Fine-grained Food Image Classification of Korean Foods

Seoyun Jang
Data Science
2023-22227
syjang115@snu.ac.kr

Sungju Jang
Landscape Architecture & Rural Systems
2022-35543
wkdtjdwn@snu.ac.kr

Yohan Yoon
Mechanical Engineering
2018-17327
strauss2204@snu.ac.kr

Abstract

Accurate dietary data is essential for health status assessment. Traditional dietary data collection methods heavily relies on human memory, thus necessitating real-time data collection through food images. However, there has been limited research on models for Korean food classification. In this study, we applied the Context-Aware Attention Pooling (CAP) model, which has demonstrated state-of-the-art performance on international food datasets, to Korean food data. Considering the prevalence of red-seasoned dishes and various types of soup in Korean cuisine, we proposed a specialized augmentation methodology tailored for Korean foods. Experimental results showed that our model outperformed the previously high-performing InceptionResNetV2 model in Korean food classification. This study highlights the significance of incorporating attention mechanisms for Korean food classification and underscores the necessity of specialized augmentation techniques for this purpose.

1. INTRODUCTION

Assessing dietary intake is essential for managing chronic diseases and delivering public healthcare services. Historically, food frequency questionnaires (ex. the Korea National Health and Nutrition Examination Survey) and 24-hour recalls have been employed to gather dietary data. Yet, these methods are susceptible to inaccuracies due to reliance on participants' memory and numerical abilities [1]. Instead, a real-time capture of diet information through meal photos could significantly enhance accuracy. This is why many studies have proposed various methods for classifying food images of diverse cuisines.

However, classifying food images poses challenges due to subtle visual differences between food classes. A slight change in ingredients or cooking method can lead to different classification results. For example, Linguine and Fettcuine just differ by the thickness of pasta [2]. These differences also vary by region, which calls for the use of localized image datasets to obtain more accurate results.

Figure 1. Images of Kimchi-Jjigae, Dongtae-Jjigae, Chueotang, Sundubu-Jjigae (from left to right)

In the case of traditional Korean recipes, the most significant challenge arises when classifying such as soups (Jeongol, Tang, Jjigae) and side dishes (Jorim, Muchim) due to combinations of different foods or similar shapes and colors [3]. This complexity arises from the diverse ingredients and cooking methods used in Korean cuisine, leading to different images for the same dish. Even with the human eye, it is difficult to accurately classify these food categories without prior knowledge of the ingredient composition. Since each ingredient contains different nutritional components, misclassifying subclasses could lead to issues in dietary assessment, emphasizing the need for a fine-grained classification model. However, an open-source model tailored for Korean recipes was yet to be found.

Since each food has different nutritional components, inaccurate food classification can lead to erroneous dietary assessments. The objective of this study is developing a "specialized fine-grained image classification model for Korean foods" to enhance the accuracy of Korean food image classification evaluations. In this study, the Context-Aware Pooling (CAP) model, currently evaluated as the SoTA model on the Food 101 dataset, was used. Then, appropriate data augmentation methods for fine-grained image classification of Korean food images were applied to evaluate the top-1 accuracy. Subsequently, the performance of the evaluated model was compared with that of existing Korean food image classification models.

2. RELATED WORK

2.1. Food Image Classification of Korean Recipes

Various studies have been conducted on the classification of Korean food images using image classification models. Jun et al. compared the performance of four pre-trained CNN models on 150,000 Korean food images classified into 150 categories, after augmenting them with horizontal flip, zoom, and shift techniques, using ImageNet data. Although DenseNet achieved the highest accuracy of 82.17%, it showed lower accuracy in Korean food images compared to applying CNN models to UEC-Food-100 and UEC-Food-256 datasets [4]. Park et al. constructed a dataset about 4,000 Korean food images classified into 23 food groups and augmented them with random contrast, brightness, sharpness, and color changes before evaluating the performance using their own CNN-based model "K-FoodNet" with DCNN (Deep Convolutional Neural Network). K-FoodNet demonstrated high accuracy of 91.3%, but the data was restricted to certain dishes (rice, soup, kimchi, pork) and the total number of the dataset was insufficient for training [1]. Chun et al. evaluated the performance of six pre-trained CNN models using approximately 150,000 images, with preprocessing to reduce network complexity and applying augmentation techniques such as brightness, saturation, contrast, and horizontal flip, using ImageNet data. InceptionResNetV2 showed the highest accuracy of 81.9% of six models. However, it suffered significant performance degradation in specific food types such as Jjigae, as shown in Figure 1. Also, there was a problem where performance decreased further when augmented compared to before [3].

