

# CalorieLLaVA: Image-based Calorie Estimation with Multimodal Large Language Models

Hikaru Tanabe<sup>[0009-0003-5752-9175]</sup> and Keiji Yanai<sup>[0000-0002-0431-183X]</sup>

The University of Electro-Communications, Chofu, Tokyo, Japan  
{tanabe-h, yanai}@mm.inf.uec.ac.jp

**Abstract.** Multimodal large language models (MLLMs) have demonstrated the ability to perform complex reasoning tasks by utilizing their language-based knowledge, including reasoning about food. Given this capability, we hypothesize that MLLMs can enhance calorie estimation from food images by incorporating this language-based reasoning, which is lacking in existing calorie estimation models. However, the effectiveness of these models, particularly when generating text-based outputs for calorie estimation, has not been fully explored. In this work, we introduce CalorieLLaVA, a model fine-tuned on paired food images and calorie data based on an MLLM. By fine-tuning the LLaVA model on the Nutrition5k dataset, we evaluate its performance in calorie estimation. Our experiments demonstrate that CalorieLLaVA surpasses the baseline models, including GPT-4V and FoodLMM, achieving superior results on the Nutrition5k dataset.

**Keywords:** Multimodal Large Language Models · Image-based Calorie Estimation · Nutrition5k

## 1 Introduction

Tracking daily food intake is essential for achieving health-related goals such as dieting and bodybuilding. Accurately estimating the caloric content of meals is a key factor in maintaining a healthy lifestyle. Traditional methods for calorie estimation, such as food diaries, 24-hour recalls, and food frequency questionnaires, are time-consuming and depend heavily on participants' ability to recall or measure their food accurately. This can be difficult, leading to errors and inconsistencies in calorie estimation.

With the advent of Multimodal Large Language Models (MLLMs), there is potential to overcome these challenges. MLLMs possess strong reasoning capabilities and can recognize diverse types of food from images, enabling high-quality calorie estimation. However, the challenge remains in accurately estimating the volume of food, which is crucial for precise calorie computation. Existing MLLMs have yet to fully address this issue.

In this study, we introduce CalorieLLaVA, a novel model designed for calorie estimation from food images using MLLMs. By fine-tuning an MLLM on food image-calorie pairs, we aim to explore the effectiveness of MLLMs for this

task. Our approach not only offers a strong baseline for calorie estimation using MLLMs but also improves upon previous methods.

The main contributions of this study are as follows:

- We propose a new model CalorieLLaVA that employs LLaVA as an MLLM, and fine-tune it on Nutrition5k dataset to enable accurate calorie estimation from food images.
- We validate the performance of CalorieLLaVA on Nutrition5k and demonstrate its superiority over existing methods, achieving higher scores than the original MLLMs and previous calorie estimation approaches.
- We provide a new baseline in MLLM-based calorie estimation, surpassing recent models like GPT-4V and domain-specific models like FoodLMM, and discuss the efficacy of adjusting the training setting.

## 2 Related Work

### 2.1 Size-Based Methods of Calorie Estimation

There are two primary methods for estimating caloric content from food images: size-based methods and direct estimation methods.

In size-based methods, a pipeline is constructed that combines multiple image recognition modules to estimate caloric content. The basic procedure involves first extracting the food regions from the food image, then estimating the food category, and subsequently estimating the volume or mass of the food regions. Based on these results, the caloric content is estimated. By going through several stages prior to calorie estimation, this approach enables estimation that specifically takes the quantity of food into consideration.

As a method for obtaining the actual area of the food in the image, there are methods based on estimating the actual size of objects included in the food image. To determine the actual size, there are approaches such as the one by Okamoto *et al.* [15] that refers to a credit card or long wallet, the one by Akpa *et al.* [8] that refers to chopsticks, and the one by Ege *et al.* [6] that refers to rice grains. Additionally, there is a method by Tanno *et al.* [20], which uses anchors set in an AR space to obtain the actual size through interaction with the user.

