Caloric Prediction from RGB Image

Maor Moshe, David Oriel, and Shachar Ashkenazi
Submitted to: Dr. Yuval Benjamini, Lee Carlin

15/03/2025

Abstract

Accurate nutritional analysis is essential for advancing health and dietary awareness. In this work, we explore the problem of predicting the caloric value of a dish from its RGB image, a task that typically relies on complex and detailed datasets such as Nutrition5k. Such datasets are complex and include RGB images, depth images and videos for a single prediction, our approach on the other hand aims to use RGB images only, with the goal of simplifying the prediction process. To enhance model performance, we propose incorporating ingredient lists as an additional side task, providing textual information that aids in better understanding the dish's composition. We investigate various Vision-Language Models (VLMs) and other architectures that jointly process image and text embeddings. Through varied experimentation, we show that combining image-based features with ingredient list data leads to significant improvements in the accuracy of caloric prediction, demonstrating the effectiveness of multi-modal learning for this task. Our results suggest that even without access to intricate datasets, a multi-modal approach can offer a robust solution for predicting the caloric content of food dishes from images only. This research highlights the potential for democratizing nutrition information, offering a scalable solution that could support healthier eating habits on a global scale

1 Introduction

Accurately estimating the nutritional content of meals is integral to maintaining overall health and well-being, yet even professional nutritionists find it challenging to assess caloric and macronutrient values based on images alone. Variations in portion sizes, cooking methods, and hidden ingredients (e.g., oils and sauces) make purely visual estimation highly error-prone. As a result, manual approaches to food tracking are labor-intensive and frequently yield imprecise dietary records. Recent advances in machine learning, particularly multimodal models, offer a promising path forward by automating the estimation of meal nutrition. These methods can reduce human error and streamline dietary logging by analyzing meal images and incorporating additional signals such as ingredients. In 2021 Google research has published a paper called Nutrition5k, demonstrating nutrition facts prediction of real world dishes at an accuracy that outperforms professional nutritionists. However, the datasets used in Nutrition5k, composed of videos, depth images, RGB images, and ingredient annotations, such large-scale data sets can be computationally restrictive for many real world uses. In this work, we focus on predicting caloric values from meal images, thus eliminating video-based analysis and depth images while aiming to maintain high predictive accuracy. Our proposed methodology employs a two-stage approach: first, we train Efficient.Net (neural network architect) to extract key meal components from the images, and then we integrate both the segmented features and the overall image for caloric prediction. The additional information of meal ingredients should allow for more accurate estimation of calories, and could also give reason for the calculation, as it has access to its predicted ingredients. We also explore the use of Vision-Language Models (VLMs) to improve performance by extracting ingredient-level textual information. We then provide a comparative analysis of various neural network architectures, comparing multimodal approaches with purely image-based models in the purpose of quantifying the value of integrating language and vision for calorie estimation. By offering an efficient, scalable framework for automated caloric prediction, this study underscores the potential of machine learning to make real-time food logging more accessible. Ultimately, these advances may enhance personal nutrition management and public health efforts worldwide.

2 Data

Our data set is RGB images (256X256 pixels) from Nutrition5k dataset, originally containing 5,000 examples. However, during our cleaning process we discovered several images that were completely black, and we found that the test set was empty. After contacting the dataset creation team at Google, we learned that the test set was intentionally hidden. As a result, our final dataset comprised a training set of 2,600 examples and a test set of 200 examples. The dataset includes information on meal caloric values, ingredients, and other pertinent details.The dishes in Nutrition5k vary drastically in portion sizes and dish complexity, with dishes ranging from just a few calories to over 1000 calories and from a single ingredient to up to 35 with an average of 5.7 ingredients per plate. It is designed to reflect real-world conditions through its diversity—capturing a wide range of ingredients, portion sizes, and dish complexities—as well as its realistic nature, with photographs taken in actual campus cafeterias. Additionally, the images present inherent challenges, such as partially or fully occluded ingredients, that mirror the ambiguities found in everyday meals. Notably, the dataset is limited to Western cuisine and does not include certain types of foods, such as mashed dishes or soups, where accurately estimating caloric content can be particularly challenging.

