大規模視覚言語モデルと食品体積推定による食事画像からのカロリー量推定

田邊 光[1,a)]　柳井 啓司[1,b)]

概要

本研究では，多様な知識に基づく推論が可能な大規模視覚言語モデルを食事画像からのカロリー量推定に活用する二つの方法の有効性を検証する．一つ目は大規模視覚言語モデルをファインチューニングする方法であり，Nutrition5k における評価で既存手法に匹敵する結果となった．二つ目は食品体積推定モデルを導入する方法であり，ゼロショットカロリー量推定について複数の指標で優れた結果となった．

1. はじめに

食品に含まれるカロリー量を把握することで，ダイエットのための食事管理や健康のための食生活分析などにその情報を活用することができる．このため，食事画像から食品のカロリー量を推定する研究が取り組まれているが，推定対象となる食品の種類の制限やアノテーションの負担が双方の課題として存在する．一方で，近年の画像認識分野においては，高い推論性能をもつ大規模言語モデルを視覚のモダリティに拡張した大規模視覚言語モデルが多様な知識に基づいた視覚的推論を実現している．

そこで，本研究では食事画像からのカロリー量推定タスクに対して大規模視覚言語モデルを用いる 2 つのアプローチの有効性を検証する．一つ目は大規模視覚言語モデルをファインチューニングする方法であり，正確なカロリー量推定を実現する．二つ目は食品体積推定モデルを構築して大規模視覚言語モデルと組み合わせる方法であり，上述の課題を改善するゼロショットカロリー量推定を実現する．

2. 関連研究

食事画像からのカロリー量推定には，サイズベースの方法と直接推定による方法がある．安蒜ら [1] は，深度カメラと領域分割モデルにより実施される食品体積推定の過程を経ることで，高品質なサイズベースカロリー量推定を実現した．會下ら [2] は，VGG16 に対して食品のカロリー量・カテゴリ・食材・料理手順に関するマルチタスク学習を適用することで，高品質な直接カロリー量推定を実現した．しかし，双方には推定対象となる食品の種類の制限やアノテーションの負担が課題として存在する．本研究では食品体積推定モデルと大規模視覚言語モデルによるゼロショットカロリー量推定によりこれらの課題に対応する．

食事ドメインにおける大規模視覚言語モデルとしては FoodLMM [3] があり，食事画像からのカロリー量推定を含めた多様な食事タスクにおいて SOTA を達成した．本研究では特に食事画像からのカロリー量推定に焦点を当てて推定性能の向上を図る．

3. 手法

3.1 大規模視覚言語モデルのファインチューニングによるカロリー量推定

第一の手法では，大規模視覚言語モデルである LLaVA-v1.5 [4] をファインチューニングすることでカロリー量推定を行う（**図 1**）．まず，OpenAI CLIP-ViT-L [5] からなる視覚エンコーダによって入力画像を視覚特徴に変換する．次に，2 層 MLP からなる視覚言語接続層によって視覚特徴をトークンの埋め込みの次元に変換する．そして，その視覚特徴とテキストトークンの埋め込みを大規模言語モデルである Vicuna-v1.5 [6] に入力することで，カロリー量の推定値が出力される．ファインチューニングでは，食事画像とカロリー量のペアを指示応答形式に変換したものを用いて，接続層と大規模言語モデルを学習する．大規模言語モデルの学習にはランク 128 の LoRA [7] を適用する．