As a method for obtaining the volume of food, there is DepthCalorieCam by Ando *et al.* [2]. This method significantly reduced the error in calorie estimation compared to existing methods by estimating the volume of food using a depth camera and a region segmentation model. Furthermore, Naritomi *et al.* [14] reconstructed high-quality 3D meshes of plates and food using an implicit surface representation.

However, these size-based methods have the limitation on the types of food that can be estimated. In particular, a noted issue with DepthCalorieCam is that the food types subject to estimation are limited to only three categories.

## 2.2 Direct Methods of Calorie Estimation

Direct estimation methods involve deep learning models that have achieved high performance in general image recognition are selected, and transfer learning related to caloric estimation is applied to achieve high-quality predictions.

Ege *et al.* [7] applied a multi-task learning approach based on VGG16 [19], to estimate not only the caloric content but also food category, ingredients, and cooking procedures. This method achieved higher quality estimation compared to simple transfer learning.

However, a challenge with the direct estimation methods is the difficulty in considering the quantity of food when estimating the caloric content. As a result, even with the same type of food, accurate estimation becomes difficult if the quantity differs. Furthermore, training calorie estimation models requires a large amount of training data, and the burden of data annotation remains a challenge.

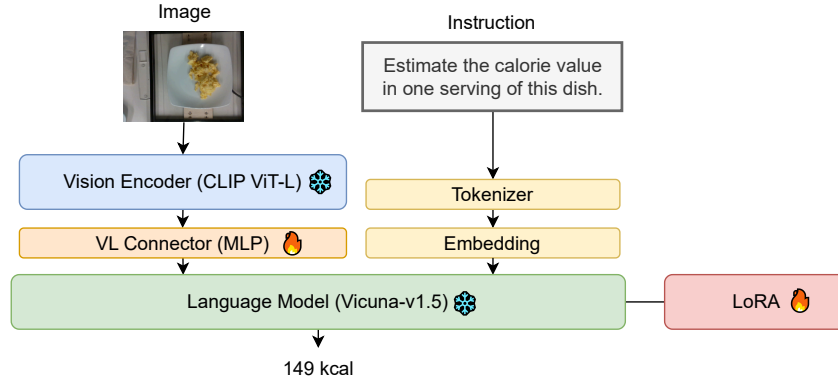
In this study, we aim to address the limitations of both approaches by introducing MLLMs, which have recently seen significant improvements in performance. By leveraging their reasoning capabilities based on prior knowledge of food, we aim to achieve accurate calorie estimation. Additionally, we apply fine-tuning specific to caloric estimation to further enhance the model’s performance in this task.

## 2.3 Large Language Models (LLMs)

In recent years, the field of natural language processing has seen the emergence of Large Language Models (LLMs), which are trained under conditions with large model parameters, extensive data, and substantial computational resources. These models have achieved high performance across various tasks. It has been reported that the performance of these models improves according to a power law with the scale of training conditions [10], and they exhibit emergent capabilities, where performance significantly improves at a certain scale [22], highlighting new aspects not observed in conventional language models.

## 2.4 Multimodal Large Language Models (MLLMs)

When LLMs are extended to multiple modalities such as vision, they are referred to as Multimodal Large Language Models (MLLMs). To extend these models to visual modalities, visual encoders that convert visual information, such as images or videos, into visual features are used. Flamingo [1] created a single model capable of handling various tasks related to images and videos by integrating visual features obtained from a visual encoder with textual features through a gated cross-attention layer. LLaVA [13] adopted a model structure that transforms visual features using a connector layer built with linear layers or MLPs, which are then input into a LLM. It also employed a training framework called Visual Instruction Tuning, which used instruction-following data including visual information, to achieve high-quality task generalization performance related



**Fig. 1.** Model architecture of CalorieLLaVA (based on LLaVA-v1.5)

to images. Other models include BLIP-2 [11], which introduced Q-Former in the connector layer, and MiniGPT-4 [24] and InstructBLIP [4], which applied similar learning as Visual Instruction Tuning based on the same model structure.