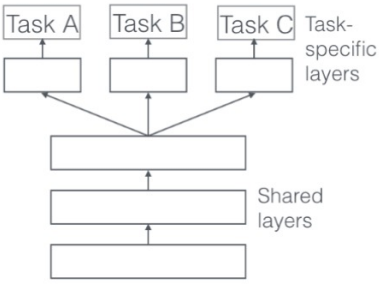
3 Methods

In this chapter, we review methods used in the nutrition 5k model by Google, as well as methods we employed.

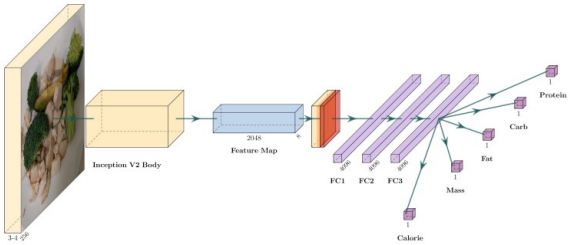
3.1 Nutrition5k Model

The original Nutrition5k model uses RGB frames and depth images from rotating video sequences of each dish. It is a multi-task learning model aimed at predicting the total mass of the food item alongside its caloric value and macronutrient composition (i.e., protein, carbohydrate, and fat content). The model processes data through an InceptionV2 encoder to extract a high-level feature map, followed by fully connected layers that produce predictions for each task. This structure is trained end-to-end using mean absolute error (MAE) loss on each regression sub-task, with parameters optimized via RMSProp. By splitting the network's final fully connected layers, a shared feature representation is maintained while allowing each task to learn domain-specific adjustments.

subcaption



(a) Multi Task Learning models concept.



(b) Overview of Google's network architecture for our multi-task learning, the model receives the dataset, and outputs predictions on calorie, mass and macronutrient composition.

Figure 1: Original concept (left) and implementation of the concept (right).

In addition to predicting the absolute caloric and macronutrient values ("direct prediction"), the authors also introduced a portion-independent model. In this setup, the target values were normalized by the dish's total mass, effectively transforming the problem into predicting calories and nutrients per gram rather than absolute totals. Despite reducing the complexity of portion-size variations, the portion-independent model still required multiplying the predicted per-gram values by the ground-truth mass to reconstruct the final caloric and macronutrient estimates. Lastly, for experiments incorporating depth data, the authors combined RGB and depth images in an effort to capture more accurate geometric information about the portion sizes.

Through these approaches, the Nutrition5k paper demonstrated the viability of large-scale, automated nutritional analysis pipelines. However, the reliance on multi-modal data—particularly videos and depth images—posed significant computational demands. This, in turn, served as motivation for lighter, image-only methods such as the ones presented in our work.

3.2 Proposed Methodologies

We present results from four models that explore different approaches to achieving improved performance over the original paper's results

Model 1 – Updated Original Architecture: Our first model strictly follows the architecture proposed in the paper, with one key modification: we replace the outdated Inception v2 with the more advanced Inception v3. No additional changes were made, ensuring that this model mirrors the original process precisely. Inception is a complex CNN, that allows the model to learn filters of different sizes directly from the image, increasing the information each filter has access to; in inception architecture one connects the different convolutional layers in parallel, in contrast to regular CNN where convolutional layers are connected in series (one after the other)

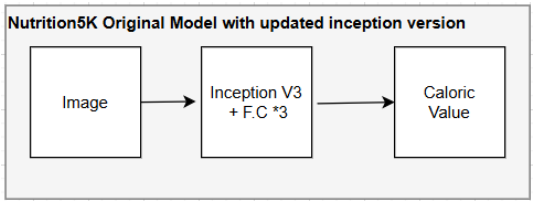


Figure 2: Nutrition 5k original model with updated inception version

Model 2 – Enhanced Complexity with ResNet and Attention: In the second model, we pursue a more complex architecture by integrating a ResNet with 50 layers along with an attention layer. This design aims to better capture and emphasize critical features within the input images. We optimized the depth of ResNet and found 50 layers to be optimal.

