

大規模視覚言語モデルを用いた 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。安藤ら [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。

近年、自然言語処理分野において、モデルパラメータ数・データ数・計算量を大規模な条件にして学習された言語モデルである大規模言語モデルが、様々なタスクにおいて高い性能を達成している。こうしたモデルには、学習条件の大規模化に対して性能がべき乗則に従い向上することや [3]、特に学習条件を大規模化していくとある段階で性能が飛躍的に向上する創発的能力 [4] が報告されており、従来の言語モデルには確認されない新たな側面に注目が集まっている。

大規模言語モデルを視覚に関するモダリティに拡張したものは大規模視覚言語モデルと呼ばれる。大規模言語モデルを視覚に関するモダリティに拡張するためには、画像やビデオのような視覚情報を視覚特徴に変換する視覚エンコーダが用いられる。Flamingo [5] は、視覚エンコーダにより得られる視覚特徴とテキスト特徴をゲート付きクロスアテンション層で融合することで、画像やビデオに関する様々なタスクをこなす単一のモデルを作成した。LLaVA [6] は、線形層または MLP により構築した接続層で視覚特徴を変換し、大規模言語モデルに入力するモデル構造をとった。また、視覚情報を含めた指示形式のデータによる学習フレームワークである Visual Instruction Tuning により、画像に関する高品質なタスク汎化性能を獲得した。他にも、接続層に Q-Former を導入した BLIP-2 [7] や、そのモデル構造を基として Visual Instruction Tuning と同様

¹ 電気通信大学
UEC, Chofu, Tokyo 182-8585, Japan
a) tanabe-h@mm.inf.uec.ac.jp
b) yanai@cs.uec.ac.jp

の学習を適用した MiniGPT-4 [8] や InstructBLIP [9] などが存在する。

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 [10] 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 [11] 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 [12] 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 [13] 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。なお、大規模言語モデルの学習には LoRA [14] を適用する。

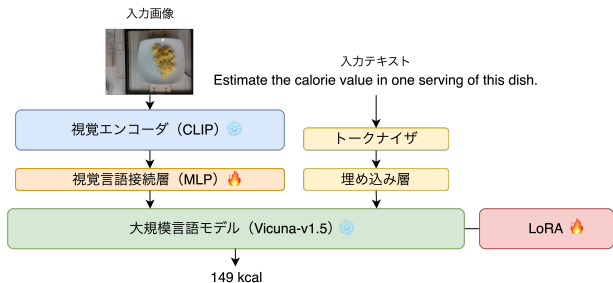


図 1 ファインチューニングによる手法のモデル構造

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)。ただし、食事画像に次の仮定を置くことに注意する。まず、ピクセルあたりの実面積は既知であるものとする。また、撮影は食品の真上から行い、そのときの撮影位置から皿上面の最深点までの距離が既知であるものとする。

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 [15] により、皿の矩形領域を得る。次に、これに関心領域として 3 つに処理が分かれる。第一に、Segment Anything (SAM) [16] 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