In the domain of food, FoodLMM [23] has achieved SOTA performance across various food-related tasks, including caloric content estimation from food images. This study focuses specifically on improving the performance of caloric content estimation from food images.

### 3 Methods

In this study, we propose a framework called CalorieLLaVA. We first explain the method for estimating caloric content using MLLMs (Section 3.1), and then describe how to fine-tune them (Section 3.2).

#### 3.1 Calorie Estimation using MLLMs

We perform calorie estimation using MLLMs like LLaVA-v1.5 [12] (Fig. 1), and perform fine-tuning based on this model. The input image is encoded into visual features via the OpenAI CLIP ViT-L visual encoder [16]. These visual features are then transformed to match the dimension of text token embeddings through a two-layer MLP-based vision-language connector. Both the visual features and the text token embeddings are then fed into the LLM Vicuna-v1.5 [3], which produces the estimated calorie value as output. The content of the text instructions input to the LLM is designed to prompt the estimation of the calorie value from the food images.

### 3.2 Fine-tuning MLLMs on Image-Caloric data

In this study, we use the Nutrition5k [21] dataset for training and evaluating the model. This dataset, created by Google Research, is focused on nutritional understanding of food images and includes 3,265 top-down food images along with nutritional information, including caloric content. The dataset is split into 2,759 images for training and 506 images for testing, and we use the training split for this fine-tuning. To fine-tune the MLLM, we convert the training split of the dataset into an instruction-following format. The instruction-following format represents the instructions the model is supposed to follow and the expected responses. For this conversion, we prepare a text template that transforms the caloric content into an instruction-response format. We uniformly set the instruction as follows:

`Estimate the calorie value in one serving of this dish.`

The response is structured by enclosing the calorie value in `[[[]]]` followed by the word `calories`. By prompting the model to output calorie values in a single consistent format, we aim to make it easier to extract caloric content from text with regular expression.

The fine-tuning process involves training the vision-language connector and LLM using pairs of food images and calorie values converted into an instruction-response format. The LLM is fine-tuned using LoRA [9] for efficient adaptation with reduced computational overhead.

## 4 Experiments

### 4.1 Experiment Setting

We trained the LLaVA-7B and 13B models using the training set of Nutrition5k and referred to the trained models as CalorieLLaVA-7B and 13B. The 6th and 5th epoch checkpoints were selected for CalorieLLaVA-7B and 13B respectively, based on monitoring the loss on the validation split using the model trained on the validation training split (Fig. 2 and 3).

During training, the AdamW optimization algorithm was used with a linear warmup and cosine decay of the learning rate, with a peak value of  $2 \times 10^{-4}$ . The rank for the LoRA matrix decomposition was set to 128, and the batch size was set to 64. The temperature parameter for text generation was set to 0, ensuring deterministic output of tokens. Additionally, data parallelism using the DeepSpeed [18] and memory optimization via ZeRO-3 [17] were implemented during training. The training was conducted on  $4 \times$  RTX 3090 for up to 10 epochs, taking approximately 5 hours.

For models without fine-tuning including LLaVA and GPT-4V, we evaluated only the entries where the estimation was successfully completed, and the calorie value was correctly extracted. For cases where the calorie value could not be extracted with the temperature parameter set to 0, we adjusted the temperature

parameter to 0.2 and repeated the same query up to five times to extract the calorie value. The entries that could not be extracted after these attempts were excluded from the evaluation. Specifically, 16, 4, and 28 entries were excluded from the evaluation datasets for LLaVA-7B, LLaVA-13B, and GPT-4V, respectively. Additionally, if the calorie estimation was output as a range, we used the average of the range as the estimated value.

