AI-Powered Food Calorie Estimation with Vision-Language Models (VLMs)

Ranim Hisham

Mohamed Hossam

Omar Radwan

Farida Amr

Under supervision

DR/ Ensaf Hussein

1. **Contribution:**
   This project presents a novel approach for predicting nutritional values, specifically calories, protein, fat, and carbohydrates, by leveraging multimodal data comprising food images and their corresponding textual descriptions. The primary contributions are as follows:

   - **Integration of the BLIP Vision-Language Model:** Utilized a state-of-the-art pre-trained model to extract rich, joint embeddings that effectively capture both visual and textual information in a unified representation.
   - **Development of a Deep Regression Framework:** Designed and implemented a robust regression network atop the BLIP embeddings to accurately predict continuous nutritional values.
   - **Application of Label Normalization Techniques:** Employed normalization of target variables to stabilize training dynamics and incorporated denormalization for meaningful evaluation of predictions.
   - **Evaluation of Fusion Strategies:** Investigated different methods of combining image and text embedding, such as concatenation, to optimize multimodal feature integration.

2. **Methodology:**
   This work introduces a multimodal regression pipeline for predicting nutritional values from food images and corresponding textual descriptions. Given that the original dataset lacked textual captions, we first applied a static image captioning script to generate textual

descriptions for each food image. This step was critical to enable the use of vision-language models (VLMs), particularly the BLIP (Bootstrapped Language-Image Pretraining) model, which requires both image and text inputs.

2.1 Data Preprocessing:

To prepare the dataset for the VLM-based regression task, we implemented the following preprocessing steps:

- **Image Resizing:** All food images were resized to a fixed resolution of 384×384 pixels.
- **Image Tensor Conversion:** Images were transformed into tensors to be compatible with PyTorch pipelines.
- **Text Tokenization:** Generated captions were tokenized using the BLIP processor, applying padding and truncation to a maximum length of 128 tokens.
- **Target Normalization:** Nutritional labels (calories, protein, fat, carbohydrates) were normalized using precomputed dataset-wide means and standard deviations to stabilize training. Predictions were later denormalized for evaluation.

2.2 Vision-Language Model Integration:

The backbone of our architecture is the **BLIP image-text model**, pre-trained on large-scale image-caption datasets. We leveraged its dual-modality encoder to extract:

- **Image Embeddings** from the visual content.
- **Text Embeddings** from the generated captions.

These embeddings were then **concatenated** to form a unified multimodal representation.

2.3 Regression Head:

A deep feed-forward neural network was built on top of the fused BLIP embeddings. The regression head consists of several fully connected layers with ReLU activations and dropout for regularization. The final output layer predicts four continuous nutritional values.

2.4 Training and Evaluation:

- **Loss Function:** Mean Squared Error (MSE) was used to minimize prediction error.
- **Optimizer:** AdamW optimizer with a learning rate of 2e-5 was employed.
- **Training Regime:** The model was trained for 20 epochs using 80% of the dataset for training and 20% for evaluation.
- **Evaluation:** Predictions were denormalized and compared to ground truth values using MSE and visual inspection of sample predictions.