2.2. Fine-grained Food Image Classification Models

Various approaches exist for fine-grained image classification, including localization-classification subnetworks, end-to-end feature encoding, and leveraging external information [5]. Localization-classification subnetworks involve designing subnetworks to identify key parts through techniques like detection or segmentation, utilizing deep filters, and leveraging attention mechanisms to obtain part-level feature vectors. End-to-end feature encoding learns integrated and distinctive representations through performing high-order feature interactions, designing new loss functions, etc., to model subtle differences between fine-grained categories. Recognition with external information utilizes external information through noisy web data, multi-modal data, humans-in-the-loop methods, etc.

Several studies have applied fine-grained image classification to the food domain [6]. To improve classification accuracy, methods have been proposed using the Cross-Shaped Window-Large architecture (CSWin-L), a transformer that uses horizontal and vertical self-attention based on learning from Multi-subset classes [7]. The Dining on Details (DoD) model utilizing Large Language Models (LLM) demonstrated outstanding accuracy of 94.9% when SwinV2-T was used as the backbone model [2]. The A Large-scale ImaGe and Noisy-text embedding (ALIGN) model extends the learning of visual and vision-language representations [8]. Sharpness-Aware Minimization (SAM) improves model generalization by finding parameters with uniformly low loss values around them instead of parameters with low loss values, minimizing both loss value and loss sharpness [9]. Context-aware Attentional Pooling (CAP) applies context-aware attention to feature pooling to recognize subcategories, enabling the identification of important areas for classification without the need for separate bounding boxes or annotations [10]. According to benchmark results on the Food 101 dataset, the current state-of-the-art (SoTA) model is CAP and it achieves an accuracy of 98.6% [6].

2.3. Data Augmentation for Fine-grained Recognition

The Vision Transformer (ViT-B) requires large datasets and may encounter overfitting when datasets are small. To address this, augmentation is necessary to supplement the data. For food image data, augmentation methods include flip, rescale, rotation, shift, uncertainty-aware data augmentation (UDA), generative adversarial network (GAN), and class activation mapping (CAM)-based sample enhancement [11–15]. Previous studies on Korean food image classification have applied augmentations such as horizontal flip, zoom, shift, contrast, brightness, sharpness, color change, and saturation [1, 3, 4]. Behera et al. applied random rotation, scaling, and cropping augmentation to the CAP model [10]. Random data augmentation presents efficiency issues and can generate noisy data [16]. To address this, methods like Auto-Augmentation [17] and Adversarial Data Augmentation [18], which consider the distribution of data to select appropriate augmentations, have been proposed. Typically, automation methods like AutoAugment [17] and RandAugment [19] are used with ViT-B [20]. However, they incur significant experimental costs. Methods like TrivialAugment [21], 3-Augment [22], RandomErasing [23], and the combination of these methods known as Augmentplus [20] have been proposed to address these issues. Furthermore, when objects are small, excessive noise from the background can reduce efficiency. Recognizing the spatial information of target objects is necessary to address this. For fine-grained image classification, methods like Weakly Supervised Data Augmentation Network (WS-DAN) generate attention maps based on the spatial information of objects and perform attention cropping and dropping on these areas [16]. Subsequent research has

Figure 2. Grad-CAM results of Korean food classification using an existing state-of-the-art model, InceptionResNetV2

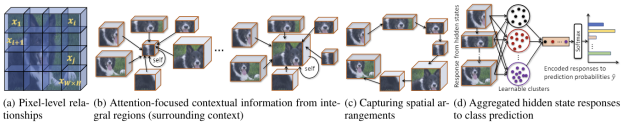


Figure 3. Novel approaches proposed in CAP for fine-grained image classification [10]

applied attention concepts to existing augmentation methods, leading to methods like Attention-Guided CutMix Data Augmentation Network (AGCN) [24] and Attention-based Cropping and Erasing Network (ACEN) [25].

2.4. Limitations of Previous Studies

Before commencing our investigation, we aimed to evaluate the performance of the InceptionResNetV2 model, which had the highest accuracy in previous studies focusing on the classification of Korean foods using the same dataset sourced from the AI-Hub platform. To assess the classification performance, we implemented Grad-CAM. Grad-CAM addresses the limitation of losing spatial information of features due to the flattening of the last fully connected layer in CNN architectures, making it impossible to discern which parts of an image contributed to the classification. By utilizing the gradient information flowing through the last CNN layer, Grad-CAM generates heatmaps to visually understand the importance of each neuron related to decision-making. Upon training with InceptionResNetV2, our results corroborated the previously mentioned findings of significantly low accuracy in classifying stew-type dishes. This was further confirmed by Grad-CAM analysis of instances where Kimchi-Jjigae and Sundubu-Jjigae was wrongly classified as Dongtae-Jjigae (Figure 2), which revealed a lack of focus on crucial parts of the food. This underscores the necessity for a model tailored specifically to Korean cuisine classification.