## 4.2 Results of Calorie Estimation

Table 1 shows the results of calorie estimation on the test split of Nutrition5k. The proposed model CalorieLLaVA achieved higher scores in mean absolute error (MAE) compared to the baseline Google-nutrition-monocular [21] and other MLLMs. It also achieved a better MAE score compared to the fine-tuned FoodLMM model. Furthermore, when comparing fine-tuned models, we observe that CalorieLLaVA-13B, with its larger number of parameters, outperforms the 7B model across all metrics.

According to the research on Nutrition5k [21], the mean absolute percentage error (MAPE) of nutritionists’ estimations on Nutrition5k dataset is reported to be 41 %. In comparison, CalorieLLaVA-13B achieves a better score. This indicates that the estimation capability of CalorieLLaVA-13B surpasses that of human nutritionists. However, it should be noted that this result is based on evaluations conducted on a subset of the test split, and the test split used in this study does not perfectly match the one used in the Nutrition5k research.

## 4.3 Distribution of the Results

Fig. 4 and Fig. 5 show the distribution of estimated calorie values by LLaVA-13B and CalorieLLaVA-13B. LLaVA-13B tends to make many inaccurate estimations, especially around 200 kcal. On the other hand, CalorieLLaVA-13B shows a significant reduction in incorrect estimations around that range, with more estimates concentrated in the intervals where the ground truth data is densely located. Fig. 6 and Fig. 7 show that the correlation between estimated and actual values improved due to fine-tuning.

In these figures, the set of points where the true values and estimated values match is represented by red lines, while the 95 % confidence ellipses are shown in black. When comparing both the red lines and the confidence ellipses, we observe that the CalorieLLaVA-13B’s predictions exhibit less variance. This is also supported by the higher correlation coefficient values for the CalorieLLaVA-13B. Additionally, in the scatter plot of LLaVA-13B, there are instances where the model estimates around 700 kcal and 180 kcal for true values of approximately 50 kcal and 900 kcal, respectively. In contrast, the scatter plot of the CalorieLLaVA-13B shows that such significant deviations, particularly in the lower true value ranges, are not present.

From these observations, we can conclude that the fine-tuned model provides more accurate calorie estimations.

**Table 1.** Results of calorie estimation

| Method                          | MAE / kcal ↓ | MAPE / % ↓  | r ↑          |
|---------------------------------|--------------|-------------|--------------|
| Google-nutrition-monocular [21] | 70.6         | 26.1        | -            |
| LLaVA-7B                        | 178.8        | 129.5       | 0.637        |
| LLaVA-13B                       | 177.1        | 92.8        | 0.656        |
| GPT-4V                          | 106.6        | 54.8        | 0.688        |
| FoodLMM FT [23]                 | 67.3         | <b>26.6</b> | -            |
| CalorieLLaVA-7B (Ours)          | 74.2         | 41.5        | 0.927        |
| CalorieLLaVA-13B (Ours)         | <b>64.3</b>  | 39.8        | <b>0.934</b> |

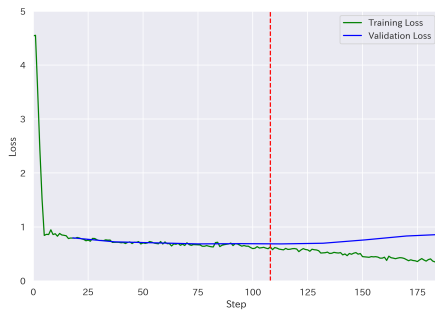
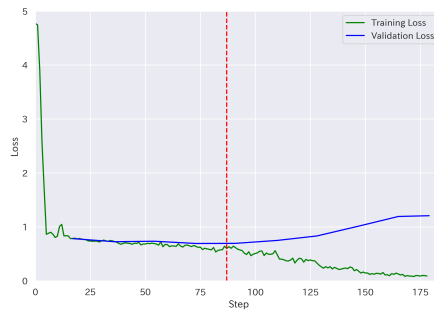
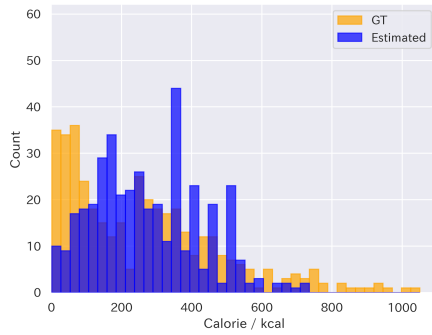
**Fig. 2.** Loss curves with respect to the number of training steps in CalorieLLaVA-7B training. Red line: selected checkpoint.**Fig. 3.** Loss curves with respect to the number of training steps in CalorieLLaVA-13B training. Red line: selected checkpoint.

