LEVERAGING DEEP LEARNING MODELS FOR NUTRITION CAPTIONING OF A DISH

Stella Dong, Sofia Marin-Quiros, Zixiang Qu

Computer Science, Columbia University

1. INTRODUCTION AND MOTIVATING WORK

In the recent years, image captioning models utilizing deep learning have reached impressive results, especially as LLM have improved. This groundbreaking technology can be applied to several fields including healthcare and nutrition. Sometimes, when people cook a plate with many different ingredients or buy a dish at a cafeteria, the nutritional information is not easily accessible. We wanted to design a model which could give users an overview of their meals in a sentence. Our idea was to include the top ingredient of the meal as well as information about the macronutrients. In order to generate such a model, we used state-of-the-art pretrained image captioning models.

We looked at 4 different models including CLIP Multimodal Learning Model, ALIGN Multimodal Model, ResNet50-T5 Vision-Text Model, and ViT-GPT2-image-captioning Model. After careful testing and analysis, this project shows that the encoder-decoder models outperform all of the other models at this task. This project discusses the strengths and failures of each of the models as well as future work to improve them.

The structure of this report follows. The Data Selection and Preprocessing section describes the Nutrition 5K dataset that was used. It also discusses the process of utilizing this dataset to create the image captions for training and testing. The Model Selection and Training section has 4 subsections, each describing the pretrained model architectures, the training process, the results, as well as strengths and weaknesses. The conclusion summarizes these findings and proposes two options for final models. Author contribution and references section can also be refrenced at the end.

1.1. Data Selection and Preprocessing

Nutrition5k dataset was selected from Github [1]. Nutrition5k is a dataset of visual and nutritional data for $\sim 5k$ realistic plates of food captured from Google cafeterias using a custom scanning rig. To tailored the dataset to our task, 1000 overhead RGB-D food plate images and their corresponding nutrition information was extracted. This includes per-ingredient mass; total dish mass and calories; fat, protein, and carbohydrate macronutrient masses. To generate image-caption pairs from the dataset, a data pre-processing pipeline was created to extract ingredients with the highest mass value, and top two nutrients with overall highest mass value. The information was then patched to a pre-designed sentence structure: "This meal contains a lot of {ingredient} and is high in {nutrient1} and {nutrient2}." The dataset was cleaned by removing ingredients listed as "deprecated" and ingredient information was matched with food images. The data was split into train, validation, and test in $8:1:1$ proportion. The dataset contains 117 unique ingredients and 3 macronutrients: fat, protein, and carbs. Below are some example images from the dataset (Fig. 1). (include images) Fig. 2 shows the mean, standard deviation, and average derivation from mean for each metric in Nutrition5k [2].

Fig. 1: Example of RGB images from Nutrition 5k.

1.2. Model Selection and Training

1.2.1. CLIP Motimodal Learning Model

CLIP (Contrastive Language-Image Pre-Training) model was created by OpenAI in 2021 as a novel approach to image-

	Mean	Standard Deviation	Average Deviation from Mean
Calorie	255	220	136
Total Mass(g)	215	161	114
Fat(g)	12.7	13.5	6.93
Carbs(g)	19.4	21.6	10.3
Protein(g)	18.0	20.0	10.7

Fig. 2: Mean, standard deviation, and average deviation from mean for each metric in Nutrition5k

captioning tasks. Instead of predicting images based on a fixed set of predetermined object categories limiting its scalability and geeneralizibility, CLIP trained image classifiers with natural language supervision at large scale. The model was trained on 400 milion (image, text) pairs. It is instructed in natural langauge to predict the most relevant text snippet, given an image, without directly optimizing for the task, similarly to the zero-shot capabilities of GPT-2 and 3 (Fig. 3).

