Understanding K-Food Semantics in Resource-Constrained Environment

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

Abstract

The global rise of K-Content has increased tourism in Korea, yet visitors often face a significant culinary barrier due to language and a lack of contextual information about Hansik (Korean food). Addressing this challenge requires an intelligent system capable of operating in offline, resource-constrained environments, such as a mobile device. We propose a hybrid **Classification-Retrieval-Generation (CRG)** pipeline designed to provide accurate and fluent food descriptions from an image. This pipeline first uses a lightweight visual classifier to identify the dish, then retrieves fact-grounded information from a curated knowledge base, and finally employs a small language model (TinyLlama-1.1B) to synthesize this information into a natural-language explanation. We conduct a comparative analysis of classifier architectures on the 150-class AI-Hub Korean Food Image dataset, identifying EfficientNet B3 as the optimal model with **86.36%** test accuracy. Qualitative results demonstrate that our CRG pipeline successfully generates factually accurate and contextually rich descriptions, effectively mitigating the hallucination common in larger models. We also analyze the pipeline's primary limitation: a dependency on the classifier's accuracy and a lack of zero-shot capability for unseen dishes. Future work will focus on integrating a CLIP-based model to enable open-set recognition.

1 Introduction

1.1 Background and Problem Definition

The global spread of K-Content has led to a rapid increase in foreign visitors to Korea. However, many tourists face a culinary barrier when encountering Hansik (Korean food), primarily due to insufficient information on dish names and ingredients, coupled with language barriers. This issue is difficult to solve with simple translation or basic image recognition alone, necessitating an intelligent approach that incorporates the dish's cultural context.

Our project aims to resolve this Hansik comprehension barrier by developing an AI system that provides accurate and natural descriptions upon receiving a dish's image. A primary requirement is that this system must operate on a tourist's mobile device, enabling real-time inference in an offline environment where internet connectivity may be unavailable. This necessitates a key focus on designing an efficient architecture capable of operating within a resource-constrained environment, without relying on massive, server-dependent models. To achieve this, we adopt a **Classifier–Retrieval–Generation**

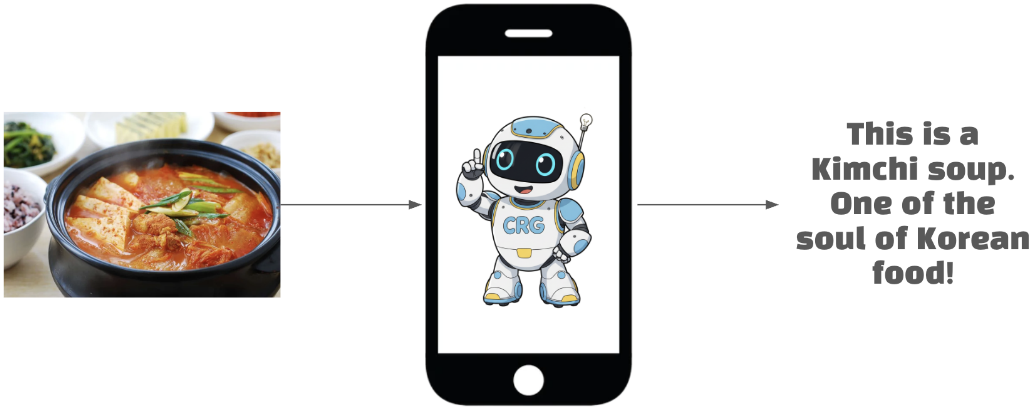


Figure 1: High-level description of our **CRG (Classification-Retrieval-Generation)** pipeline. It is devised for the resource constrained environment such as a mobile phone.

(CRG) pipeline: using the classification result to retrieve information from a Knowledge Base (DB), and leveraging a small Language Model (LLM) to generate factual and fluent explanations.

1.2 Proposed Approach: The Hybrid CRG Pipeline

To address the inherent limitations of standard text-only LLMs, namely their inability to understand image inputs, high computational cost, and the hallucination problem, we propose and implement a **Hybrid Classification-Retrieval-Generation (CRG)** pipeline.

The pipeline is structured as follows:

1. **Classification**: A lightweight classifier quickly and accurately identifies the food item from the input image.
2. **Retrieval**: Using the identified food name as a key, the system retrieves fact-grounded descriptions from a carefully curated Knowledge Base (DB).
3. **Generation**: A small Language Model (`TinyLlama/TinyLlama-1.1B-Chat-v1.0`) is fed the retrieved factual information to generate a fluent, user-friendly, and contextual final response.