Fig. 8 to Fig. 10 represent the changes in the values of evaluation metrics on the test split as the number of training epochs increases. For MAE and MAPE, there is a tendency for the values to fluctuate and not decrease significantly even as the number of epochs increases. On the other hand, the correlation coefficient shows a tendency to decrease monotonically with the number of epochs up to about 5 epochs, which is particularly noticeable in CalorieLLaVA-13B. In contrast, the values of each evaluation metric for CalorieLLaVA-7B appear relatively unstable, with a temporary decline observed around the 5th epoch. Additionally, it is noted that CalorieLLaVA-7B achieves better scores in the first epoch.

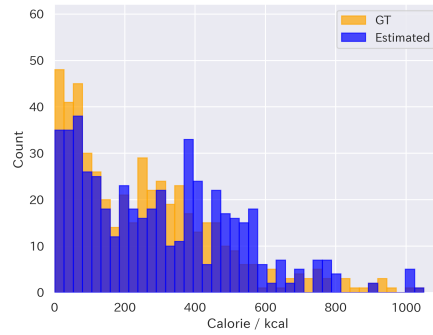
Fig. 11 and Fig. 12 are examples of responses for calorie estimation by each model. For LLaVA-13B and GPT-4V without fine-tuning, the output includes both the process of calorie estimation and the estimated value. In contrast, for CalorieLLaVA-13B, the estimated value is directly outputted.

## 5 Discussion

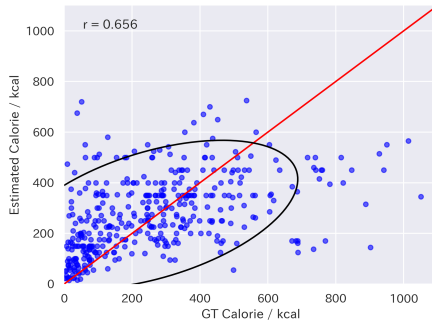
In Section 4, we confirmed that increasing the number of training epochs during fine-tuning of MLLMs tends to improve or maintain the correlation coefficient



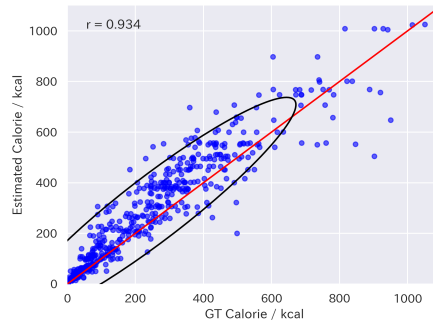
**Fig. 4.** Distribution of estimated calorie values by LLaVA-13B



