



Applying Bootstrapping Language-Image Pre-training for Nutrition Detection from Food Images

Ida Bagus Putra Manuaba¹, I Wayan Suasnawa²,
and Komang Ayu Triana Indah³

^{1,2,3} Information Technology Department, Politeknik Negeri Bali, Bali, Indonesia
manuabaputra@pnb.ac.id

Abstract. Accessing accurate and comprehensive food calorie and nutrition data remains a challenge, as existing references are often incomplete, outdated, or difficult to access. This limitation reduces the effectiveness of daily dietary monitoring for both individuals and healthcare professionals. This study proposes a method that integrates Bootstrapping Language-Image Pre-training (BLIP) and the Large Language Model (LLaMA) to automatically detect food calories, providing broader coverage of nutritional data. Bootstrapping Language-Image Pre-training generates textual descriptions from food images uploaded by users, which the Large Language Model then processes to estimate key nutritional values, including calories, protein, fat, and carbohydrates. The experimental results demonstrate that this multimodal approach improves both accuracy and efficiency compared to conventional methods. The proposed method offers a practical tool for individuals to independently track daily nutrition and provides healthcare professionals with adaptive AI-based nutritional analysis, contributing to the advancement of digital health technologies in nutrition monitoring and dietary management.

Keywords: Bootstrapping Language-Image Pre-training, Food Calorie Detection, LLaMA, Serverless Architecture

1 Introduction

Obtaining accurate and relevant information on food calories or nutritional content remains a significant challenge. Most nutritional data retrieval processes still rely on manual references or online sources that are often incomplete, outdated, or difficult to access. These limitations hinder users' ability to accurately monitor their daily nutritional intake, ultimately affecting the effectiveness of dietary management and overall health. Previous studies have attempted to address this issue by developing various methods and applications, such as deep learning-based platforms for automatic calorie estimation (Chen & Chiang, 2025) and computer vision-based applications for nutritional composition analysis (Chung et al., 2021). However, most of these solutions still face limitations in terms of data coverage, estimation accuracy, and adaptability to food variations across different regions. In response to these challenges, this study

© The Author(s) 2025

A. A. N. G. Sapteka et al. (eds.), *Proceedings of the International Conference on Sustainable Green Tourism Applied Science - Engineering Applied Science 2025 (ICOSTAS-EAS 2025)*, Advances in Engineering Research 280,

https://doi.org/10.2991/978-94-6463-878-3_35

proposes a method that integrates Bootstrapping Language-Image Pre-training (BLIP) with Large Language Model (LLaMA) to automatically detect food calories, providing broader and more comprehensive coverage of calories and nutritional data. In this approach, BLIP generates accurate textual descriptions from food images uploaded by users. LLaMA then analyzes these descriptions to estimate key nutritional parameters, including calories, protein, fat, and carbohydrates. The method is implemented within an AWS serverless environment, utilizing services such as AWS Lambda, Amazon API Gateway, and Amazon S3 to ensure system scalability, efficiency, and flexibility.

To evaluate the performance of the proposed method, this study employs Confusion Matrix-based testing to measure the accuracy, precision, recall, and F1-score of the nutritional estimations. The synergy between BLIP and LLaMA, supported by a serverless cloud computing infrastructure, is expected to produce more precise nutritional estimations, adapt to a wide variety of food types, and expand the availability of nutritional information for end users..

2 Methodology

2.1 Application Architecture Design

The development of this food calorie detection application adopts a serverless architecture based on Amazon Web Services (AWS) to support scalability, cost efficiency, and ease of service management (Nandula & Padmanabhan, 2023; Menéndez et al., 2023). The selection of a serverless architecture enables the application to automatically handle image processing tasks without the need for manual server infrastructure management. Users will upload food images through the web application interface. The uploaded image data will be directly processed in the cloud using a pipeline built on AWS Lambda, with Amazon S3 utilized for data storage.