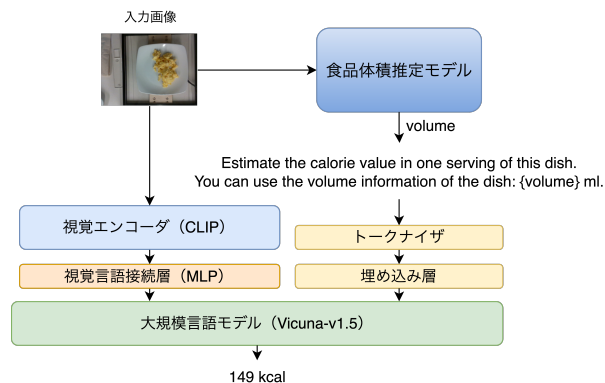


図 2 食品体積推定モデルを導入する手法の全体図

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 [17] を関心領域に適用し、深度マップを得る。

こうして得られた 2 つのマスクと深度マップに基づいて、実体積を推定する。まず、深度マップと各マスクとのピクセルごとのアダマール積をとることで、各マスクの領域を切り出す。次に、深度マップの皿領域内における最大値を取得し、これを皿基準面の深度とする。そして、食品領域の深度マップの各値に対して、皿上面の基準の深度との差を取る。これにより、各ピクセルにおける皿基準面から食品上面までの高さや撮影位置までの高さの比が得られる。さらに、皿基準面から撮影位置までの実寸高さとの比により、各ピクセルにおける皿基準面から食品上面までの実寸高さを得る。最後に、この高さやピクセルあたりの実面積を基に、食品領域の深度マップの全てのピクセルに対する体積積分を実施する。なお、このときの体積積分の値は (1) による計算により求められる。ただし、 V は食品の実体積、 H_{ij} は i 行 j 列のピクセルの実寸高さ、 A_{ij} は i 行 j 列のピクセルの実面積を表す。

$$V = \sum_{i=1}^n \sum_{j=1}^m D_{ij} A_{ij} \quad (1)$$

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。

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 [18] の overhead データセットを学習と評価に用いた。これは、3265 枚の真上から撮影された食事画像とカロリー量をはじめとした栄養素の情報が含まれており、2759 件と 506 件でそれぞれ訓練用とテスト用に分割されている。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.

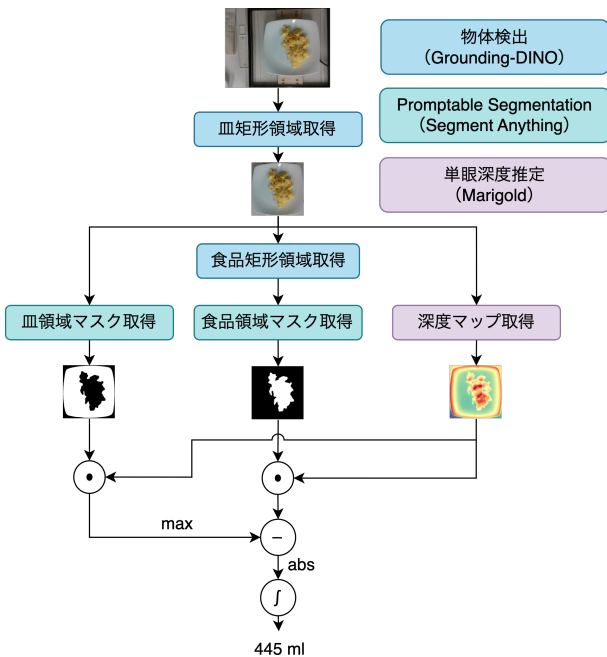


図3 食品体積推定モデルの構造

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。さらに、訓練用分割を 8 : 2 で分割してそれぞれを検証用訓練分割と検証用分割とした。そして、検証用訓練分割で学習したモデルに対する検証用分割の損失の監視に基づいて、使用する LLaVA-7B, 13B のチェックポイントとしてそれぞれ 6, 5 エポック目のものを選択した。

学習時には、最適化アルゴリズムとして AdamW を使用し、学習率スケジューリングとして最大値 2×10^{-4} の linear warmup および cosine decay を適用した。また、LoRA の行列分解で使用するランクは 128 とし、バッチサイズは 64 とした。テキスト生成における温度 (temperature) パラメータは 0 とし、トークンが決定的に出力されるようにした。さらに、学習時には DeepSpeed ライブラリ [19] によるデータ並列化と、ZeRO-3 [20] によるメモリ最適化を実施した。学習は $4 \times$ RTX 3090 で 10 エポックまで学習を行い、約 5 時間がかかった。

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) [18] 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。

図 4 および図 5 は、ファインチューニング前後の LLaVA-13B の推定カロリー量の分布である。ファインチューニング前のモデルについては、特に 200 kcal 周辺に多くの誤推定を行っていることがわかる。一方で、ファインチューニング後のモデルは、その周辺の誤推定が著しく減ってお