**Fig. 5.** Distribution of estimated calorie values by CalorieLLaVA-13B



**Fig. 6.** Scatter plot of estimated calorie values by LLaVA-13B



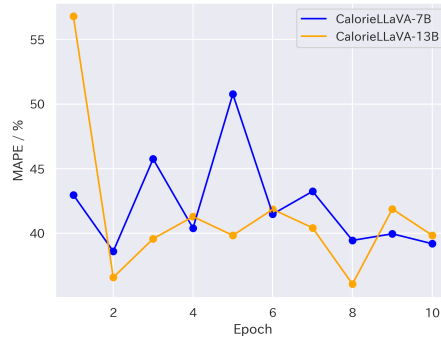
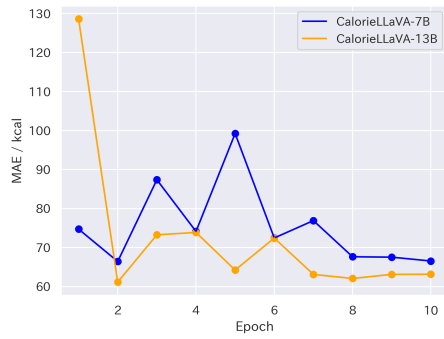
**Fig. 7.** Scatter plot of estimated calorie values by CalorieLLaVA-13B

scores. Additionally, the number of training samples in the Nutrition5k training split used in this study is around 2,700, which is relatively small compared to the number of data points used for Visual Instruction Tuning in the base model, LLaVA. Based on this, it can be said that when applying fine-tuning on a relatively small dataset for domain adaptation of MLLMs, increasing the number of training epochs is effective in improving the correlation between ground truth and estimated calories.

On the other hand, an excessive increase in the number of epochs was found to cause overfitting to the training split. In Fig. 2 and Fig. 3, the gap between the training and validation losses as the number of training steps increased was observed. Therefore, when increasing the number of epochs, it is important to monitor the validation loss during training and select appropriate checkpoints using methods like early stopping to maintain the generalization performance of the model.

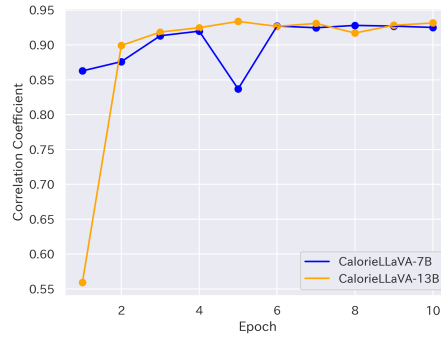
This trend is also confirmed by referencing other studies that applied fine-tuning to LLMs. Dinh *et al.* [5] applied LIFT (Language-Interfaced Fine-Tuning),





**Fig. 8.** Changes in MAE of calorie estimation results with the number of training epochs

**Fig. 9.** Changes in MAPE of calorie estimation results with the number of training epochs



**Fig. 10.** Changes in correlation coefficient of calorie estimation results with the number of training epochs



Answer: **164 kcal**

User

Estimate the calorie value in one serving of this dish.

LLaVA-13B

To estimate the calorie value of this dish, we need to consider the ingredients and their respective calorie counts. The dish consists of a salad, which typically contains a mix of vegetables, and a fruit salad with melon, strawberries, and blueberries.

1. Salad: The salad is made of lettuce, which is low in calories. Assuming the salad also contains other vegetables, we can estimate the calorie count to be around 100-200 calories.
2. Fruit salad: Melon is low in calories, while strawberries and blueberries are rich in nutrients and antioxidants. Assuming a serving of fruit salad contains 1/2 cup of each fruit, we can estimate the calorie count to be around 100-150 calories.

Combining the salad and fruit salad, we can estimate the calorie value of one serving to be around **200-350 calories**. However, this is a rough estimate, and the actual calorie count may vary depending on the specific ingredients and portion sizes used in the dish.

GPT-4V



The image shows a plate with a variety of foods that include a leafy green salad, some fruit pieces (looks like melon and possibly strawberries), blueberries, and roasted potatoes. Estimating the calorie content of this meal can be challenging without knowing the exact quantities and any dressings or oils used, but I can provide a rough estimate based on typical serving sizes.