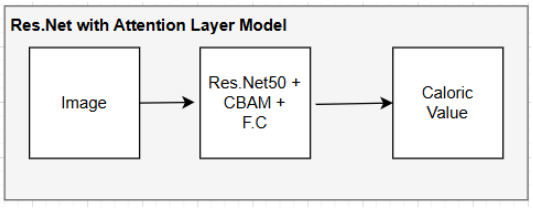


Figure 3: Res.Net with Attention Layer Model

Model 3 – Augmented Input with Object Detection and Embeddings: The third and final model begins by detecting the ingredients in a dish and encoding them using an embedding model. For ingredient identification, we employ EfficientNet, which primarily consists of convolutional layers. We trained the EfficientNet-B3 model using encoded ingredients as target variables and experimented with two encoding methods: one-hot encoding, which represents each ingredient as a binary vector of length equal to the total number of ingredients, with a 1 indicating the presence of a specific ingredient and 0s for all others, and Word2Vec encoding, which embeds ingredients into a latent vector space that captures semantic relationships between words. Our evaluation showed that one-hot encoding outperformed Word2Vec on the test set, achieving 98.5 % accuracy. Therefore, we chose one-hot encoding as our final approach. A possible explanation is that the embedded ingredient space may not have been well-suited to this specific problem. One-hot encoding introduces a binary multi-classification problem, whereas Word2Vec embeddings produce continuous vectors that require additional learning. Additionally, one-hot encoding provided another advantage: it is easily integrated with the next submodel in our pipeline, as the second submodel also performed best when trained on one-hot encoded ingredient representations. The second submodel is based on ResNet18, which, despite being relatively basic, outperformed InceptionV3 and was not far behind the more advanced ResNet50 with attention mechanisms. We experimented with three different ingredient embedding

methods: one-hot encoding, index-based encoding where each ingredient is assigned a unique integer ID, and Word2Vec embeddings. Our evaluation showed that one-hot encoding outperformed the other embedding methods, so we selected it as the final approach for representing input ingredients. In total, the third model consists of two submodels. The first submodel, EfficientNet-B3, takes an image as input and outputs a one-hot encoded ingredient representation, while the second submodel takes both the image and the encoded ingredients as input and predicts the total calorie count. The plot below represents the third model architecture.

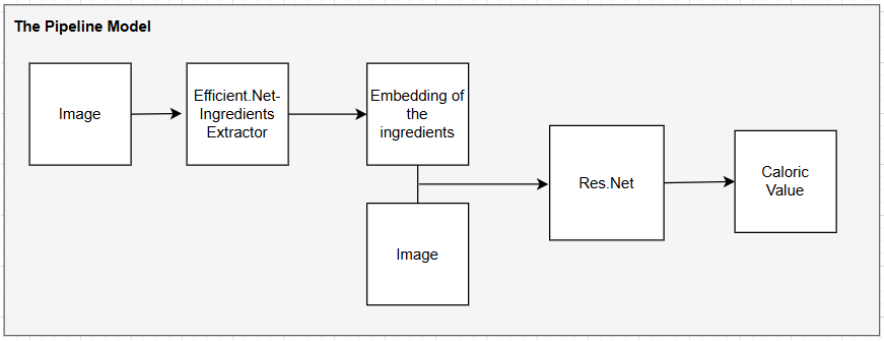


Figure 4: The Four Stages Model: From Vision to Language to Caloric Value

4 Results

In what follows, we present results from two models. One is inception V3, a more advanced version of Inception V2, with input being RGB images. We view inception as a strong neural network, which is a benchmark for models that have access only to images (no text). The second model we use is an image-text fusion model designed specifically for calorie estimation, We call this model pipeline, and it has four stages:

Image-Text Fusion (Pipeline Model): This model uses a four-stage process:

1. **Image:** Extract features using ResNet18.
2. **Ingredient List:** Convert ingredients to indices, obtain embeddings, and average them.
3. **Fusion:** Concatenate image and ingredient features.
4. **Regression:** Predict calories with a small multilayer perceptron (MLP).

Our experiments compare models using metrics such as mean absolute error (MAE) and mean absolute percentage error (MAPE). The pipeline model (image-text fusion) significantly outperforms the pure image-based Inception model in MAPE (49% vs. 117%) and shows competitive performance in MAE. Although our pipeline model's MAPE (48.69%) does not match the Nutrition5k benchmark (26.1%), it achieves superior MAE performance. These differences are largely attributed to variations in dataset size, preprocessing, and computational constraints.