3. METHOD

3.1. Context-Aware Attention Pooling (CAP)

In this paper, we examine the effectiveness of CAP, the SoTA fine-grained food image classification model, for Korean food image recognition.

CAP begins by taking the convolutional feature map x of size $W \times H \times C$, which is the output of a base CNN model. For our study, we used the Xception architecture as the base CNN, as it demonstrated the best performance among six architectures across multiple datasets, as reported in [10].

Next, CAP integrates contextual information hierarchically from pixel-level to image-level features. First, self-attention is applied directly to x to encode pixel-level relationships (Figure 3a).

Second, multiple integral regions with varying sizes and aspect ratios are proposed to capture higher-level contextual information. To handle size variations between regions, bilinear pooling is applied. After representing the regions as feature vectors with a fixed size of $w \times h \times C$, an attention mechanism is applied again to focus on relevant integral regions, generating more comprehensive contextual information (Figure 3b).

Third, CAP uses LSTM to extract structural knowledge from internal hidden states (Figure 3c). This approach is inspired by the way humans attend to different parts sequentially to extract relevant information. To improve generalization and reduce computational complexity, global average pooling (GAP) is applied beforehand, reducing the feature map size to $1 \times C$.

Finally, to further guide the model in distinguishing subtle differences, learnable pooling is applied (Figure 3d). CAP utilizes an adapted version of a feature encoding method named NetVLAD, proposed in [26]. This method effectively groups similar hidden state responses into clusters, which are then mapped into the final prediction probabilities of each class.

In summary, CAP integrates contextual details across scales by leveraging attention mechanisms and LSTM. We expect CAP to improve fine-grained food recognition accuracy for Korean food images beyond the 81.91% achieved with InceptionResNetV2 in [10].

3.2. Augmentation Method for Korean Food

To facilitate data augmentation for fine-grained classification of Korean food images, it is imperative to comprehend the distinctive features of Korean cuisine. First, Korean food is typically served in bowls, and most food photographs are taken with the bowls placed on a dining table. For soups, it is common to plate the dish by placing the solid ingredients in the center of the bowl before taking the picture. Consequently, we applied a center cropping augmentation method. However, not all food images have the food positioned exactly in the center. Therefore, before performing a center crop, we applied random rotations within ±30 degrees, followed by random translations within ±30 pixels along both the x and y axes.

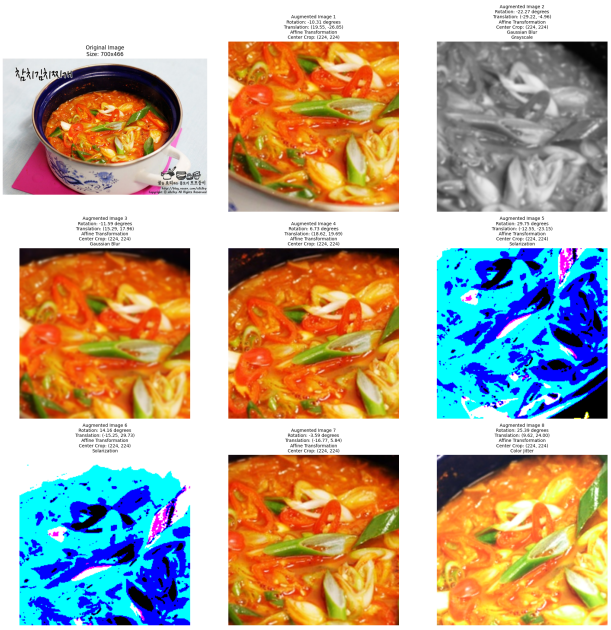


Figure 4. Examples of applied augmentations. Eight different augmentation results from a single kimchi-jjigae photograph are shown. After applying rotation, translation, and center cropping, the color augmentation techniques were applied randomly.

Second, many Korean dishes use red pepper powder or red pepper paste as seasoning, giving them a distinctive red color. Despite having the same color, the type of dish varies depending on the ingredients. For instance, Kimchi-Jjigae and Dongtae-Jjigae cannot be distinguished by their broth color but by the presence of their main ingredients: Kimchi and Dongtae, respectively. This issue is not limited to Jjigae but is also common among various types of Kimchi. To differentiate between dishes with similar colors but different ingredients, we need augmentations that focus less on color information and more on identifying the ingredients as in [20]. Hence, we applied color augmentation techniques that prioritize the shapes of the ingredients over their colors. The color augmentations included random applications of grayscale conversion, solarization, color jittering, and Gaussian blurring.