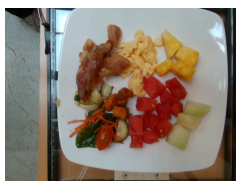
- Leafy greens: A cup of mixed greens is typically around **[[10]]** calories.
- Melon: A cup of diced melon is approximately **[[60]]** calories.
- Strawberries: A cup of strawberries is about **[[50]]** calories.
- Blueberries: A cup of blueberries comes to around **[[85]]** calories.
- Roasted potatoes: A half-cup of roasted potatoes can be roughly **[[140]]** calories, depending on the preparation.

Assuming this is a single serving and the fruits and potatoes are in half-cup portions, the total estimated calorie content for this plate might be around **[[345]] calories**. However, this is a rough estimate and the actual calorie content could be higher or lower.

CalorieLLaVA-13B  
(Ours)

**[[180]] calories.**

**Fig. 11.** Examples of responses by the MLLMs



Answer: 339 kcal

User

Estimate the calorie value in one serving of this dish.

LLaVA-13B

It is difficult to provide an accurate calorie estimate for this dish without knowing the specific ingredients and their quantities. However, we can make a rough estimate based on the types of food present in the image. The dish consists of a variety of foods, including fruits, vegetables, and possibly some meat.

Assuming that the dish contains a mix of fruits, vegetables, and meat, we can estimate the calorie value based on the average calorie content of each food group. For example, a serving of mixed vegetables might contain around 50-100 calories, a serving of mixed fruits might contain around 50-100 calories, and a serving of meat might contain around 100-200 calories.

Taking these estimates into account, a single serving of this dish could contain anywhere from **150 to 400 calories**. However, this is a rough estimate and the actual calorie value may vary depending on the specific ingredients and their quantities.

GPT-4V



The image shows a plate with a variety of foods that include what appears to be some bacon, scrambled eggs, pineapple, watermelon, honeydew melon, and a small salad with mixed vegetables such as carrots, tomatoes, and possibly some leafy greens.

To estimate the calorie content, I'll provide a rough estimate for each component:

- Bacon (2 strips): Approximately 80-100 calories
- Scrambled eggs (about 1 egg's worth): Approximately 90-100 calories
- Pineapple (1 small slice): Approximately 40 calories
- Watermelon (1 small slice): Approximately 30 calories
- Honeydew melon (1 small slice): Approximately 20 calories
- Mixed vegetables salad (small portion): Approximately 20-30 calories

Adding these up gives a total of approximately 280-320 calories. However, this is a rough estimate and actual calorie content can vary based on the exact serving sizes and preparation methods.

**[[300]] calories**

CalorieLLaVA-13B  
(Ours)

**[[368]] calories.**

Fig. 12. Examples of responses by the MLLMs

a method for solving classification and regression tasks without changing the text interface, to LLMs and conducted analyses on their estimation performance and inductive bias. In this experiment, regression tasks on basic functions were evaluated. As a result, the relative absolute error (RAE) used as an evaluation metric improved or was maintained as the number of epochs increased. However, it was confirmed that from around 10 epochs, the validation error during training began to diverge from the training loss.

In this study, we referred to these results on training LLMs and selected checkpoints that did not exhibit overfitting to the training split during fine-tuning of MLLMs. As a future direction, it is promising to achieve higher performance by conducting a more extensive hyperparameter search, not limited to the number of epochs. Furthermore, verifying the impact of hyperparameter changes, including the number of epochs, on the performance of MLLMs in general domain adaptation is expected to accelerate the exploration of further methods.

## 6 Conclusion

In this study, we verified the effectiveness of the approach using MLLMs for the task of estimating caloric content from food images. We created a model CalorieLLaVA by fine-tuning LLaVA, achieving performance that surpassed the baseline and contemporary MLLMs as evaluated on the Nutrition5k dataset.

