performance evaluation.
- **10/16–10/30:** Development of the first approach: CRG.
- **10/30–11/4:** Write and submit Progress report.
- **11/5–11/12:** Development of the second approach: Encoder-based Generative Modeling
- **11/13–11/26:** Comparative evaluation of CRG and optimized encoder-based models to identify the best-performing architecture.
- **11/27–12/10:** Model refinement, error analysis, and integration of user feedback.
- **12/11–12/16:** Write and submit Final report.

4.2 Team Roles

- **Changwoo Yoo:** Model architecture design and deep learning implementation.
- **Jaeseung Lee:** Data collection, preprocessing, and retrieval system integration.
- **Jiyoon Lee:** Evaluation design, preprocessing, and interpreting results.