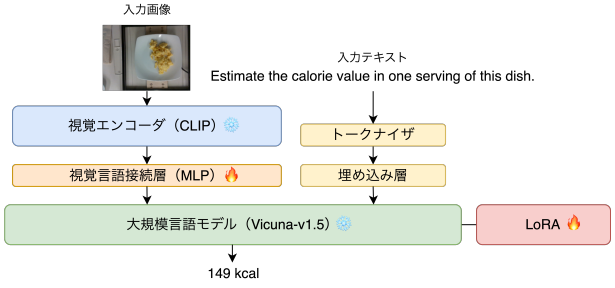

図 1　ファインチューニングによる手法のモデル構造

[1] 電気通信大学
[a)] tanabe-h@mm.inf.uec.ac.jp
[b)] yanai@cs.uec.ac.jp

3.2 食品体積推定による方法

第二の手法では，食品体積推定モデルと大規模視覚言語モデルを用いてカロリー量推定を行う（**図 2**）．体積推定の過程は次のとおりである（**図 3**）．まず，オープンセット物体検出モデルである Grounding-DINO [8] により，皿の矩形領域を得る．次に，これを関心領域として 3 つに処理が分かれる．第一に，Segment Anything（SAM） [9] を関心領域に適用し，皿の領域マスクを得る．第二に，Grounding-DINO を関心領域に適用して食品部分の矩形領域を取得し，さらに SAM を適用して領域マスクを得る．第三に，単眼深度推定モデルである Marigold [10] を関心領域に適用し，深度マップを得る．そして，これらの結果と画像の実寸に基づいて食品の体積を得る．最後に，食事画像の視覚特徴と入力テキストのトークン埋め込みとともに，体積値のトークン埋め込みを大規模視覚言語モデルに入力し，カロリー量の推定結果を得る．

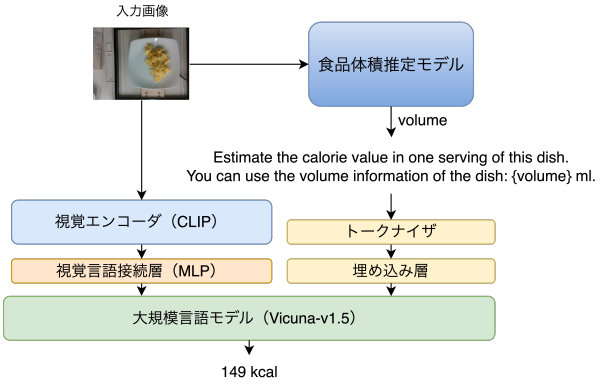

図 2 食品体積推定モデルを導入する手法の全体図

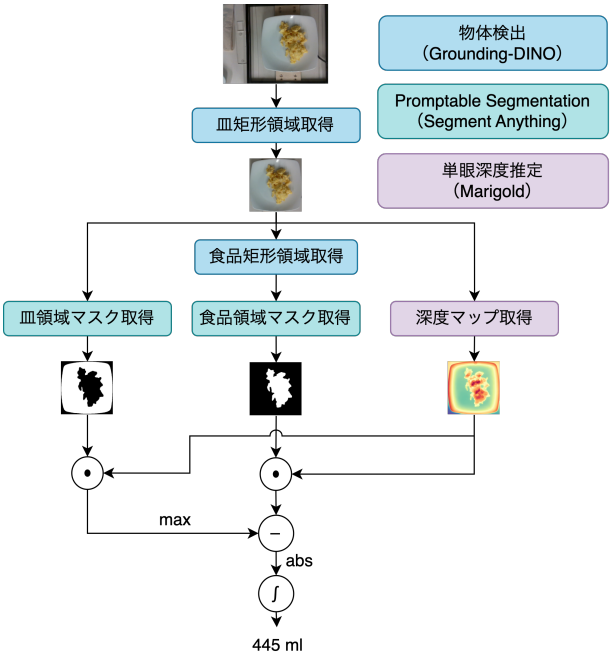

図 3 食品体積推定モデルの構造

4. 実験

4.1 ファインチューニングによる方法

本実験では，Nutrition5k [11] の overhead データセットを学習と評価に用いた．この訓練用分割に対してテンプレートによる質問応答形式への変換を行った後に，LLaVA-7B, 13B に対して訓練を行った．ただし，テンプレートの指示は Estimate the calorie value in one serving of this dish. で共通とした．また，応答はカロリー量の数値を [[]] で囲み，その背後に calories とつけたものとした．そして，検証損失の監視に基づいてそれぞれ 6, 5 エポック目のチェックポイントを用いた．**表 1** は Nutrition5k のテスト用分割に対するカロリー量推定の結果である．提案手法は，ベースラインとして設定した Google-nutrition-monocular（GNM） [11] や各大規模視覚言語モデルに対して，平均絶対誤差（MAE）と平均絶対パーセント誤差（MAPE）について高いスコアを達成した．また，FoodLMM のファインチューニング後のモデルに比べて MAE について高いスコアを達成した．さらに，**図 4**, **図 5** より，ファインチューニングによって推定値と正解値の相関が向上したことがわかる．

表 1 カロリー量推定の結果

項目	MAE / kcal ↓	MAPE / % ↓
GNM [11]	70.6	26.1
LLaVA-7B	178.8	129.5
LLaVA-13B	177.1	92.8
GPT-4V	106.6	54.8
FoodLMM FT [3]	67.3	**26.6**
LLaVA-7B FT (Ours)	74.2	41.5
LLaVA-13B FT (Ours)	**64.3**	39.8