2.2 Implementation of the BLIP Model

The uploaded images will be processed using the Bootstrapping Language-Image Pre-training (BLIP) model. Prior studies by Li et al. (2022), Li et al. (2023), Singh et al. (2022), Wang et al. (2022), and Yu et al. (2022) have demonstrated that the application of BLIP in various domains significantly enhances accuracy in visual understanding and natural language processing across diverse contexts. BLIP is specifically designed to synergistically integrate visual and language processing, enabling it to comprehend image content and generate accurate textual descriptions. Recent studies have consistently shown that BLIP and similar vision-language models exhibit superior performance in tasks such as image captioning, visual question answering, and zero-shot image classification. These advantages position BLIP as a foundational model in the development of vision-language applications that require automated image analysis. Leveraging a pre-training approach on large-scale datasets, BLIP is capable of learning robust multimodal representations, thereby providing more stable and accurate results

in visual food recognition tasks. These strengths form the basis for selecting BLIP in the present study to detect food types from user-uploaded images.

In this study, the BLIP model introduced by Li et al. (2023) is implemented to improve computational efficiency by adopting a frozen image encoder. The direct integration of the frozen image encoder with a Large Language Model (LLM) was further advanced in the BLIP-2 framework proposed by Li et al. (2023). The BLIP model offers enhanced visual grounding capabilities and effectively links visual outputs to natural language understanding, making it particularly well-suited for vision-language tasks that require precise visual-linguistic alignment. The textual description produced by the BLIP model serves as the primary input for the subsequent stage, where it is processed by the Large Language Model within the AWS-based environment to perform calorie estimation and classification.

2.3 Integration of AWS-Based LLaMA Model

After obtaining the visual description from the BLIP model, the subsequent stage involves integrating the Large Language Model (LLaMA) to perform food calorie estimation. In this stage, LLaMA receives the textual description generated by BLIP as input and processes it to estimate caloric content. This integration is deployed within an AWS serverless environment, enabling flexible and efficient model execution without the need for manual server management. To facilitate inter-service communication and real-time data processing, the system utilizes serverless microservices built on AWS Lambda, Amazon API Gateway, and Amazon S3 (Nandula & Padmanabhan, 2023).

The adoption of AWS architecture provides high scalability, allowing the system to process multiple user requests concurrently while maintaining low latency and cost efficiency (Menéndez et al., 2023). Furthermore, the design of this serverless microservices architecture incorporates modern security principles aimed at strengthening system protection (Wang et al., 2022). Each request undergoes rigorous authentication and verification procedures to ensure data integrity and safeguard the overall service.

The effectiveness of this approach is reinforced by the findings of (Yu et al., 2022), which demonstrates that migrating complex applications to serverless platforms such as AWS can improve management efficiency and enhance operational flexibility across diverse application contexts. At the same time, the architecture acknowledges potential limitations identified by Singh et al. (2022), including considerations for auto-scaling and resource optimization in AWS Lambda deployments. By addressing these factors, the integration of LLaMA within the AWS ecosystem is expected to reduce data processing time, increase service availability, strengthen end-to-end security, and optimize cloud infrastructure expenditure. This design aligns with best practices for serverless and microservices architecture within the AWS environment, as recommended by Menéndez et al. (2023) and Nandula & Padmanabhan (2023).

2.4 Confusion Matrix

The performance evaluation of the application was conducted by comparing the calorie estimation results generated by the BLIP and LLaMA models against the actual food calorie data obtained from a validated nutritional database. The model accuracy was assessed using a confusion matrix to calculate precision, recall, and F1-score in classifying calorie levels (low, medium, high), following the evaluation approach adopted in previous studies (Cherti et al., 2023; Tanabe & Yanai, 2025a). The confusion matrix is an effective method for measuring classification performance by considering the distribution of correct and incorrect predictions. In this evaluation, the true positive (TP), false negative (FN), false positive (FP), and true negative (TN) values were calculated to assess the classification accuracy of the model comprehensively.