Figure 4 demonstrates examples of the applied augmentations. It shows eight augmentation results from a single Kimchi-Jjigae image. After rotation, translation, and center cropping of all train images, color augmentation techniques were randomly applied, resulting in eight different augmented images. This novel augmentation approach is expected to effectively differentiate subtle differences between Korean food sub-classes and prevent overfitting.

대분류	소분류	대분류	소분류
구이	갈비구이,갈치구이,고등어구이,곱창구이,닭갈비,더덕구이,떡갈비,불고기,삼겹살,장어구이,조개구이,황태구이,훈제오리	국	계란국, 떡국/만두국, 무국, 미역국, 북엇국, 소고기무국, 시래기국, 육개장, 콩나물국
김치	깻김치, 깍두기, 나박김치, 무생채, 배추김치, 백김치, 부추김치, 열무김치, 오이소박이, 총각김치, 파김치	나물	가지볶음, 고사리나물, 미역줄기볶음, 숙주나물, 시금치나물, 애호박볶음
떡	경단	만두	만두
면	막국수, 물냉면, 비빔냉면, 수제비, 열무국수, 잔치국수, 쫄면, 칼국수, 콩국수, 라면, 자장면, 짬뽕	무침	고추튀김무침, 도라지무침, 콩나물무침, 콩어무침
밥	김밥, 김치볶음밥, 비빔밥, 새우볶음밥, 알밥, 잡곡밥, 주먹밥, 유부초밥	볶음	건새우볶음, 오징어채볶음, 감자채볶음, 고추장진미채볶음, 두부김치, 떡볶이, 라볶이, 멸치볶음, 소세지볶음, 어묵볶음, 제육볶음, 쭈꾸미볶음
쌈	보쌈	음청류	수정과, 식혜

Figure 5. Classes in the AIHub Korean Food Image Dataset

4. EXPERIMENTS

4.1. Dataset

This study utilizes an open-source Korean food image dataset from AIHub (https://www.aihub.or.kr/), specifically curated based on the Korean Food Menu Foreign Language Guide published by the Korea Advanced Institute of Science and Technology (KAIST) in collaboration with the Ministry of Agriculture, Food and Rural Affairs and the Korean Food Foundation. The dataset categorizes Korean food images into 27 main classes and 150 sub-classes, comprising approximately 1,000 images per sub-class, totaling 150,000 images. This comprehensive dataset serves as a valuable resource for training and evaluating deep learning models targeting food recognition tasks, particularly in the context of Korean cuisine.

4.2. Image Preprocessing

Image data of each sub-class in the AIHub dataset was partitioned into training and testing subsets using an 8:2 ratio. Before training and testing, each image was loaded using OpenCV (`cv2.imread`) and resized to a fixed target size of 224×224 pixels. This resizing operation ensured consistent input dimensions across all images, facilitating compatibility with convolutional neural network architectures. During image loading, each image was inspected for potential data inconsistencies. If errors such as NaN values or file corruption occurred during image processing, they were captured and specifically handled to ensure the integrity of the dataset. Undetected anomalies such as video clips, were also filtered out before model training. Out of 150 sub-classes, 2 were deleted because less than 10% were available after deleting errors. Thus, a total of 148 classes were used for training model on full dataset.

4.3. Experimental Settings

In our experimental setup, we made several modifications inspired by the original CAP paper [10]. We resized

all input images to 224×224 pixels and utilized the Adam optimizer instead of Stochastic Gradient Descent (SGD) for training the model, with an initial learning rate of 1×10^{-3}. The Adam optimizer was configured with default parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-7}$). The training duration was limited to much fewer epochs than 1,000 from [3] in both the baseline and CAP due to computational constraints. Specifically, we set the epoch to 50 for the three main food classes and 20 for the full dataset. We also applied a learning rate schedule that multiplied the initial rate by 0.1 after every 10 epochs as stated in [10]. For CAP model, image augmentation techniques including random rotation (±15 degrees), random scaling (1 ± 0.15), and random cropping were applied, consistent with [10]. However, unlike the original paper, we maintained the input image size instead of resizing to 256×256 before augmentation. Model training was conducted on an NVIDIA Titan RTX GPU (12 GB) with CUDA Version 11.8. The algorithm was implemented using TensorFlow 2, updating the software framework from TensorFlow 1 to TensorFlow 2 for improved performance and compatibility.