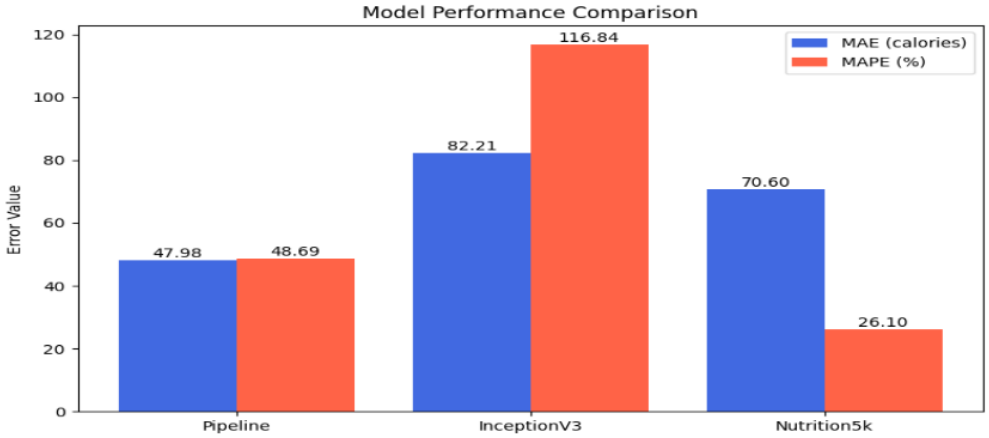


Figure 5: Bar plot of two models, comparable in running time and complexity. Pipeline model, receives image of a dish, identifies the ingredients and then uses an image-text fusion model, Inception receives images and tries to predict calories directly from the image. For each model we print the mean absolute error (MAE) measured in calories, as well as the mean absolute percentage error (MAPE). Evaluations are run on test set.

Pipeline model outperform Inception by quite a bit (49% vs 117% on the MAPE metric), showcasing a huge benefit of using image-text fusion model.Our pipeline model accuracy is 48.69% (MAPE), which is worse than the one achieved in Nutrition5k at 26.1%, however it performs better on the MAE metric. One reason why Nutrition5k outperform our model on the MAPE metric is that the Nutrition5k model is very complex, utilizes a huge and complex dataset (over 200 GB), compared with the data we used which consists of images only and weighs about 2GB. In addition, the google team has put a lot of effort in preprocessing. For example, they crop the images to exclude non-food elements, ensuring that the dataset primarily contains the food items themselves, which the authors claim helped them facilitate more accurate nutritional content prediction. Another reason we could not improve on the Nutrition5k model is that its complexity is much larger than what we can run on our computers.

To elaborate on the last point, we examined training a larger model, the Resnet with attention (which has about 100 times more degrees of freedom compared with inception v4). We could not do the same with an image- text fusion model, as the additional text increases the complexity of the model making it not feasible to run on our computers.