For comparison, previous studies have proposed various methods to improve the accuracy of food recognition and calorie estimation, such as the use of deep learning for automatic food identification on mobile devices (Tanabe & Yanai, 2025b) and volume-based calorie estimation methods as proposed by (Cherti et al., 2023; Tanabe & Yanai, 2025a). In addition, other approaches that have been applied include the use of hybrid transformer models to enhance food recognition accuracy (Chen & Chiang, 2025), lightweight and parameter-optimized models for real-time calorie estimation (Jagadesh et al., 2025), and the implementation of the RT-DETR model for fruit calorie estimation from digital images (Haque et al., 2022). Furthermore, deep learning-based food recognition benchmarks have also been developed to support more accurate nutritional assessments (Tang & Yan, 2024).

3 Result and Discussion

3.1 Result

The model performance evaluation was conducted on a total of 349 test samples, consisting of 200 food items and 149 beverage items. Based on the evaluation results presented in Figure 1, the distribution of the model's predictions is as follows: True Positive (TP) = 287 samples, False Negative (FN) = 42 samples, False Positive (FP) = 15 samples, and True Negative (TN) = 5 samples. The calculated performance metrics of the model are as follows: Accuracy: 83.67%, Precision: 95.03%, Recall: 87.23%, and F1-Score: 90.97%. These results indicate that the model achieves a high overall classification accuracy. The high precision (95.03%) suggests that most of the model's positive predictions are correct, indicating a low rate of false positive errors. This is particularly important in the context of calorie estimation for food and beverages, where positive classification errors could lead to inaccurate nutritional information being provided to users. The performance metrics of the proposed model, including accuracy, precision, recall, and F1-score, are visually summarized in Figure 1, which illustrates the distribution of classification results across true positive, false negative, false positive, and true negative categories.

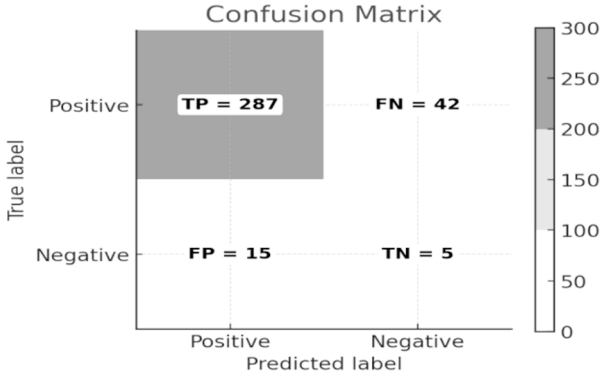


Figure 1. Confusion Matrix of Model Testing Results of Applying Bootstrapping Language-Image Pre-training for Nutrition Detection from Food Images

The recall value of 87.23% indicates that the model performs reasonably well in detecting all data that should be classified as positive, although some instances remain undetected (false negatives). The F1-score of 90.97% demonstrates a good balance between the model's precision and sensitivity. The visualization of the confusion matrix in Figure 1 shows a dominance of correct predictions in the true positive quadrant, which is illustrated with a darker blue color gradient. The distribution in the false negative and false positive quadrants indicates that, although the model has demonstrated strong performance, the potential for misclassification can still be further minimized through continued development and optimization.

The evaluation using the confusion matrix emphasizes the importance of measuring accuracy, precision, recall, and F1-score in assessing the performance of image classification models. The models evaluated in this study, namely BLIP and LLaMA, have demonstrated strong capabilities in classifying the calorie levels of food and beverages based on the testing conducted on 349 test samples.

3.2 Discussion

The application architecture developed in this study adopts a distributed approach involving two main servers: the BLIP (Bootstrapping Language-Image Pretraining) server, which is responsible for image captioning processing, and the Amazon Web Services (AWS)-based architecture, which hosts the Large Language Model (LLM), specifically LLaMA. The system workflow begins when the user uploads a food image through the web application interface. The uploaded image is then processed by the BLIP server, which performs visual analysis and automatically generates a textual description. This process aims to identify the food items present in the image and provide structured descriptive information in the form of coherent sentences. The overall workflow and integration of the BLIP and LLaMA models within the AWS serverless environment are illustrated in Figure 2, providing a detailed view of the system components and data flow.