4.4. Results

We initially compared the test accuracies of CAP with the baseline model (InceptionResNetV2) for sub-class classification across three main food classes: Jjigae, Kimchi, and Namul. These classes were chosen for their representation of unique characteristics in Korean cuisine and were ranked as bottom, middle, and top tiers, respectively, in terms of True Positive Rate in [3].

Main Class	**Baseline (%)**	**CAP (%)**
Jjigae	68.75	**88.21**
Jjigae (new aug)	-	74.19
Kimchi	77.01	**92.02**
Namul	87.42	**99.10**
Full (ours)	54.55	80.82
Full [3]	**81.91**	-

Table 1. Test accuracies of main food classes and full dataset

Overall, our CAP model outperformed the baseline model across three main classes. In Table 1, we present a comparison of performances by main food class and tested model. While the order of accuracy remained consistent, all classes demonstrated significant improvement. It is worth highlighting that while the accuracy of the Jjigae class remained relatively low at 88.21%, both Kimchi and Namul achieved accuracies above 90%, with Namul particularly impressive at 99.10%.

In Figure 6, we compare the accuracy evolution between the Baseline and CAP models by main food classes. Both models demonstrated convergence around epoch 20 across all classes, with a slight initial instability observed in the CAP model's performance.

Additionally, we sought to improve the accuracy of the Jjigae class by applying our proposed augmentation method to the CAP model as outlined in Section 3.2. Prior to the experiment, we hypothesized that this augmentation would enhance performance by enabling the model to focus on the shapes of ingredients rather than the similar colors of the dishes. Furthermore, we expected an increase in performance due to the enlarged training dataset resulting from the augmentation.

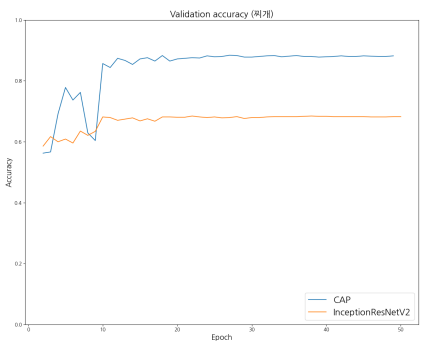
However, as illustrated in Table 1, the accuracy decreased from 88.21% to 74.19% when using the new augmentation compared to the default augmentation of the CAP model. The degradation in performance observed with color augmentation including grayscale conversion and solarization suggests that color information plays a crucial role in food classification than we first expected. Therefore, further analysis and refinement of the proposed augmentation method are necessary.

Lastly, when fitted on the full dataset of 148 food main classes, the model achieved an accuracy of 80.82%, which closely followed but did not exceed the best accuracy of 81.91% reported by the baseline model in [3] (Table 1). It is important to note that our experimental setting was much more limited, being restricted to only 20 epochs, and accuracy of InceptionResNetV2 remained at a very low 54.55% in the same setting.

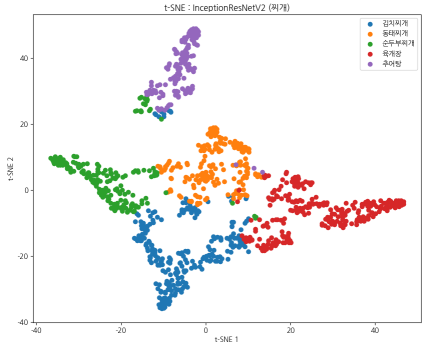
4.5. Qualitative Analysis

For qualitative analysis, t-SNE was applied to both baseline and CAP model outputs to compare and visualize the class separability and compactness in features. As depicted in Figure 6, it is evident that the clusters show greater separation and cohesion in the CAP model compared to the baseline. This distinction contributes to a clearer identification of clusters representing sub-classes in CAP than baseline.

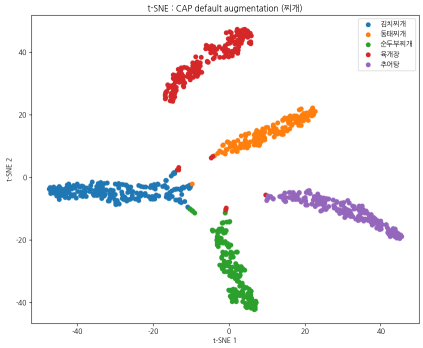
Also, in the t-SNE visualizations of CAP on three main classes, we observed notable clusters that were close in proximity or insufficiently distinguished. Within the Jjigae category, clusters representing Kimchi-Jjigae and Sundubu-Jjigae were observed to be closely grouped together (Figure 6c). This can be attributed to their similar color and overlapping ingredients, with both dishes incorporating tofu (sundubu) and kimchi. Similarly in the Kimchi category, some instances of Chongak-Kimchi were found to be located closer to cluster of Ggakdugi (Figure 6f), possibly due to similar color and texture of main ingredients. In the Namul category, the clusters were well separated (Figure 6i), likely contributing to the high accuracy of 99.1%.