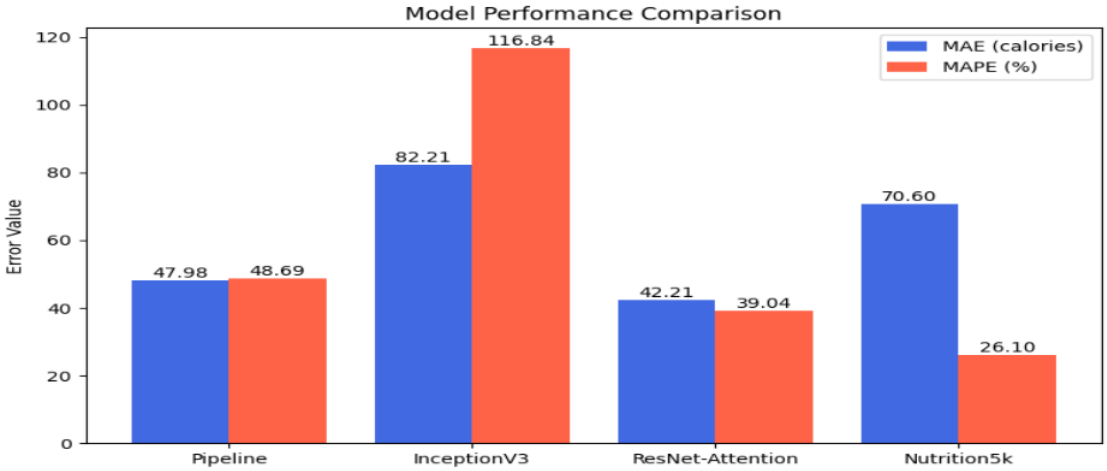


Figure 6: Bar plot of all models, showcasing huge improvement for a more complicated neural network that includes attention mechanism

5 Summary

In this study, we explored the problem of caloric prediction from RGB images using deep learning models, focusing on improving accuracy through ingredient-based multimodal learning. Unlike the Nutrition5k model, which leverages videos, depth images, and additional metadata, rendering its application very restrictive, our approach aims to achieve accurate predictions with RGB images alone, making it computationally more feasible. We have demonstrated that incorporating ingredient information significantly enhances caloric estimation performance, with our image-text fusion model (Pipeline model) outperforming purely image-based models like InceptionV3 and out performing Nutrition5k model on the MAE metric.. One key aspect of our methodology was ingredient encoding. Initially, we used one-hot encoding due to its high classification accuracy. However, alternative methods such as using RNNs or Transformers for ingredient representation could potentially capture complex interactions between ingredients more effectively, leading to further improvements in caloric estimation. Additionally, integrating attention mechanisms within our fusion model could refine feature selection, ensuring that critical aspects of both image and text inputs receive higher weights. While our approach demonstrated promising results, our model's accuracy still falls short of the Nutrition5k benchmark, primarily due to dataset constraints and computational limitations of our computers. Nonetheless, our work underscores the potential of multimodal learning for practical nutrition analysis, offering a scalable solution for real-world applications where only RGB images are available. Future work should focus on refining ingredient embeddings and incorporating advanced architectures such as transformers to further improve performance.

6 Additional Information

Table 1 in the Nutrition5k paper reports the MAE and MAPE for caloric prediction. The Nutrition5k model, utilizing over 200 GB of multi-modal data, outperforms our image-only model (which uses approximately 2 GB of data). Preprocessing techniques such as cropping non-food elements were critical in the Nutrition5k approach and highlight the challenges of direct model comparisons.

	Calorie MAE
Baseline	150.8 / 60.2%
2D Portion Independent Model	24.1 / 9.5%
2D Direct Prediction	70.6 / 26.1%
Depth as 4th Channel	47.6 / 18.8%
Volume Scalar	41.3 / 16.5%

Figure 7: The results obtained in Nutrition5k paper. Mean absolute error (MAE) and mean absolute error as a percent of the respective mean for that field. Caloric MAE is measured in calories. It shows the results of the Nutrition 5k models.

article hyperref

7 References

- Thames, Q., Karpur, A., Norris, W., Weyand, T., Xia, F., Sim, J., & Panait, L. (2021). *Nutrition5k: Towards Automatic Nutritional Understanding of Generic Food.* Retrieved from

  https://arxiv.org/pdf/2103.03375

- Brital, A. (n.d.). *Inception-v4 CNN Architecture Explained.* Retrieved from

  https://medium.com/@AnasBrital98/inception-v4-cnn-architecture-explained-23a7fe12c727

- Tan, M., & Le, Q. V. (2019). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.* Retrieved from

  https://arxiv.org/abs/1905.11946

- GeeksforGeeks. (n.d.). *EfficientNet Architecture Overview*. Retrieved from

  https://www.geeksforgeeks.org/efficientnet-architecture/

- Encord. (n.d.). *Vision-Language Models Guide*. Retrieved from

  https://encord.com/blog/vision-language-models-guide/

- Tayyebi, A.Z. (n.d.). *Bridging Vision and Language: Exploring CLIP, BLIP, and OWL-ViT*. Retrieved from

  https://medium.com/@az.tayyebi/bridging-vision-and-language-exploring-clip-blip-and-owl-vit-f8f6a99a263e