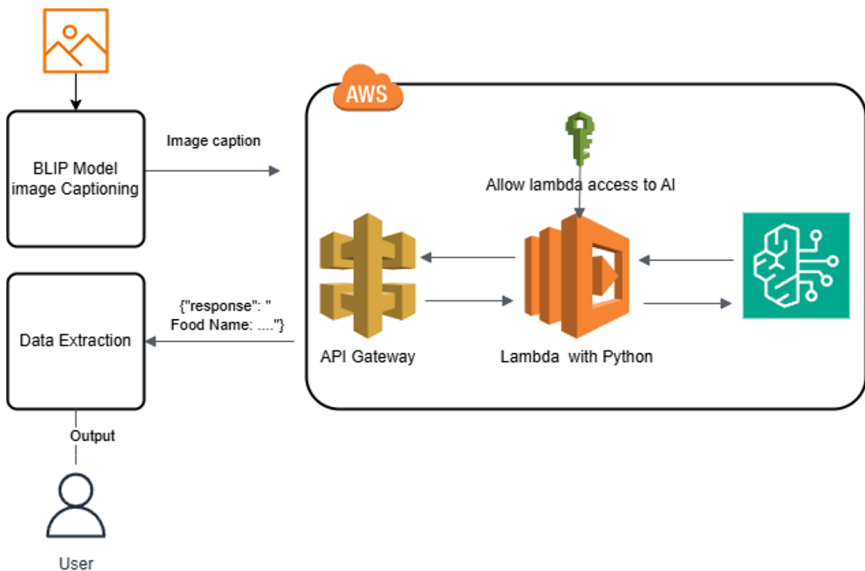


Figure 2. Application Architecture of Applying Bootstrapping Language-Image Pre-training for Nutrition Detection from Food Images

The output generated by BLIP, which contains the food image description, is transmitted to the AWS architecture for further inference processing. On the AWS side, the API Gateway serves as the communication entry point between the application and the cloud backend services. The data received by the API Gateway is processed by AWS Lambda, which is executed using Python and has access to the LLaMA model responsible for calorie estimation and calorie level classification. The LLaMA model analyzes the descriptive output from BLIP to generate more accurate predictions for calorie estimation.

The output generated at this stage is presented in a structured JSON data format, containing essential information such as the food name, description, estimated calories, and other nutritional components. The JSON data is then parsed to extract key elements, including the food name, food description, calorie count, and additional nutritional information. The final extracted results are subsequently displayed on the web interface, which users can access in real time.

The implementation of this architecture offers significant advantages in terms of scalability, processing efficiency, and seamless integration with cloud services. The use of AWS Lambda serverless services enables the system to operate dynamically and resource-efficiently, as it is only executed upon request. Furthermore, separating the image captioning process and calorie inference into two independent servers enhances the modularity and flexibility of the system. Therefore, the integration of BLIP and LLaMA within the cloud-based architecture developed in this study provides an effective and efficient solution to support automated and accurate food detection and calorie estimation based on image inputs.

4 Conclusion

Based on the evaluation conducted on 349 test samples, consisting of 200 food items and 149 beverage items, it can be concluded that the calorie detection model developed through the integration of the BLIP and LLaMA methods demonstrates strong performance in classifying calorie levels of food and beverages. The evaluation using the confusion matrix yielded an accuracy of 83.67%, precision of 95.03%, recall of 87.23%, and an F1-score of 90.97%. The high precision value indicates that the model is highly capable of generating relevant positive predictions. In contrast, the relatively high recall suggests that the model is effective in detecting most of the data that should be classified as positive. Nevertheless, several misclassifications, particularly false negatives, remain, indicating that the model's sensitivity needs to be improved to further reduce classification errors.

Overall, the model implemented in this study demonstrates significant potential for application in image-based calorie estimation systems for food and beverages, with a high degree of reliability. For future development, it is recommended to optimize the model parameters and increase the diversity of the test dataset to minimize prediction errors and enhance the model's generalizability across various types of food and beverages.