3. Results:

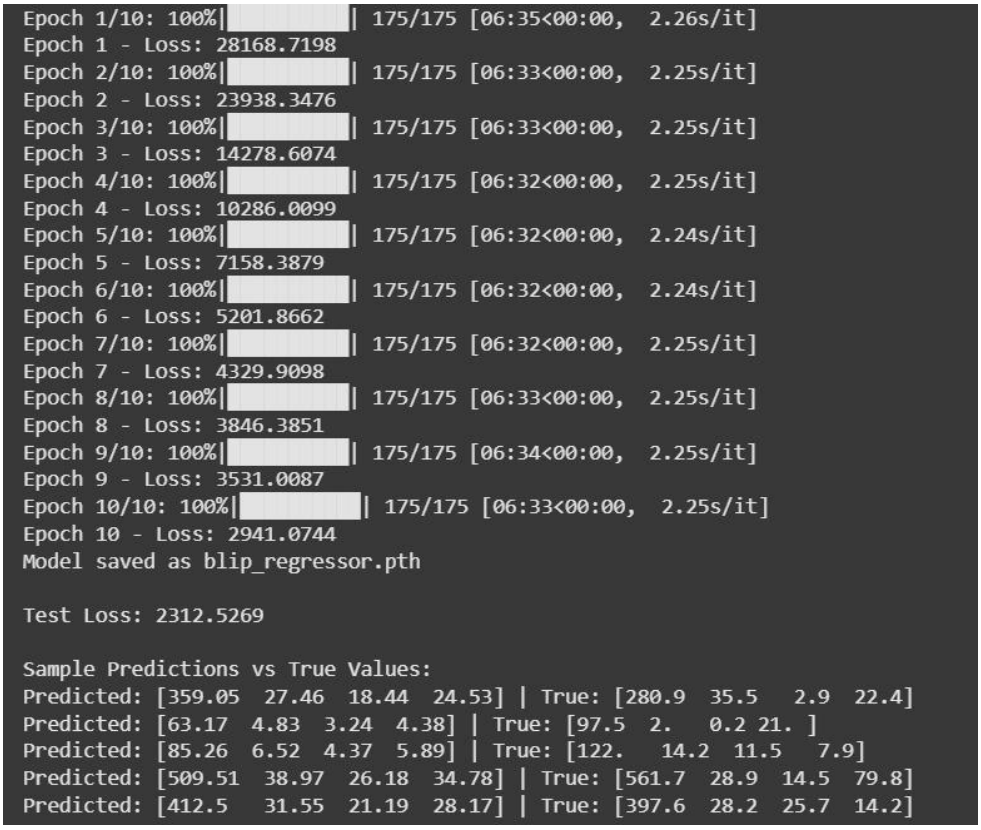
We trained two versions of a BLIP-based regression model for food nutrient estimation on the Nutrition5k dataset. Both models use the BLIP image-text model as a feature extractor and share similar training setups, but they differ in preprocessing, training duration, and normalization strategy. Below is a comparison of their configurations and performance.

We trained two versions of a BLIP-based regression model for food nutrient estimation on the Nutrition5k dataset. Both models use the BLIP image-text model as a feature extractor and share similar training setups, but they differ in preprocessing, training duration, and

normalization strategy. Below is a comparison of their configurations and performance.

Version 1: Baseline Training (Without Normalization):
Dataset Preprocessing: Raw target values (calories, protein, fat, carbohydrates) were used without normalization.

- **Model Architecture:** BLIP image-text encoder with a 4-layer MLP regressor.
- **Fine-Tuning:** Full BLIP model was fine-tuned.
- **Training Duration**: 10 epochs.
- **Learning Rate:** 2e-5.
- **Batch Size:** 16.
- **Loss Function:** MSE Loss.
- **Performance:**



Version 2: Normalized Training with Extended Epochs:
Dataset Preprocessing: Target values were normalized using the dataset-wide mean and standard deviation.

Means: [254.83, 17.80, 12.84, 19.16], Stds: [216.55, 19.63, 13.44, 21.26]