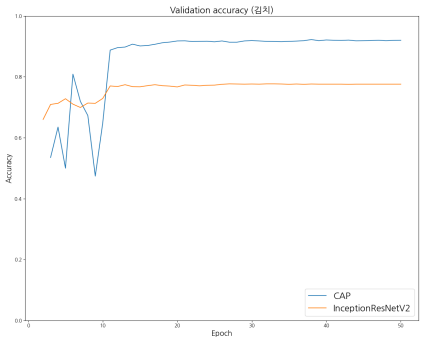
(a) Accuracy by epoch (Jjigae)



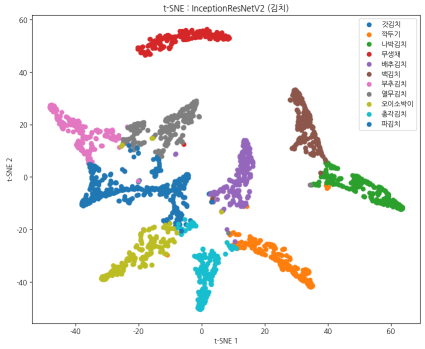
(b) t-SNE of InceptionResNetV2 (Jjigae)



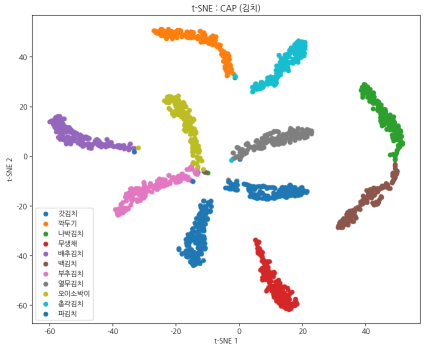
(c) t-SNE of CAP (Jjigae)



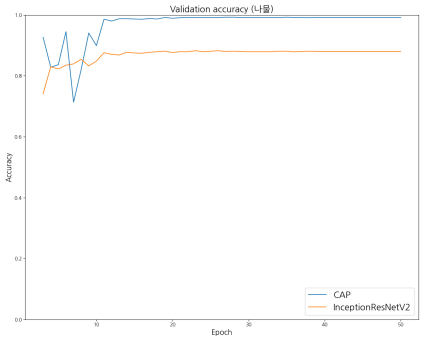
(d) Accuracy by epoch (Kimchi)



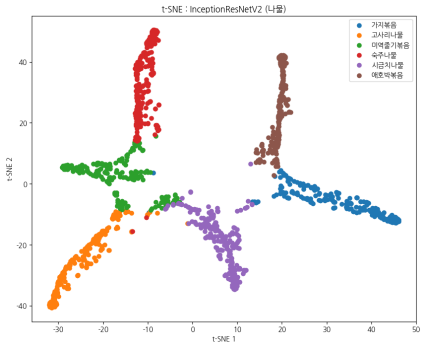
(e) t-SNE of InceptionResNetV2 (Kimchi)



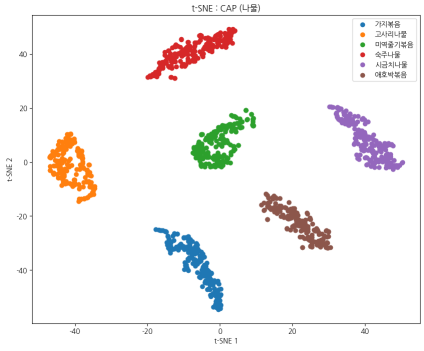
(f) t-SNE of CAP (Kimchi)



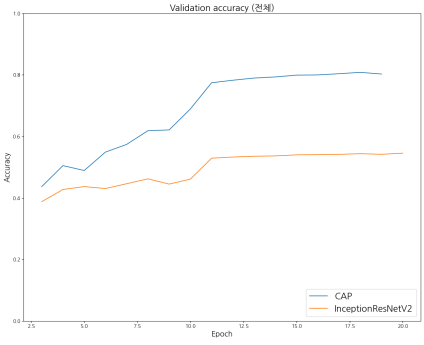
(g) Accuracy by epoch (Namul)



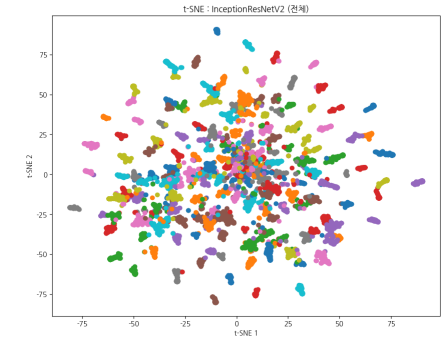
(h) t-SNE of InceptionResNetV2 (Namul)