1.3 Key Contributions and Report Structure

This report details the preliminary results from the construction and optimization of the CRG pipeline. Our main contributions include:

1. **Optimal Classifier Selection**: A comparative analysis of classifiers (CNN vs. ViT) to select the most efficient model that provides optimal performance under resource constraints.
2. **Qualitative Results for CRG pipeline**: We present a detailed analysis of generated text for sample images. This analysis will evaluate the fluency, factual accuracy, and contextual relevance of the outputs, demonstrating the effectiveness of the hybrid CRG pipeline.
3. **Limitations for CRG and future works**: We will discuss the current constraints of the CRG model, including its dependency on the initial classifier's accuracy (error propagation) and the lack of zero-shot inference ability. Future work will focus on mitigating these issues.

While the current CRG model achieves high accuracy and reliable, fact-based responses, it faces a challenge in **zero-shot generalization** for unseen classes outside its training dataset. Therefore, the subsequent phase will be a CLIP-based approach to address this limitation.

2 Related works

Our research aims to develop an efficient and fact-grounded system for semantic K-Food understanding. This goal sits at the intersection of key studies in Fine-grained Classification and the Retrieval-Augmented Generation (RAG) paradigm.

2.1 Food Classification and Domain-Specific Datasets

Initial developments in food recognition established benchmarks like the Food-101 dataset [2], driving the progress of CNN-based classification models. While these classification approaches achieve high accuracy for seen classes, they often struggle with the fine-grained differentiation required for visually similar dishes in specialized domains like Hansik. Furthermore, they fundamentally lack the ability to generate descriptive text for classes outside their training distribution, highlighting the limitation of a purely discriminative approach.

2.2 Retrieval-Augmented Generation (RAG) and Hybrid Approaches

To mitigate the hallucination problem and enhance the factuality of generated text, the **Retrieval-Augmented Generation (RAG)** paradigm [5] was introduced. RAG ensures that the generation process is grounded in information retrieved from a trusted, external knowledge store. Our Classifier–DB–LLM (CRG) Hybrid Pipeline adapts this philosophy to the vision domain. By decomposing the task into a lightweight Classification step, a Retrieval step from a curated database, and a small Generation step, our approach seeks to achieve an optimal tradeoff between the efficiency of small models and the factual reliability provided by external knowledge.

3 CRG Pipeline

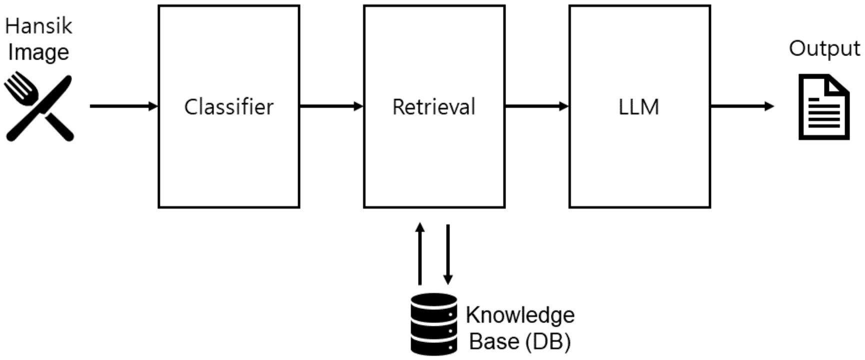


Figure 2: Description for the CRG pipeline

The CRG pipeline operates in three sequential, specialized stages, as illustrated in Figure 2.

3.1 Stage 1: Classification

The process begins with an input image being fed into a lightweight visual classifier. The primary task here is rapid and accurate identification of the food item. To identify the optimal model for this resource-constrained task, we conducted preliminary evaluations considering major high-performance architectures, including ResNet [4], EfficientNet [6], and the Vision Transformer (ViT) [3]. The result of this stage is the predicted dish name.

3.2 Stage 2: Retrieval

The identified dish name from the Classification stage acts as the key for the Retrieval module. This module accesses a carefully curated **Knowledge Base (DB)**, which stores structured, fact-grounded information about each dish. This knowledge base is implemented as a simple Key-Value mapping, where the dish name is the key. The stored data follows a rich structure, including fields such as `korean_name`, `english_name`, `description`, `category`, `ingredients`, `cooking_method`, and `cultural_note`. The retrieved factual information is then formatted as a structured prompt (e.g., a JSON or structured text block) to be passed to the generative model. This explicit retrieval step is crucial for guaranteeing high factual accuracy and mitigating the hallucination inherent in large language models.