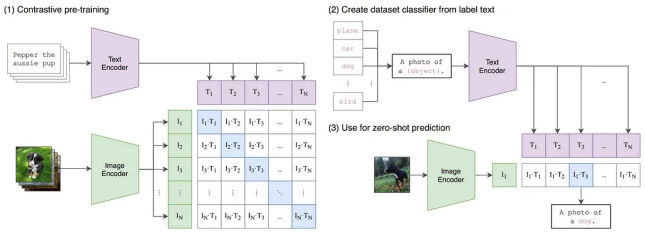
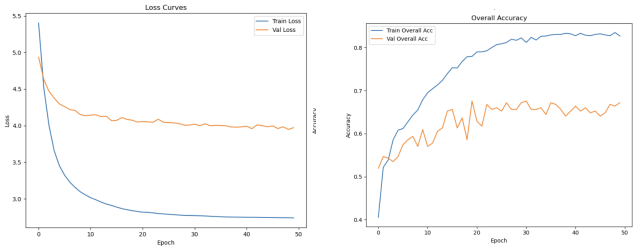


Fig. 3: The architecture of CLIP[3]

CLIP is a multi-modal vision and language model. It uses a ViT like transformer to extract visual features and a causal languagage model to extract text features. Both features are then projected to a latent space with identical dimension. Then the dot product between the projected image and text features is used as a similarity score. The model optimizes image-text pairs with the highest similarity score, thus generate the most fitting caption of a given image.

To apply the model to the nutrient description task, image-description pairings from the pre-processed dataset is passed into the model to create image features and text features. We use the names of all unique ingredients and macronutrients and generate all possible combination of sentences as the image, text pairing comparison. Cosine similarity scores are calculated and CrossEntropyLoss is used for the training. The model was trained with 50 epochs and an early stopping patience of 10 with a $1e-5$ learning rate. The model reached to a 69% validation accuracy and 3.9 validation loss (Fig. 4a,b).



(a) Validation Loss Curve (b) Validation Accuracy Curve

Fig. 4: Validation Loss and Accuracy Curve trained on CLIP model for 50 epochs

Below are some example image-description pairs and the predicted output based on similarity scores (Fig. 5):

From the output it's evident that the generated similarity scores are generally very close to each other, and the error in predictions mainly lies in nutrient output, and the granularity of ingredient names. The model achieved a test accuracy of 55%, with prediction in ingredient accuracy of 60% and nutrient accuracy 51%, separately. The model was then fine-tuned by adding weight decay and a cosine learning rate scheduler with warm-up. The model reached an overall accuracy of 58%, with ingredient accuracy of 64% and nutrient accuracy of 51%. The slightly higher accuracy in ingredient prediction can be explained by the fact that there are two possible outputs for nutrient matching, which makes an exact matching even more challenging. Top 5 ingredient mistakes is printed to check if the model is able to pick up slight variations in ingredient names (Fig. 6).

```
Top 5 Ingredient Mistakes:
Actual: fish, Predicted: grilled chicken, Count: 2
Error Rate: 2.00%
Actual: sweet potato, Predicted: roasted potatoes, Count: 2
Error Rate: 2.00%
Actual: chicken, Predicted: pork, Count: 1
Error Rate: 1.00%
Actual: chicken breast, Predicted: salmon, Count: 1
Error Rate: 1.00%
Actual: mixed greens, Predicted: kale, Count: 1
Error Rate: 1.00%
```

Fig. 6: Top 5 Ingredient Prediction Mistakes Generated by CLIP

The mistakes in ingredient predictions are due to the granularity of the ingredient names, such as "sweet potato" vs. "roasted potato", "mixed greens" vs. "kale". This can be due to the nature of the model's poor performance on fine-grained classification tasks [3]. This shows that CLIP lacks the capability to discern images with complex contents and highly