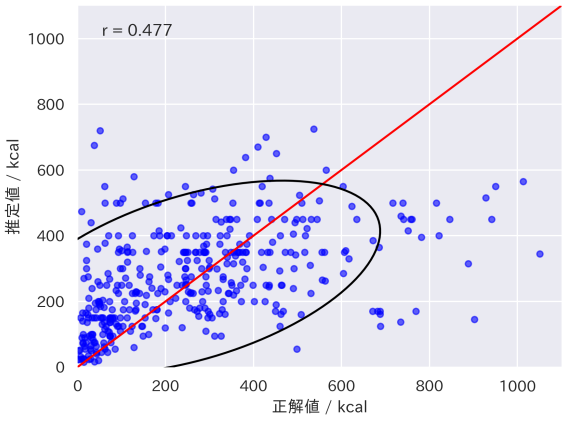

図 4 ファインチューニング前の LLaVA-13B の推定カロリー量の散布図

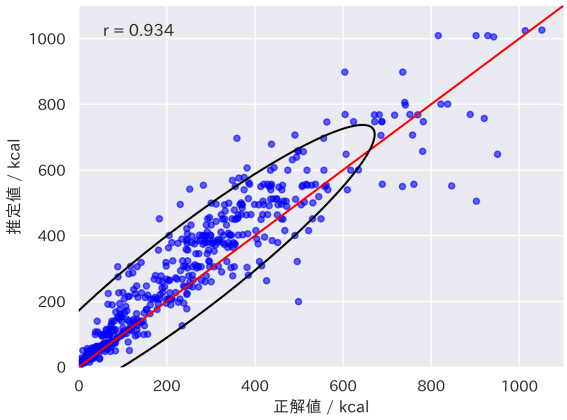

図 5 ファインチューニング後の LLaVA-13B の推定カロリー量の散布図

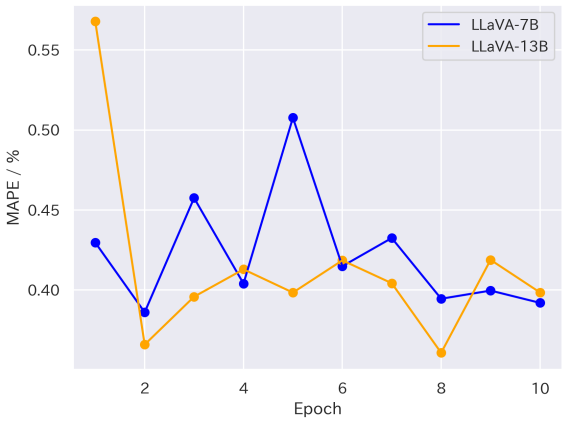

図 7 学習エポック数に対するカロリー量推定結果の平均絶対パーセント誤差（MAPE）の変化

図 6 から**図 8** は，学習エポック数の増加に伴うテスト用分割に対する各評価指標の値の変化が表された図である．MAE と MAPE については，エポック数が上がっても振動したように変化しながらあまり減少しない傾向がわかる．一方で，相関係数については，4 エポック程度まではエポック数の増加に伴って値が単調に増加しているような傾向があり，特に LLaVA-13B の方でそれが顕著である．これに対して，LLaVA-7B の各評価指標の値は，5 エポック目で一旦悪化しているように比較的安定していない様子がわかる．また，特に 1 エポックにおいては LLaVA-7B の方が良いスコアを達成していることがわかる．

図 A·1 および**図 A·2**(共に参考資料に示す) は，カロリー量推定における各モデルの応答例である．ファインチューニングが実施されていない LLaVA-13B と GPT-4V による推定では，カロリー量の推論過程が出力されながら最終的なカロリー量が導かれている様子がわかる．これに対して，ファインチューニング後の LLaVA-13B では，推定値が直接出力されている様子がわかる．