3.3 Stage 3: Generation

In the final stage, a small Language Model (LLM) is used to generate the final, fluent explanation. We utilize a `TinyLlama/TinyLlama-1.1B-Chat-v1.0` as our generator. By supplying the LLM with the retrieved factual data, the resulting generation is augmented: its task shifts from generating potentially ungrounded knowledge to merely fluently articulating the provided facts. This addition of ground truth data is crucial for achieving high description quality and factual richness while adhering to our resource constraints.

4 Experiments

4.1 Dataset

For our experiments, we utilize the image collection from the `AI-Hub Korean Food Image Dataset` [1]. This medium-scale dataset, originally intended for fine-grained classification, is composed of 150,000 images evenly distributed across 150 distinct categories of Korean dishes with 1,000 images per category.

The original dataset provides all metadata and class labels in Korean. As a preliminary step, we translated all necessary textual information into English. It is crucial to note that the original AI-Hub dataset consists only of images and their corresponding class labels. It does not contain the detailed, descriptive paragraphs that our CRG model is tasked with generating. Therefore, We generated the additional information for Korean food using OpenAI GPT-5.

4.2 Determining optimal classifier

Model Name	Train Acc (%)	Test Acc (%)
ResNet50 (No techniques for overfitting)	90.61	77.49
ResNet50 (Regularization + LS)	99.67	84.78
ResNet50 (Regularization + Mixup + LS)	74.36	84.03
EfficientNetB0 (Regularization + LS)	98.85	85.96
EfficientNet B1 (Regularization + LS)	98.47	85.44
EfficientNet B2 (Regularization + LS)	99.20	86.32
EfficientNet B3 (Regularization + LS)	99.32	**86.36**
EfficientNet B4 (Regularization + LS)	97.66	85.08
EfficientNet B5 (Regularization + LS)	99.39	85.12
ViT base (Regularization + LS)	99.70	77.60
ViT Small (Regularization + LS)	81.57	84.11
ViT Tiny (Regularization + LS)	74.79	79.17

Table 1: Model Performance: LS means Label Smoothing

We finetuned a baseline classifiers including ResNet, EfficientNet, and ViT using AI-Hub Korean Food Image Dataset [1]. We evaluated those architectures with different regularization techniques to identify the optimal model for our dataset. The results are summarized in Table 4.2.

4.2.1 The Importance of Regularization

The ResNet50 architecture provides a clear case study in the necessity of preventing overfitting. Our baseline ResNet50 (No techniques for overfitting) achieved a high training accuracy of 90.61%

but a low test accuracy of 77.49%, a performance gap of over 13%. This indicates the model was memorizing the training data rather than learning generalizable features.

To combat this, we introduced standard regularization techniques (e.g., weight decay) and Label Smoothing (LS). This combination in ResNet50 (Regularization + LS) dramatically improved test accuracy to 84.78%. However, the training accuracy also rose to 99.67%, widening the gap and signaling that the model was still severely overfitting.

Finally, we experimented with adding Mixup, a strong augmentation and regularization method. As seen in the ResNet50 (Regularization + Mixup + LS) model, this addition significantly reduced the training accuracy to 74.36%, successfully closing the gap between training and test performance. However, the final test accuracy saw a slight decrease to 84.03% (compared to 84.78% without Mixup). This suggests that while Mixup was effective at reducing overfitting, it was not beneficial for improving the model's final generalization performance in this task.

4.2.2 Comparative Analysis of Architectures

ResNet: As noted, applying Regularization + LS achieved a strong test accuracy of 84.78%. While adding Mixup reduced overfitting, it did not improve the final score (84.03%), indicating that the simpler regularization scheme was optimal for ResNet.

EfficientNet: The EfficientNet family of models, trained with Regularization + LS, proved to be the most effective. Performance scaled positively from EfficientNetB0 (85.96%) up to EfficientNet B3 (86.36%). Performance plateaued and slightly decreased with larger models like B4 and B5. This suggests that EfficientNet B3 represents the best capacity-performance trade-off for this specific task. All models in this family showed high training accuracy (94-99%), indicating they were also overfitting, yet their architectural efficiency translated to better test performance.