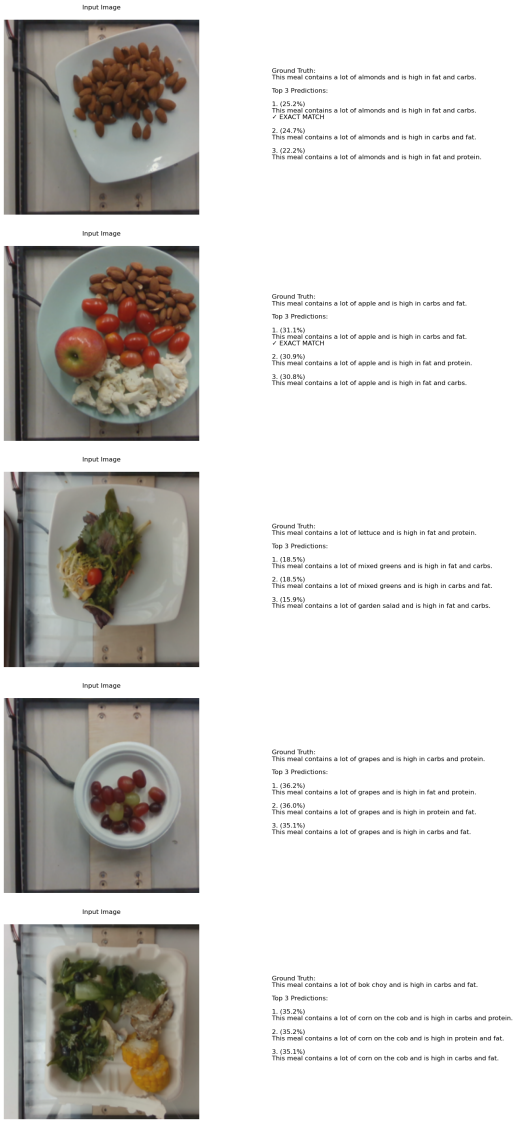


Fig. 5: Example predictions

similar texts. In addition, recent developments in deep learning models targeted at image-captioning tasks mainly uses models with an encoder-decoder structure [4], which adds a generative capability for text generation. In the following part of the report, we can also see how models combined with decoder is able to perform the task much better than a simple CLIP model. For future works, we can keep exploring ways to implement a CLIP-decoder model that pairs CLIP's image-text encoding ability with a generative text additional layer.

1.2.2. ALIGN Multimodal Model

The ALIGN model was initially used for the nutrient and ingredient description task. ALIGN is a state-of-the-art multimodal model that pairs an EfficientNet-L2 image encoder with a BERT text encoder, projecting both of them into a shared embedding space. It is trained with a contrastive loss to maximize the similarity between aligned image-text pairs while minimizing the similarity of misaligned pairs [5].

For this task, ALIGN was adapted to process the preprocessed dataset, where each image was paired with its corresponding caption. Training aimed to align visual features with textual descriptions. The training and validation loss and accuracy curves, shown in Figure 7, reveal the steady improvement of the model's performance during training. However, the ALIGN model struggled to perform well on the test set, with a test accuracy of 59.42% and a test loss of 0.0600.

One of the main problems was that ALIGN had difficulty generating text descriptions from the shared embedding space. It also overfit to the small Nutrition5k dataset. As seen in Fig. 7, the training and validation curves for "Loss Over Epochs" and "Accuracy Over Epochs" demonstrate this issue. The training accuracy improved over epochs, but the validation accuracy started to lagged significantly at around 59%. This disparity indicates that the model failed to generalize effectively to unseen data, likely overfitting to the small Nutrition5k dataset. Furthermore, the validation loss plateaued at a relatively high value compared to the training loss after a few epochs, reinforcing the observation of overfitting.

On the test set, the ALIGN model achieved a test accuracy of 59.42% with a test loss of 0.0600 after 10 epochs, highlighting its limitations in capturing detailed information required for precise ingredient and nutrient description. The model often made mistakes when distinguishing small differences in the dataset, such as similar ingredients or slight changes in nutrient levels. These issues made it clear that a different model was needed to do this task more effectively.