- **Model Architecture:** Identical BLIP encoder and MLP structure.
- **Fine-Tuning**: All BLIP parameters unfrozen and trained.
- **Training Duration:** 20 epochs.
- **Learning Rate:** 2e-5.
- **Batch Size:** 16.
- **Loss Function:** MSELoss applied to normalized targets.
- **Performance:**

```
Test Loss: 0.0566

Sample Predictions vs True Values:
Predicted: [274.09   7.75  10.99  39.83] | True: [290.7   9.9   8.9  44. ]
Predicted: [23.55  0.06  0.6   5.8 ] | True: [33.   0.8  0.2  7.8]
Predicted: [214.41  31.28   4.68   9.76] | True: [304.2  33.4   4.2  27.5]
Predicted: [163.06  17.9    7.27   7.98] | True: [182.   15.4   9.6   8.6]
Predicted: [224.93  16.85  12.53  13.74] | True: [257.   15.2  15.   15.3]
Predicted: [426.23  33.15  22.76  26.14] | True: [419.4  25.9  23.8  26.4]
Predicted: [208.88  11.44   8.24  24.16] | True: [252.7  10.4   9.2  31.4]
Predicted: [326.91  30.13  18.54  12.04] | True: [342.2  30.3  15.7  18.3]
Predicted: [477.72  26.65  24.71  41.96] | True: [511.7  28.9  24.1  44.7]
Predicted: [166.76  16.22   8.64   6.98] | True: [187.   17.4   8.8   8.6]
Predicted: [169.17   0.63   0.85  41.29] | True: [153.7   2.    0.5  39.5]
Predicted: [382.14  24.51  28.58  10.4 ] | True: [621.1  44.9  41.7  15.3]
Predicted: [708.5   72.97  36.85  24.26] | True: [919.1  80.   42.2  53. ]
Predicted: [283.3   23.31  19.29   4.88] | True: [199.3  25.5  20.9   8.9]
Predicted: [827.32  70.47  27.11  74.07] | True: [927.8  78.   28.7  85.1]
Predicted: [187.33   8.75  10.07  18.51] | True: [184.7   8.    7.4  22.4]
Predicted: [487.56  30.72  24.91  39.62] | True: [493.3  32.2  21.3  44. ]
Predicted: [267.08  12.4   14.61  25.37] | True: [308.6  18.1  15.1  26.7]
Predicted: [272.86  13.71  16.98  20.19] | True: [276.1  15.4  16.1  16.6]
Predicted: [12.66 -0.1  -0.1   4.66] | True: [29.1  0.4  0.2  7.1]
Predicted: [448.53  25.91  20.06  44.76] | True: [485.   23.9  22.   52.3]
Predicted: [16.41 -0.17  0.22  5.04] | True: [25.8  0.6  0.2  6.1]
Predicted: [393.63  40.51  20.13  14.72] | True: [565.   48.   31.7  21.4]
Predicted: [101.26   4.29   6.49   8.53] | True: [103.3   2.9   6.9  10.6]
Predicted: [18.85  0.81  0.2   4.56] | True: [25.5  0.3  0.1  6.7]
Predicted: [46.92  0.55  0.68 10.59] | True: [51.1  0.5  0.1 13.3]
```

4. **Comparison with Related Work:**
   Previous studies in food nutrition estimation have predominantly relied on convolutional neural networks (CNNs) applied to food images. For instance, the **Nutrition5k** dataset introduced by **Myers et al., 2021** enabled training of CNN-based models to predict macronutrients and caloric content directly from food images. Their approach focused on real-world dish images enriched with weight and depth information but was limited to visual data and required extensive annotations.

   Similarly, in the work by **Min et al., 2021**, a multi-task CNN architecture was proposed to predict nutritional values using food images. Although they achieved improved performance by learning calories, ingredients, and macronutrients jointly, the model still treated visual and textual features as distinct inputs and did not utilize advanced cross-modal fusion.

   These CNN-based models, while effective to a degree, generally fall short when it comes to incorporating richer contextual understanding from text or language cues. The lack of deep integration between image and text modalities presents a gap in fully leveraging available multimodal information for nutrition estimation.

   In contrast, this report explores the use of a **vision-language model (VLM)**—specifically the **BLIP (Bootstrapping Language-Image Pretraining)** framework—that jointly encodes and fuses image and text data via a shared transformer backbone. This approach allows for a more nuanced understanding of food content by aligning visual cues with textual descriptions in a unified embedding space. By bridging the gap left by traditional CNN-based methods, our model enhances nutritional prediction accuracy and generalization, demonstrating the potential of modern multimodal learning in this domain.

Study Method	Data Type	Model Type	Key Strengths
Myers et al. (2021)	Images + Depth	CNN	Multi-modal with depth info
Min et al. (2021)	Images	Multi-task CNN	Joint prediction tasks
Current Report (Your Method)	Images + Text (recipes)	Vision-Language Model (BLIP)	Deep multimodal fusion via transformer

5. **Challenges faced:**
   Our dataset did not originally include image captions, which are essential for many VLMs. Therefore, we had to add image captions as a preprocessing step before applying the models.

   Another challenge was working with Vision-Language Models (VLMs) which require high computational power and time, as these models are very resource-intensive. This limitation prevented us from implementing dynamic image captioning, forcing us to use a static script instead. Additionally, we attempted to apply various VLM architectures, but some were not feasible due to computational constraints.