り、その前後に存在する正解値データが密集している区間において多くの推定を行っていることがわかる。

図 6 および図 7 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 [18]	70.6	26.1
LLaVA-7B	178.8	129.5
LLaVA-13B	177.1	92.8
GPT-4V	106.6	54.8
FoodLMM FT [10]	67.3	26.6
LLaVA-7B FT (Ours)	74.2	41.5
LLaVA-13B FT (Ours)	64.3	39.8

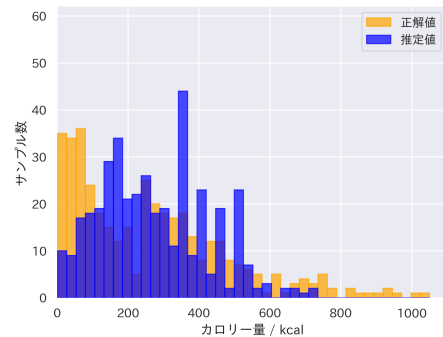


図 4 ファインチューニング前の LLaVA-13B の推定カロリー量の分布

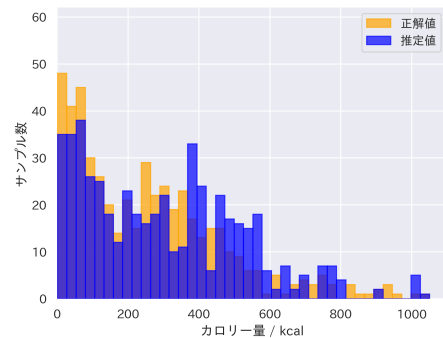


図 5 ファインチューニング後の LLaVA-13B の推定カロリー量の分布

図 8 から図 10 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減少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

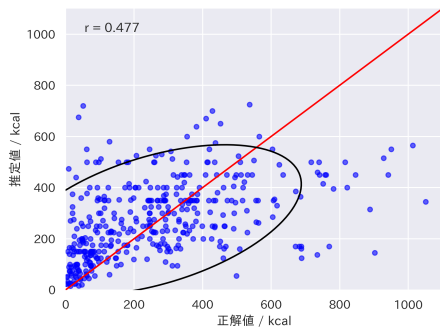


図 6 ファインチューニング前の LLaVA-13B の推定カロリー量の散布図

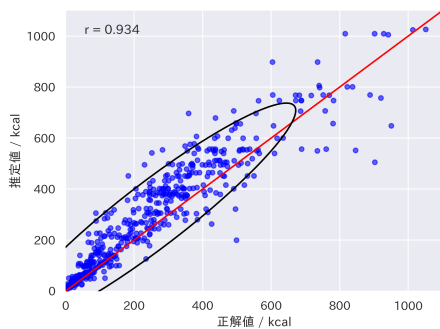


図 7 ファインチューニング後の LLaVA-13B の推定カロリー量の散布図

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 13 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見られる。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見られる。

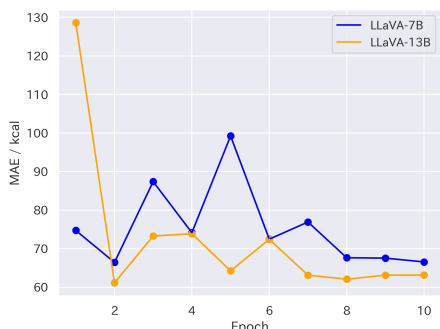


図 8 学習エポック数に対するカロリー量推定結果の平均絶対誤差 (MAE) の変化

4.2 食品体積推定による方法

本実験では、節 4.1 とは異なり追加の学習は行っていないことに注意する。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