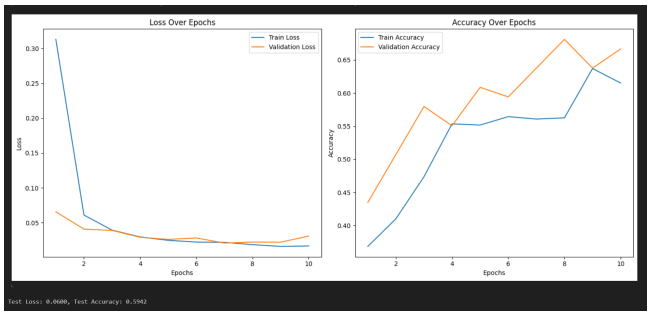


Fig. 7: Loss and Accuracy Curves for ALIGN Model. Test Loss: 0.0600, Test Accuracy: 0.5942 after 10 epochs.

These limitations underscored the need for a more tailored approach to address the challenges of the dataset and task. Consequently, I explored encoder-decoder structures in subsequent models to improve both the ingredient and nutrient prediction accuracy.

1.2.3. *ResNet50-T5 Vision-Text Model*

After encountering some difficulties by using ALIGN, the ResNet50-T5 Vision-Text Model was introduced to improve performance on the nutrient and ingredient description task. This model combines ResNet50(Residual Network), a neural network for extracting image features, with T5 (Text-to-Text Transfer Transformer), a generative language model for text creation, into an encoder-decoder architecture. This integration provides a more tailored approach to understanding both visual and textual modalities, solving the shortcomings of ALIGN.

ResNet50 acts as the visual encoder of the model, effectively extracting high-level features from food images. ResNet50 is renowned for its deep residual learning framework, which uses shortcut connections to address the vanishing gradient problem, enabling the model to learn increasingly complex features [6]. These extracted features are then projected into a feature space compatible with the T5 decoder. Figure 8 shows the structure of ResNet50 and demonstrates how its residual connections improve its ability to process visual data.

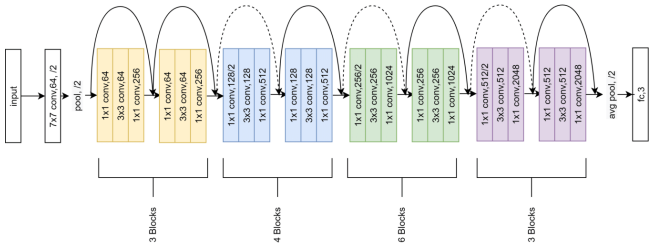


Fig. 8: The architecture of ResNet50, highlighting the use of residual connections to facilitate deep feature extraction [6].

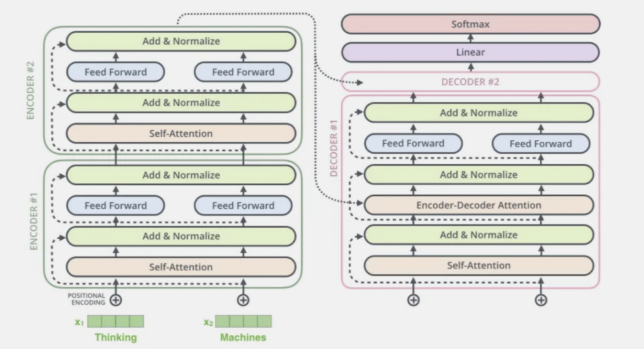


Fig. 9: The structure of T5, a transformer-based model used for text generation [7].