Based on comparisons with recent methods, it is considered promising to introduce a monocular depth estimation model and apply multitask learning to models estimating caloric content based on special tokens and regression heads, like FoodLMM. Moreover, the Nutrition5k dataset used in this study includes only a limited range of food images within the food domain. The effectiveness of the method has not been fully verified for food domains representative of everyday meals, such as sushi or ramen, which are popular in Japan but not included in the dataset. Verification in these domains is crucial for the practical application of the method.

## References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: *Advances in Neural Information Processing Systems*. vol. 35, pp. 23716–23736 (2022)
2. Ando, Y., Ege, T., Cho, J., Yanai, K.: DepthCalorieCam: A mobile application for volume-based foodcalorie estimation using depth cameras. In: *Proc. of the 5th International Workshop on Multimedia Assisted Dietary Management*. p. 76–81 (2019)

3. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing GPT-4 with 90%\* ChatGPT quality (2023), <https://lmsys.org/blog/2023-03-30-vicuna/>, <https://lmsys.org/blog/2023-03-30-vicuna/>
4. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: InstructBLIP: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500 (2023)
5. Dinh, T., Zeng, Y., Zhang, R., Lin, Z., Gira, M., Rajput, S., Sohn, J.y., Papailiopoulos, D., Lee, K.: LIFT: Language-interfaced fine-tuning for non-language machine learning tasks. In: Advances in Neural Information Processing Systems. vol. 35, pp. 11763–11784 (2022)
6. Ege, T., Shimoda, W., Yanai, K.: A new large-scale food image segmentation dataset and its application to food calorie estimation based on grains of rice. In: Proc. of ACM MM Workshop on Multimedia Assisted Dietary Management (2019)
7. Ege, T., Yanai, K.: Image-based food calorie estimation using knowledge on food categories, ingredients and cooking directions. In: Proc. of the on Thematic Workshops of ACM Multimedia 2017. pp. 367–375 (2017)
8. Elder Akpro Hippocrate Akpa, Hirohiko Suwa, Y.A., Yasumoto, K.: Smartphone-based food weight and calorie estimation method for effective food journaling. SICE Journal of Control, Measurement, and System Integration **10**(5), 360–369 (2017)
9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: Proc. of International Conference on Learning Representations (2022)
10. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)
11. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proc. of International Conference on Machine Learning (2023)
12. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
13. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems (2023)
14. Naritomi, S., Yanai, K.: Hungry Networks: 3d mesh reconstruction of a dish and a plate from a single dish image for estimating food volume. In: Proc. of the 2nd ACM International Conference on Multimedia in Asia (2021)
15. Okamoto, K., Yanai, K.: An automatic calorie estimation system of food images on a smartphone. Proc. of the 2nd International Workshop on Multimedia Assisted Dietary Management (2016)
16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proc. of International Conference on Machine Learning. pp. 8748–8763 (2021)
17. Rajbhandari, S., Rasley, J., Ruwase, O., He, Y.: ZeRO: Memory optimizations toward training trillion parameter models. In: SC20: International Conference for High Performance Computing, Networking, Storage and Analysis. pp. 1–16 (2020)
18. Rasley, J., Rajbhandari, S., Ruwase, O., He, Y.: DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In: Proc. of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. p. 3505–3506 (2020)

19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
20. Tanno, R., Ege, T., Yanai, K.: AR DeepCalorieCam V2: Food calorie estimation with cnn and ar-based actual size estimation. In: Proc. of the 24th ACM Symposium on Virtual Reality Software and Technology (2018)
21. Thames, Q., Karpur, A., Norris, W., Xia, F., Panait, L., Weyand, T., Sim, J.: Nutrition5k: Towards automatic nutritional understanding of generic food. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 8903–8911 (2021)
22. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., Fedus, W.: Emergent abilities of large language models. arXiv preprint arXiv:2206.07682 (2022)
23. Yin, Y., Qi, H., Zhu, B., Chen, J., Jiang, Y.G., Ngo, C.W.: FoodLMM: A versatile food assistant using large multi-modal model. arXiv preprint arXiv:2312.14991 (2023)
24. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)