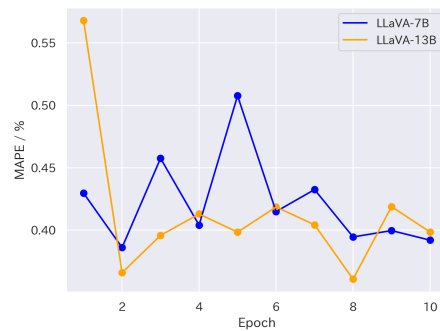


図 9 学習エポック数に対するカロリー量推定結果の平均絶対パーセント誤差 (MAPE) の変化

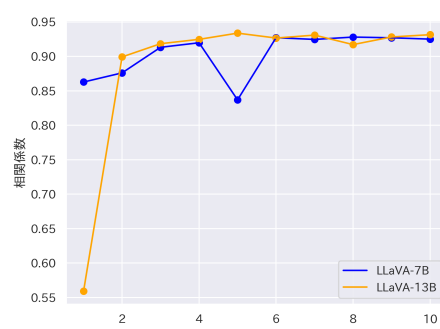


図 10 学習エポック数に対するカロリー量推定結果の相関係数の変化

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。

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 および図 12 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 14 および図 15 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 14 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見られる。図 15 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モデルである 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。

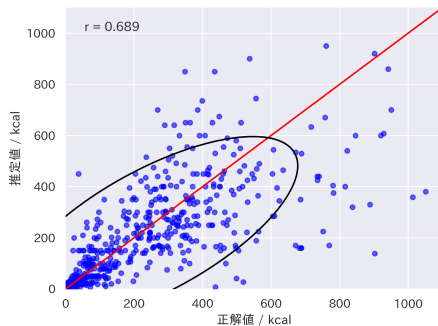


図 11 GPT-4V の推定カロリー量の散布図

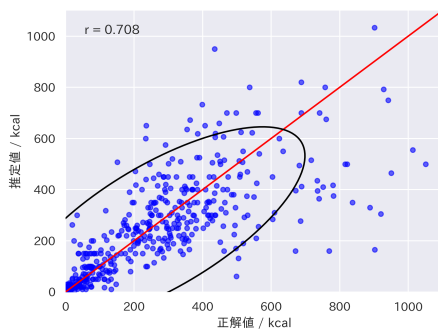


図 12 GPT-4V と食品体積推定モデルによる推定カロリー量の散布図

図 16 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 [21] 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。

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、今後は FoodLMM のように特殊トークンと回帰ヘッドに基づいてカロリー量を推定するモデルに、単眼深度推定モデルを導入してマルチタスク学習を適用することが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] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- [4] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.
- [5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information*



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.

LLaVA-13B FT
(Ours)

[[180]] calories.

図 13 各大規模視覚言語モデルの推定例

- Processing Systems*, Vol. 35, pp. 23716–23736, 2022.
- [6] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.
- [7] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proc. of International Conference on Machine Learning*, 2023.
- [8] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- [9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.
- [10] Yuehao Yin, Huiyan Qi, Bin Zhu, Jingjing Chen, Yungang Jiang, and Chong-Wah Ngo. FoodLMM: A versatile food assistant using large multi-modal model. *arXiv preprint arXiv:2312.14991*, 2023.
- [11] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.
- [12] 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.



Answer: **687 kcal**

User

Estimate the calorie value in one serving of the dish.
Return the single calorie value in the following format: "[[x]] calories".
This information is essential and must be provided in every response.

Example: "[[100]] calories"

GPT-4V



This is a bowl of almonds. A typical serving size of almonds is about 1 ounce (28 grams), which is roughly 23 almonds and contains approximately [[160]] calories. However, without a reference for the exact amount in this bowl, it's difficult to provide an accurate calorie count. If this bowl contains about one serving, then the calorie estimate would be around **[[160]] calories**. If it contains more, the calorie count would be higher accordingly.

User

Estimate the calorie value in one serving of the dish.
Return the single calorie value in the following format: "[[x]] calories".
This information is essential and must be provided in every response.

Example: "[[100]] calories"

You can use the volume information of the dish: 527 ml.