Acknowledgment

The author would like to express sincere gratitude to Politeknik Negeri Bali for providing financial support through the DIPA funding scheme. This research was conducted under the Department of Information Technology, DIII Information Management Program. The support and resources provided by the institution were essential to the successful completion of this study.

References

- Chen, Y.C., & Chiang, H.C. (2025). Deep learning-based automatic food identification with a numeric label. *Multimedia Tools and Applications*. <https://doi.org/10.1007/s11042-025-20648-x>.
- Cherti, M., Czernik, A., Kesselheim, S., Effenberger, F., & Jitsev, J. (2023). *A comparative study on generative models for high-resolution solar observation imaging* (No. arXiv:2304.07169). arXiv. <https://doi.org/10.48550/arXiv.2304.07169>.
- Chung, C.F., Ramos, A., Chiang, P.N., Wu, C.C., Tan, C. A., Khoo, W., & Crandall, D. (2021). *Computer vision for dietary assessment*.
- Haque, R. U., Khan, R. H., Shihavuddin, A. S. M., Syeed, M. M. M., & Uddin, M. F. (2022). Lightweight and parameter-optimized real-time food calorie estimation from images using a CNN-Based Approach. *Applied Sciences*, 12(19), 9733. <https://doi.org/10.3390/app12199733>.
- Jagadesh, B. N., Mantena, S. V., Sathe, A. P., Prabhakara Rao, T., Lella, K. K., Pabboju, S. S., & Vatambeti, R. (2025). Enhancing food recognition accuracy using hybrid transformer

- models and image preprocessing techniques. *Scientific Reports*, 15(1), 5591. <https://doi.org/10.1038/s41598-025-90244-4>.
- Li, D., Li, J., & Hoi, S. C. H. (2023). *BLIP-Diffusion: Pre-trained subject representation for controllable text-to-image generation and editing* (No. arXiv:2305.14720). arXiv. <https://doi.org/10.48550/arXiv.2305.14720>.
- Li, J., Li, D., Xiong, C., & Hoi, S. (2022). *BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation* (No. arXiv:2201.12086). arXiv. <https://doi.org/10.48550/arXiv.2201.12086>.
- Menéndez, J. M., Gayo, J. E. L., Canal, E. R., & Fernández, A. E. (2023). *A comparison between traditional and serverless technologies in a microservices setting* (No. arXiv:2305.13933). arXiv. <https://doi.org/10.48550/arXiv.2305.13933>.
- Nandula, L., & Padmanabhan, S. C. (2023). Serverless microservices architecture on AWS. *International Journal of Scientific and Research Publications*, 13(10), 28–37. <https://doi.org/10.29322/IJSRP.13.10.2023.p14205>.
- Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., & Kiela, D. (2022). *FLAVA: A foundational language and vision alignment model* (No. arXiv:2112.04482). arXiv. <https://doi.org/10.48550/arXiv.2112.04482>.
- Tanabe, H., & Yanai, K. (2025a). CalorieVoL: Integrating volumetric context into multimodal large language models for image-based calorie estimation. *MultiMedia Modeling*, 15523, 353–365. Springer Nature Singapore. https://doi.org/10.1007/978-981-96-2071-5_26.
- Tanabe, H., & Yanai, K. (2025b). Reasoning-driven food energy estimation via multimodal large language models. *Nutrients*, 17(7), 1128. <https://doi.org/10.3390/nu17071128>.
- Tang, S., & Yan, W. (2024). Utilizing the RT-DETR model for fruit calorie estimation from digital images. *Information*, 15(8), 469. <https://doi.org/10.3390/info15080469>.
- Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., & Wei, F. (2022). *Image as a foreign language: BEiT pretraining for all vision and vision-language tasks* (No. arXiv:2208.10442). arXiv. <https://doi.org/10.48550/arXiv.2208.10442>.
- Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., & Wu, Y. (2022). *CoCa: Contrastive captioners are image-text foundation models* (No. arXiv:2205.01917). arXiv. <https://doi.org/10.48550/arXiv.2205.01917>.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