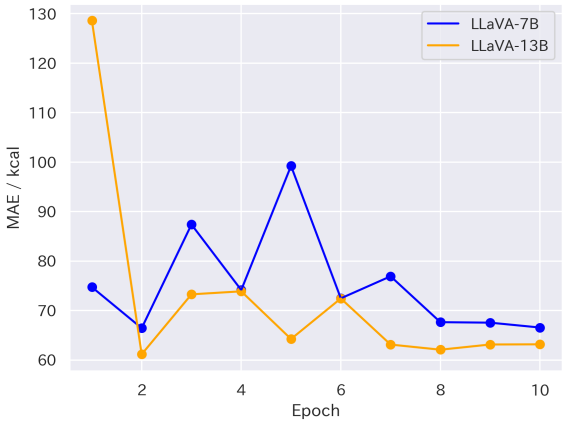

図 6 学習エポック数に対するカロリー量推定結果の平均絶対誤差（MAE）の変化

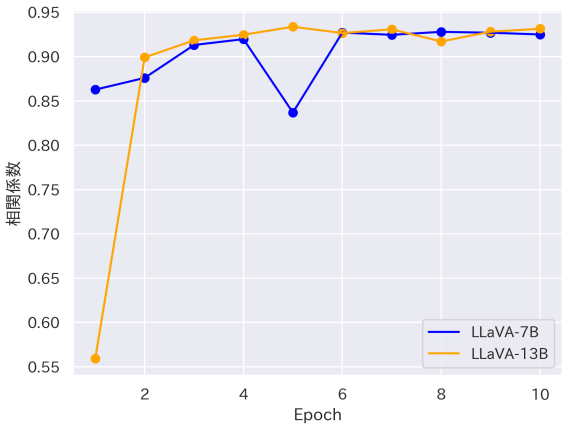

図 8 学習エポック数に対するカロリー量推定結果の相関係数の変化

4.2 食品体積推定による方法

表 2 は Nutrition5k のテスト用分割に対するゼロショットカロリー量推定の結果である．GPT-4V と食品体積推定（Vol）を組み合わせた手法が，LLaVA-13B や GPT-4V に対して MAE について高いスコアを達成した．一方で，LLaVA-13B と食品体積推定を組み合わせた手法は，外れ値の影響で各指標のスコアが著しく悪化した．

図 9 および**図 10** は，ゼロショットカロリー量推定の推定値と正解値の散布図である．全体としてはあまり大きな違いは見られないものの，相関係数の値は食品体積推定モデルを組み合わせた方が高くなっている．

図 A·3 から**図 A·5** は，ゼロショットカロリー量推定における各モデルの応答例である．図 A·3 では，本手法による体積推定の結果がカロリー量推定の過程において考慮されることで，推定結果が大きく改善されている様子がわかる．**図 A·4** では，食事画像に含まれている複数の食品について，体積が考慮されることで具体的な値が定まったカロリー量推定が行われている様子がわかる．図 A·5 は外れ値が出力されたときの推論過程である．はじめに推定され

た食品のカロリー量がカロリー量密度として誤認識された上で，その値と与えられた食品体積の値が掛け合わされるような推論が行われることで，外れ値が生じていることがわかる．こうした推論の品質は大規模言語モデルの推論性能に大きく基づいていると考えられるとともに，強力な大規模言語モデルに基づく GPT-4V ではこうした現象が生じていないことから，大規模言語モデルの推論性能の改善によりカロリー量の推定結果が改善されると考えられる．

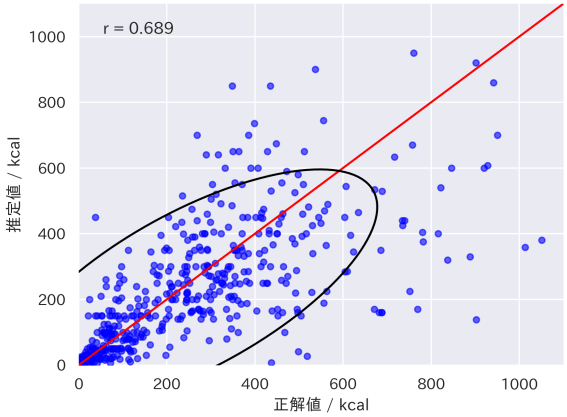

図 9　GPT-4V の推定カロリー量の散布図

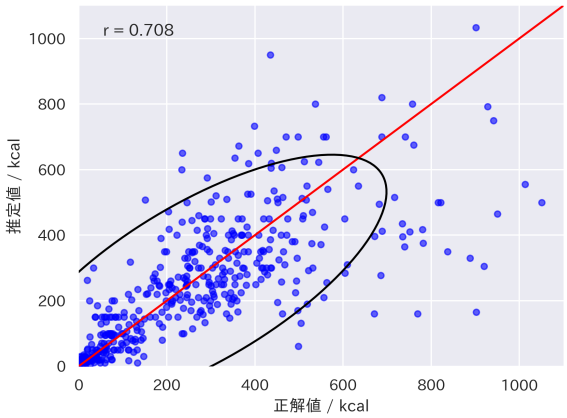

図 10　GPT-4V と食品体積推定モデルによる推定カロリー量の散布図

図 11 は体積推定の結果の例を物体検出，領域分割，深度推定の結果と合わせて示したものである．物体検出と領域分割については，皿と食品について適切な領域が抽出されており，全体として高品質な推定が実現されていることがわかる．深度マップの推定においても，同一種類の食品の中で凹凸のある部分の差が詳細に表現されていることが確認される．また，複数種類の食品が写っている画像についても，特に高さの異なる部分の深度の値が周辺の値に対して大きく異なるように推定されている様子がわかる．