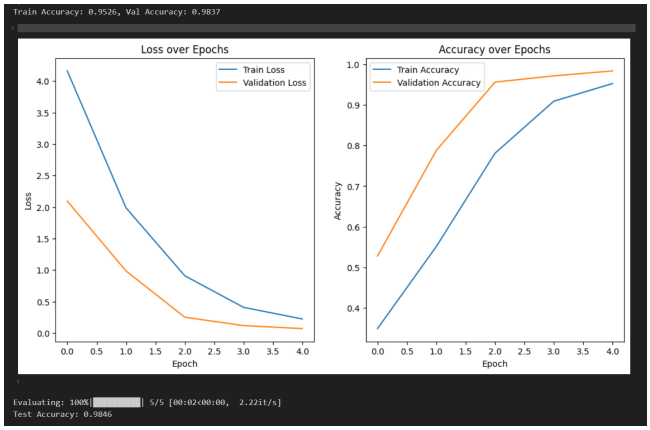


Fig. 10: Loss and Accuracy Curves for ResNet50-T5 Vision-Text Model.

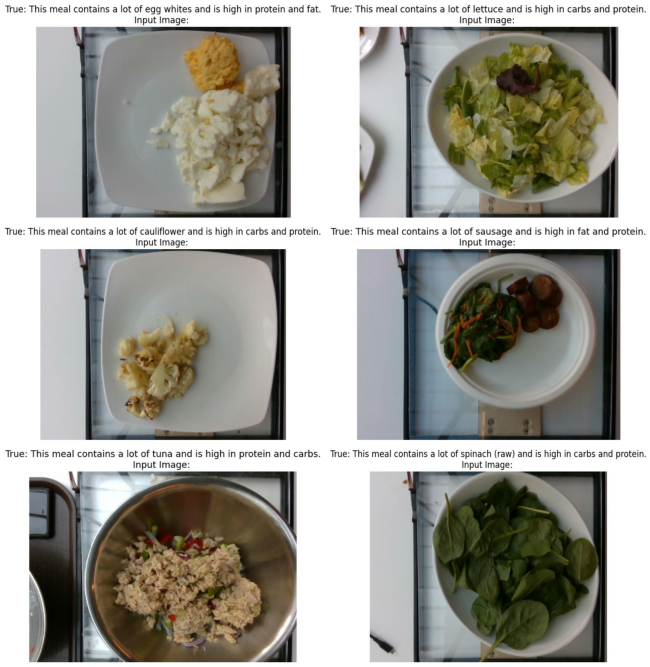


Fig. 11: Example Predictions Generated by ResNet50-T5 Vision-Text Model.

The T5 model functions as the text decoder, generating descriptive captions from the image features provided by ResNet50. T5 is a transformer-based model designed to handle diverse text-to-text tasks, such as summarization and translation [7]. Its generative capability enables the ResNet50-T5 model to produce accurate and proper descriptions. The combination of ResNet50 and T5 each uses their strengths respectively in vision and language understanding to achieve outstanding performance. Figure 9 illustrates the T5 model structure and its use of the transformer architecture for text generation.

Training and validation loss and accuracy curves, for this model, are shown in Figure 10, illustrate its superior performance. The model achieved a train accuracy of 95.26%, a validation accuracy of 98.37%, and a test accuracy of 98.46% after 5 epochs, reflecting its ability to generalize effectively across datasets.

Example predictions generated by this model further highlight its capability to produce detailed and accurate descriptions. Figure 11 showcases a selection of input meal images along with the corresponding predictions. These examples demonstrate the model's precision in identifying both the most abundant ingredients and the top two nutrients for each meal.

1.2.4. ViT-GPT2-image-captioning

The Vision Transformer (ViT)-GPT2 model, similar to the ResNet50-T5 Vision-Text Model, uses the encoder decoder architecture. It integrates ViT for image encoding with GPT-2 as the text decoder, creating image captions for all categories of images. This particular model was trained by the Common Objects in Context (COCO) dataset which contains over 120,000 images with descriptions. The pretrained model was obtained through Hugging Face and was trained by Ankur Kumar, a Senior Data Analyst at JP Morgan Chase [8]. ViT extracts high-level visual features, while GPT-2 leverages its language modeling capabilities to generate contextually accurate captions.