Vision Transformer (ViT): The ViT models showed a strong dependency on model size. The ViT base model, despite regularization, overfit severely (99.70% train vs. 77.60% test), performing worse than the baseline ResNet. In contrast, the ViT Small and ViT Tiny models showed excellent generalization, with training accuracies (81.57% and 74.79%) very close to their test accuracies (84.11% and 79.17%). The ViT Small model was competitive with the best ResNet50 result.

4.2.3 Best Performing Model

Across all experiments, the EfficientNet B3 (Regularization + LS) model achieved the highest test accuracy at 86.36%. While smaller ViT models demonstrated better generalization by avoiding overfitting, no model surpassed the peak accuracy of EfficientNet B3. Therefore, we selected EfficientNet B3 as the best-performing model for our task.

4.3 Qualitative Results

To assess the practical quality of our Classifier-Retriever-Generator (CRG) model, we present a selection of generated explanations for various Korean food images in Table 4.3. Our qualitative analysis demonstrates that the Classifier-Retriever-Generator (CRG) model is highly effective at producing accurate, fluent, and contextually rich descriptions of Korean food images. As shown in the examples for "Bibimbap" and "Fried Chicken," the system consistently executes its pipeline with precision. The classifier first correctly identifies the dish, which then allows the retriever to gather specific, relevant facts, such as key ingredients (seasonal vegetables, crispy coating) and important cultural context (harmony and balance, chimaek).

This successful retrieval of information is crucial, as it provides the factual grounding for the generator. The 1.1B LLM, despite its small size, then skillfully synthesizes these retrieved facts into a single, coherent, and encyclopedia-like paragraph. The high quality of these generated texts validates the core strength of our CRG architecture: it effectively leverages a compact generator to produce outputs that are not only fluent but also factually dense and accurate, a task that would typically be a significant challenge for a model of its size.

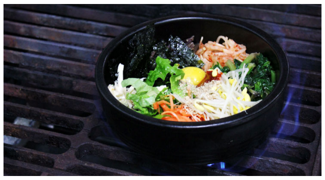
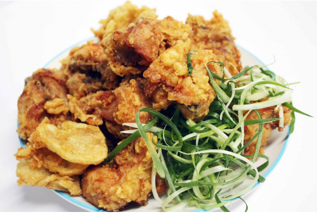
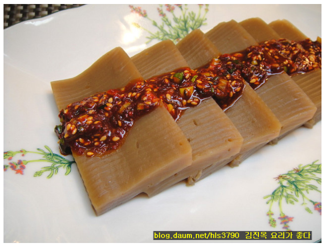
Image	Generated text
	Bibimbap is a popular Korean rice dish that is both delicious and visually appealing. It combines various seasonal vegetables, spicy beef, and crispy fried eggs on top of a hot bowl of steaming rice. This dish has become synonymous with Korean cuisine, representing harmony and balance through its vibrant colors and textures.
	Fried Chicken is a beloved Korean dish that has gained immense popularity overseas. It's known for its signature crispy coating made from a combination of flour, cornstarch, and various seasoning blends. This delicious treat is typically served with beer (chimaek), which adds to the overall experience.
	As a Korean food expert, I am thrilled to share with you my knowledge about one of Korea's most beloved dishes - Acorn Jelly. Acorn Jelly is a unique and delicious Korean dish that combines traditional Korean flavors and techniques with modern cooking methods. This dish is often served as an appetizer or side dish at traditional Korean BBQ or banchan parties.

Table 2: Generated texts from sample images

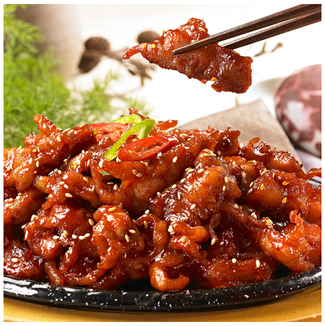
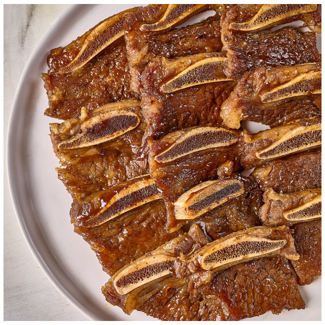
Image	Generated text
	Marinated Grilled Beef is a beloved Korean dish that has been enjoyed by generations of Koreans. The name "bear meat" refers to the distinctive smoky flavor that comes from the marinade, which is made up of a blend of soy sauce, sugar, garlic, and sesame oil.
	Grilled Short Ribs is a popular Korean BBQ dish that combines the richness of marinated beef with the savory umami taste of soy sauce. These tender cuts of beef are grilled until tender and juicy, with a crispy exterior that adds to their delicious flavors. The key to making these ribs stand out is using the right marinade recipe that combines sweet, salty, and spicy flavors.