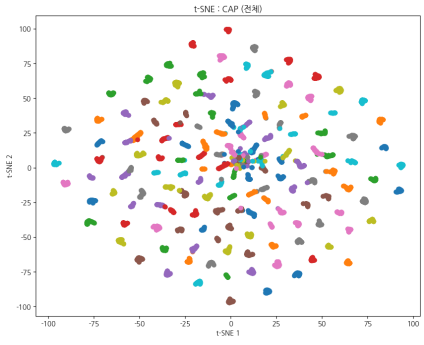
(i) t-SNE of CAP (Namul)



(j) Accuracy by epoch (Full data)



(k) t-SNE of InceptionResNetV2 (Full Data)



(l) t-SNE of CAP (Full Data)

Figure 6. Accuracy evolution and t-SNE visualization: Baseline vs. CAP model across main food classes & full dataset

5. CONCLUSION

5.1. Significance of Study

In this study, we proposed an approach to improve the accuracy of Korean food image classification models by applying the CAP model, a SoTA model for the Food-101 dataset, and incorporated augmentations specific to Korean food images. We focused on three main classes identified in prior research on Korean food image classification: "Namul" (which had relatively high classification TPR in previous work), "Kimchi" (moderate TPR in previous work), and "Jjigae" (lower TPR in previous work). For each main class, we classified the sub-classes and compared the results with those obtained using the InceptionResNetV2 model employed in a previous study on the same dataset. The significance of this study lies in the application of a model with an attention mechanism instead of traditional CNN-based models for Korean food image classification, and the introduction of augmentation methods tailored to the challenging class "Jjigae." The results of this study are expected to contribute to the development of Korean food image classification model.

5.2. Limitations and Future Works

The research encountered several limitations that affected the performance and results. First, there were errors in the dataset. The dataset was not perfectly refined, making it difficult to distinguish certain images. Examples of such errors include images of restaurant interiors without any food as in Figure 7, pictures of ingredients before they were cooked, or completely unrelated images. These errors, likely introduced during the process of crawling photos from blogs, hindered the training process. Given the difficulty of manually filtering out these erroneous images, a more refined and carefully curated dataset is necessary for future work.

Second, there are edge cases that the current model struggles as the examples in Figure 7. Since the model is designed for classification rather than object detection, it has difficulty classifying images that contain multiple dishes. Furthermore, distinguishing between dishes is challenging when Jjigae is boiling at a high temperature, as the contents are not visible and only the color of the broth can be used for differentiation. The model also faces difficulties when the food occupies only a small portion of the image.

Third, there were constraints in the experimental environment. Although a server GPU was used, the restriction on usage time (4 hours) prevented sufficient training of the CAP model on the entire food dataset. Due to memory limitations, the batch size could not be increased significantly. The total training dataset comprised approximately 120,000 images (150 classes $\times$ 800 images), but the batch size could not exceed 32, resulting in prolonged training times. Consequently, due to the GPU usage time limitation, training could only be conducted for only few epochs. It is anticipated that with adequate GPU usage time and memory, the training process could be more effective.

(a) No food image (b) Several foods (c) Hard to classify

Figure 7. Examples of an invalid image (a) and edge cases (b,c)

Lastly, as mentioned in the Section 4.4, the newly proposed augmentation methodology did not enhance performance as expected; instead, it resulted in a decrease in performance compared to the CAP with default augmentation. Therefore, for future work, it is necessary to redefine the augmentation methodology. For instance, similar to the approach in [16], we could utilize attention maps to guide the augmentation process. By using the attention maps from the CAP model trained with default augmentation, we could create new augmented data by cropping or dropping the regions where the attention is focused. Iteratively training the model with these augmented datasets could help identify and emphasize the significant features for food classification, such as color, shape, and texture.