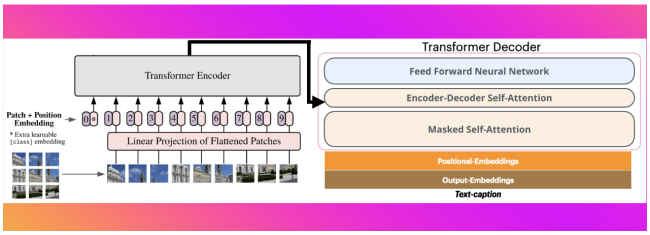


Fig. 12: The architecture of ViT-GPT2 [9]

The ViT model divides images into non-overlapping patches and represents each patch as a flattened sequence of pixel embeddings. The pixel value go through a linear projection layer that flattens these patches. These are processed through transformer layers to generate a sequence of visual embeddings. The self-attention mechanism allows for the complex relationships between different areas of the picture to be represented.[10] In the case of a food plate, there might be relationships between the different ingredients pictured on the plate that affect the result of the caption. This mechanism helps to achieve the high accuracy that the model has.

The visual embeddings represent the significant parts of the images of the food plates. These embeddings are then inputted into the GPT2 decoder. The decoder also receives the image captions with the nutrition information. However, these inputs are not the raw images, but rather, representations of these sentences through tokens. Tokenization is a common practice in Large Language Models in which words are assigned a unique number representation. GPT2 is an autoregressive decoder, meaning that it predicts tokens for the caption one at a time. It uses attention mechanisms to minimize redundant computations by focusing only on relevant context. This differentiates its accuracy from the CLIP model.

Through their unique processes, the ViT-GPT2 model is able to be faster and more memory efficient than other image captioning models. For example, through the design process, the BLIP (Bootstrapping Language-Image Pre-training) was tested but was unsuccessful due to issues with not enough memory. These factors are very important to consider since the aim of this project is to make this model usable by people outside the Machine Learning space that might have limited computational resources.

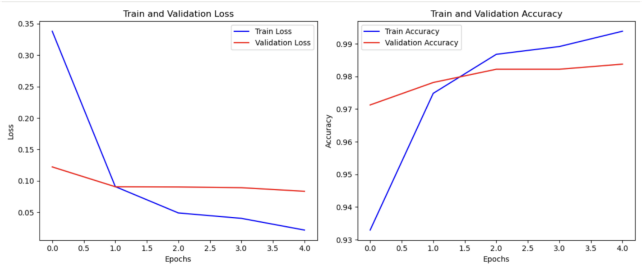
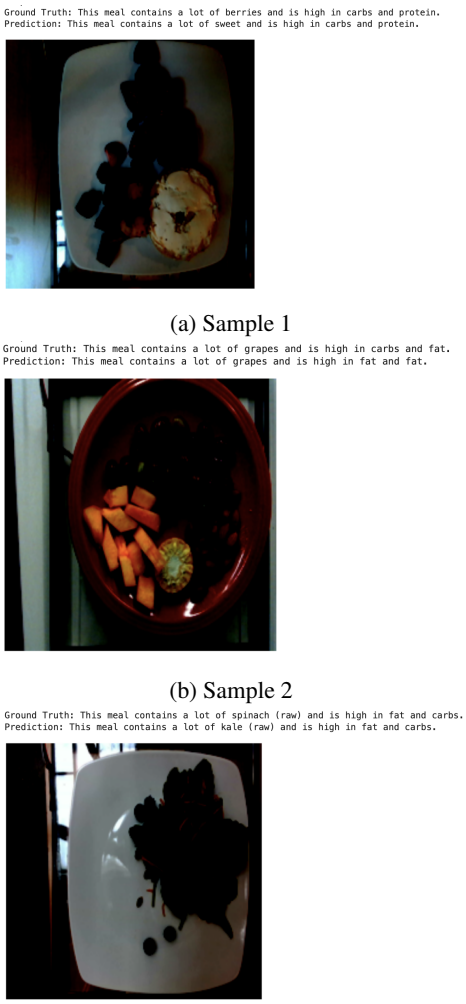