Table 3: Failure cases of our CRG pipeline.

4.4 Failure Case Analysis

The examples in Table 4.4 reveal the primary weakness of our CRG pipeline: its total dependence on the initial classifier's accuracy. Any error in this first step propagates through the system, leading to irrelevant or incorrect text generation. We observed two main types of failure:

- **Out-of-Distribution (OOD) Classes:** The first example, *Dakbal* (chicken feet), is a dish that was not included in our 150-category training dataset. The classifier is therefore forced to make an incorrect prediction, "Marinated Grilled Beef," based on visual similarity (e.g., red, grilled appearance). The CRG pipeline then correctly retrieves facts and generates a fluent description for the wrong dish.
- **High Inter-Class Similarity:** Another two example show the classifier confusing visually and semantically similar dishes. *LA Galbi* is misclassified as "Grilled Short Ribs," a closely related but different class.In these cases, the generated text is plausible but factually wrong, failing to capture the specific details of the input image.

These failures highlight that our current model operates as a closed-set system, which is a critical limitation for real-world scalability. To address this, our future work will focus on developing an open-set model by replacing the fixed classifier with a **zero-shot inference** mechanism (e.g., using a CLIP-based image-to-text embedding model). This will allow the retrieval module to find relevant information for any food image, even those not seen during training.

5 Discussions

5.1 Limitations

- **Low Accuracy of The Classifier**: Our first-stage classifier caps end-to-end performance. With 86.36% top-1 accuracy on held-out data, mistakes at this gate propagate: a wrong label retrieves the wrong knowledge entry and conditions the generator on incorrect context. Errors concentrate on fine-grained, visually similar dishes (e.g., variants that differ by garnish or broth opacity), and are exacerbated by inconsistent plating across sources.
- **Lack of Zero-shot Capability**: The current model is strictly closed-set. Unseen dishes are forced into the nearest seen class with high confidence, which harms robustness in realistic settings (regional specials, seasonal items, fusion variants). The system also lacks a calibrated mechanism to flag out-of-distribution (OOD) inputs, so uncertainty cannot be surfaced or routed to safer fallbacks.

5.2 Future Works

- **Adopt CLIP to Improve Accuracy and Zero-Shot Capability**: We will replace the current closed-set classifier with a CLIP-based recognizer to improve accuracy. CLIP evaluates images against text in a shared image–text space; by scoring each image against concise dish-name prompts, it can leverage linguistic cues—ingredients, form, serving style—to disambiguate fine-grained, visually similar dishes that currently dominate our errors. Building on this, we will operate CLIP in zero-shot mode to extend coverage beyond the original label set without task-specific training, reducing forced misclassification of unseen dishes and improving robustness on long-tail and newly introduced items.
- **Attribute-Aware Retrieval for Zero-Shot Usability**: To ensure zero-shot predictions remain useful when the knowledge-base (DB) lacks an exact entry, we will upgrade retrieval to attribute-aware matching while preserving the existing key→value structure. Each knowledge-base entry will be augmented with a compact attribute string (e.g., broth color/opacity, noodle/rice type, spiciness, primary protein, garnish, vessel). At query time, we will rank entries by a weighted combination of similarity to the predicted label and similarity between the image and each entry's attribute string (using CLIP's text encoder), then pass the top match to generation. This minimal change is designed to raise accuracy on seen classes and yield sensible, near-miss matches for novel dishes without broader schema changes.

6 References

[1] AI-Hub (National Information Society Agency). *Korean Food Image*. `https://aihub.or.kr/aihubdata/data/view.do?srchOptnCnd=OPTNCND001&currMenu=115&topMenu=100&searchKeyword=%ED%95%9C%EC%8B%9D&aihubDataSe=data&dataSetSn=79`. Accessed: 2025-11-03. 2018.

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. "Food-101–mining discriminative components with random forests". In: *European conference on computer vision*. Springer. 2014, pp. 446–461.

[3] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[4] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[5] Patrick Lewis et al. "Retrieval-augmented generation for knowledge-intensive NLP tasks". In: *Advances in neural information processing systems* 33 (2020), pp. 9459–9474.

[6] Mingxing Tan and Quoc Le. "EfficientNet: Rethinking model scaling for convolutional neural networks". In: *International conference on machine learning*. 2019, pp. 6105–6114.