References

[1] Seon-Joo Park, Akmaljon Palvanov, Chang-Ho Lee, Nanoom Jeong, Young-Im Cho, and Hae-Jeung Lee. The Development of Food Image Detection and Recognition Model of Korean Food for Mobile Dietary Management. Nutrition Research and Practice, 13(6): 521–528, 2019. 1, 2

[2] Jesús M. Rodríguez-de-Vera, Pablo Villacorta, Imanol G. Estepa, Marc Bolaños, Ignacio Sarasúa, Bhalaji Nagarajan, and Petia Radeva. Dining on Details: LLM-Guided Expert Networks for Fine-Grained Food Recognition. In Proceedings of the 8th International Workshop on Multimedia Assisted Dietary Management (MADiMa '23), pages 43–52. ACM, New York, NY, USA, 2023. 1, 2

[3] Minki Chun, Hyeonhak Jeong, Hyunmin Lee, Taewon Yoo, and Hyunggu Jung. Development of Korean Food Image Classification Model Using Public Food Image Dataset and Deep Learning Methods. In Proceedings of IEEE, pages 128732-128741, 2022. 1, 2, 5

[4] Tae Joon Jun, Nakyung Lee, Dohyeun Kim, Hyunseob Kim, and Daeyoung Kim. Comparative performance

analysis of pre-trained convolutional neural networks in Korean food image classification. In Proceedings of the Korean Institute of Information Scientists and Engineers, pages 961-963, 2018. 2

[5] Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(12): 8927-8948, 2022. 2

[6] Paperswithcode. Fine-Grained Image Classification on Food-101. https://paperswithcode.com/sota/fine-grained-image-classification-on-food-101 2

[7] Javier Ródenas, Bhalaji Nagarajan, Marc Bolaños, and Petia Radeva. Learning Multi-Subset of Classes for Fine-Grained Food Recognition. In Proceedings of the 7th International Workshop on Multimedia Assisted Dietary Management, pages 17-26, 2022. 2

[8] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning, pages 4904-4916. PMLR, 2021. 2

[9] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. arXiv preprint arXiv:2010.01412, 2020. 2

[10] Ardhendu Behera, Zachary Wharton, Pradeep R. P. G. Hewage, and Asish Bera. Context-aware attentional pooling (CAP) for fine-grained visual classification. In Proceedings of the AAAI Conference on Artificial Intelligence, 3(2): 929-937, 2021. 2, 3, 4, 5

[11] Eduardo Aguilar, Bhalaji Nagarajan, Rupali Khantun, Marc Bolaños, and Petia Radeva. Uncertainty-aware data augmentation for food recognition. In 2020 25th International Conference on Pattern Recognition (ICPR), pages 4017-4024. IEEE, 2021. 2

[12] Jiangpeng He, Luotao Lin, Jack Ma, Heather A. Eicher-Miller, and Fengqing Zhu. Long-tailed continual learning for visual food recognition. arXiv preprint arXiv:2307.00183, 2023.

[13] Sirawan Phiphiphatphaisit and Olarik Surinta. Food image classification with improved MobileNet architecture and data augmentation. In Proceedings of the 3rd International Conference on Information Science and Systems, pages 51-56, 2020.

[14] A. Sivaranjani, S. Senthilrani, B. Ashokumar, and A. Senthil Murugan. CashNet-15: An optimized cashew nut grading using deep CNN and data augmentation. In Proceedings of the 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), pages 1-5, 2019.

[15] Liu Zhang, Qing Nie, Haiyan Ji, Yaqian Wang, Yaoguang Wei, and Dong An. Hyperspectral imaging combined with generative adversarial network (GAN)-based data augmentation to identify haploid maize kernels. Journal of Food Composition and Analysis, 106: 104346, 2022. 2

[16] Tao Hu, Honggang Qi, Qingming Huang, and Yan Lu. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. arXiv preprint arXiv:1901.09891, 2019. 2, 7

[17] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 113-123, 2019. 2

[18] Xi Peng, Zhiqiang Tang, Fei Yang, Rogerio S. Feris, and Dimitris Metaxas. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2226-2234, 2018. 2

[19] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 702-703, 2020. 2

[20] Xinle Gao, Zhiyong Xiao, and Zhaohong Deng. High accuracy food image classification via vision transformer with data augmentation and feature augmentation. Journal of Food Engineering, 365: 111833, 2024. 2, 4

[21] Samuel G. Müller and Frank Hutter. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 774-782, 2021. 2

[22] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit III: Revenge of the ViT. In Proceedings of the European Conference on Computer Vision, pages 516-533, 2022. 2

[23] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, 34(7): 13001-13008, 2020. 2

[24] Wenming Guo, Yifei Wang, and Fang Han. Attention-Guided CutMix Data Augmentation Network for Fine-Grained Bird Recognition. In Proceedings of the 2021 2nd International Conference on Artificial Intelligence and Information Systems, pages 1-5, 2021. 3

[25] Jianpin Chen, Heng Li, Junlin Liang, Xiaofan Su, Zhenzhen Zhai, and Xinyu Chai. Attention-based

cropping and erasing learning with coarse-to-fine refinement for fine-grained visual classification. Neurocomputing, 501, pages 359-369, 2022 3

[26] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 3