本研究で提案した食品体積推定モデルには，次の二点において体積を過剰に推定してしまう性質があると考えられる．第一に，食品下面から皿基準面までの体積が余計に算出されてしまう点である．第二に，皿の最低面が食品により覆われている場合に，誤った皿基準面が選択されてしまう点である．こうした課題に対して，DepthCalorieCam [1] のように質量回帰モデルを作成する方法や，成富ら [12] のように皿と食品の高品質な 3D 形状を再構成する方法が考えられる．しかし，いずれも大量の食品データが必要となり，それを用意するための負担がかかる点が課題となる．

表 2　食品体積推定によるゼロショットカロリー量推定の結果

項目	MAE / kcal ↓	MAPE / % ↓
LLaVA-13B	109.6	92.8
GPT-4V	106.6	**54.8**
LLaVA-13B+Vol（Ours）	6122.7	6591.4
GPT-4V+Vol（Ours）	**101.7**	56.8

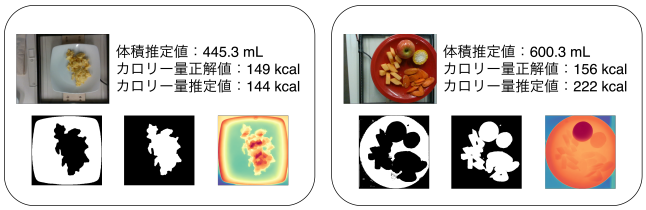

図 11　物体検出，領域分割，および深度推定の結果．各枠左上：元画像，右上：推定値は GPT-4V+Vol の場合，左下：皿領域マスク，中央下：食品領域マスク，右下：深度マップ．

5.　おわりに

本研究では，食事画像からのカロリー量推定タスクに対して，大規模視覚言語モデルを用いる 2 つのアプローチの有効性を検証した．第一には，大規模視覚言語モデルをファインチューニングする手法であり，Nutrition5k による評価でベースラインや同時期の大規模視覚言語モデルに匹敵する性能を達成した．第二には，大規模視覚言語モデルと食品体積推定モデルを組み合わせる手法であり，LLaVA や GPT-4V といったモデルに対して MAE と相関係数について優れたゼロショットカロリー量推定性能を達成した．

近年の手法との比較に基づくと，今後は FoodLMM のようなモデルに単眼深度推定モデルを導入することが特に有望であると考えられる．また，今回用いた Nutrition5k には食事ドメインの中でも限られた食事画像しか含まれておらず，ユーザーが日常的に食べる食品のドメインに関しては本研究の手法の有効性が十分に検証されていない．例えば，寿司やラーメンといった日本においてポピュラーな食品は本データセットに含まれていない．そうしたドメインにおける検証は本手法を実用化する上で重要である．

参考文献

[1] Yoshikazu Ando, Takumi Ege, Jaehyeong Cho, and Keiji Yanai. 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.

[2] Takumi Ege and Keiji Yanai. 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.

[3] Yuehao Yin, Huiyan Qi, Bin Zhu, Jingjing Chen, Yu-Gang Jiang, and Chong-Wah Ngo. FoodLMM: A versatile food assistant using large multi-modal model. *arXiv preprint arXiv:2312.14991*, 2023.

[4] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proc. of International Conference on Machine Learning*, pp. 8748–8763, 2021.

[6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality, 2023. `https://lmsys.org/blog/2023-03-30-vicuna/`.

[7] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proc. of International Conference on Learning Representations*, 2022.

[8] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[10] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. *arXiv preprint arXiv:2312.02145*, 2023.

[11] Quin Thames, Arjun Karpur, Wade Norris, Fangting Xia, Liviu Panait, Tobias Weyand, and Jack Sim. Nutrition5k: Towards automatic nutritional understanding of generic food. In *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 8903–8911, 2021.

[12] Shu Naritomi and Keiji Yanai. 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.