GPT-4V+Vol
(Ours)

The image shows a bowl of almonds. Almonds are a nutrient-dense food, and a typical serving size of almonds is about 1 ounce (28 grams), which is approximately 23 almonds. This serving size generally contains about 160 to 170 calories. However, the bowl in the image appears to contain more than a single serving.

Without knowing the exact weight or number of almonds in the bowl, it's challenging to provide an accurate calorie count. If we assume the bowl contains around 3 servings (which is a rough estimate based on visual assessment), the calorie content would be approximately 480 to 510 calories.

Given the volume information of the dish (527 ml), and assuming the bowl is filled mostly to the top, it could potentially hold around 4 to 5 ounces of almonds (since almonds are not a liquid and their volume would not directly convert to weight). This would mean the bowl might contain around 640 to 850 calories, depending on the exact amount and whether the almonds are raw or roasted, which can slightly affect the calorie content.

For the purpose of providing a single calorie value, I will estimate the calorie content of the bowl of almonds as follows:

[[700]] calories

Please note that this is a rough estimate and the actual calorie content could be higher or lower depending on the precise amount of almonds in the bowl.

図 14 GPT-4V および食品体積推定モデルを組み合わせたモデルの推定例

- 8748–8763, 2021.
- [13] 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/>.
- [14] 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.
- [15] 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.
- [16] 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.



Answer: 149 kcal

User

Estimate the calorie value in one serving of the dish.
Return the single calorie value in the following format: "[[x]] calories".
This information is essential and must be provided in every response.

Example: "[[100]] calories"

You can use the volume information of the dish: 445 ml.

LLaVA-13B+Vol
(Ours)

To estimate the calorie value in one serving of the dish, we need to know the ingredients and their respective calorie counts. Since the image shows a plate of scrambled eggs, we can assume that the dish is primarily composed of eggs.

The calorie value of one serving of scrambled eggs can vary depending on the number of eggs used and the cooking method. However, as a rough estimate, we can assume that one serving of scrambled eggs contains around 200-250 calories.

To calculate the calorie value for the entire dish, we need to know the volume of the dish, which is 445 ml. To convert the volume to weight, we can assume that the dish is made up of 445 grams of scrambled eggs.

So, the calorie value of the dish would be approximately **98,000 to 112,000 calories**. However, this is a rough estimate and may not be accurate, as the actual calorie count can vary depending on the specific ingredients and cooking method used.

図 15 LLaVA-13B および食品体積推定モデルを組み合わせたモデルの外れ値推定例

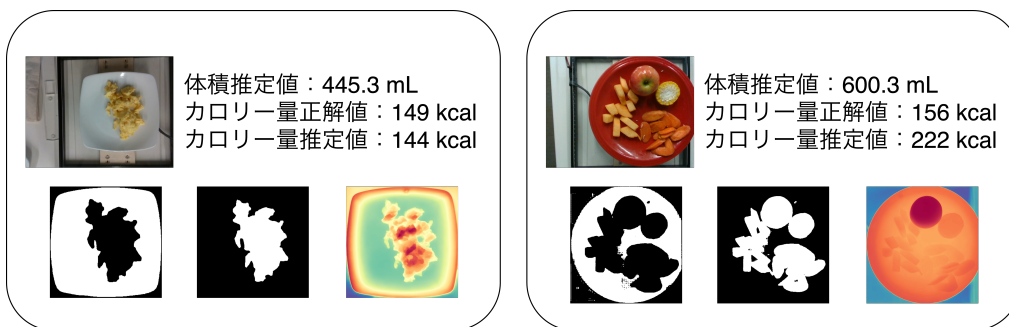


図 16 物体検出, 領域分割, および深度推定の結果. 各枠左上: 元画像, 右上: 推定値は GPT-4V+Vol の場合, 左下: 皿領域マスク, 中央下: 食品領域マスク, 右下: 深度マップ.

- [17] 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.
- [18] 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.
- [19] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 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.
- [20] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–16, 2020.
- [21] 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.