Fig. 13: Validation Loss and Accuracy Curve on ViT-GPT2 for 5 epochs

To fine tune the model, 5 epochs were trained using cross-entropy loss and the Adam Optimizer and a learning rate of 0.00005. In Figure 13, it can be clearly seen that the graph converges relatively quickly. More epochs were tested, but they did not show differences in the final accuracy of the model. In order to prevent overfitting, the model was left trained at 5 epochs. The accuracy on the testing set was 97.56%. A few samples of mislabeled data were examined and printed in Figure 14. The prediction and the ground truth are displayed.

As shown on Figure 14, most of the mislabeled images were images with poor lighting or plates with a colors other than white (as seen in the Sample 2). For these kinds of images, the model was able to build the correct sentence structure as well as get 2 out of the 3 components correct. If two ingredients looked similar (under low lighting) the model mixes them up. For example, something like raw spinach was mistaken for raw kale (Sample 3 image). These mistakes with low lighting or different color plates occur because there are not many like them in the training set. In order to fix this issue, in the future, we could create a dataset of low lighting images with different colored plates. If these images were then added to the training set, it would ensure these conditions were accurately represented in the training data, therefore increasing



(a) Sample 1

(b) Sample 2

(c) Sample 3

Fig. 14: Sample incorrect predictions

accuracy.

2. CONCLUSION

In this project, we explored several multimodal learning models to generate ingredient and nutrient descriptions from food plate images. The results clearly showed a difference between models that use text decoders and those that rely on similarity scoring.

The ALIGN model, although it is a cutting-edge multimodal approach designed to align visual and textual features, faced challenges with generalization. By depending on a shared embedding space and similarity scoring, it achieved a test accuracy of 59.42% with 10 epochs, which was not satisfactory. Similarly, the CLIP model, which also uses similarity scores to match images and text in a shared space, had limited success with an overall accuracy of 58% with 50 epochs. While CLIP is known for its versatility and scalability in zero-shot tasks, its focus on similarity scores made it less effective at creating accurate and detailed captions for this task.

However, models with encoder-decoder structures performed much better. The ResNet50-T5 Vision-Text Model combined ResNet50's strong visual feature extraction with T5's ability to generate descriptive text. This combination led to a big improvement, with the model achieving a test accuracy of 98.46% with only 5 epochs. Different from ALIGN and CLIP, this model could generate detailed descriptions, helping it to better capture the subtle differences in ingredients and nutrients.

The ViT-GPT2 model further demonstrated the advantages of using text decoders. Its self-attention mechanisms and autoregressive decoding allowed it to handle complex relationships between visual and textual data effectively. This model achieved a test accuracy of 97.56% with only 5 epochs, showing strong performance while also optimizing memory usage. Its efficiency makes it a practical solution for users with limited computational resources. In contrast, the BLIP model faced memory-related issues, which showed the importance of building models tailored to specific tasks.

In summary, our project revealed that models with text decoders, such as ResNet50-T5 and ViT-GPT2, are well-suited for generating detailed descriptions, while similarity-based models like ALIGN and CLIP struggle with accuracy and generalization. These findings also highlight how choosing the right model architecture can make a significant difference in multimodal tasks.

3. AUTHOR CONTRIBUTION

Each group member researched and trained their own selected model. 3 jupyter notebooks will be submitted for each of their own work. The author contribution for the final report is as follow:

Introduction: Sofia Marin-Quiros
Data Selection and Preprocessing: Stella Dong
CLIP Motimodal Learning Model: Stella Dong
ALIGN Multimodal Model and ResNet50-T5 Vision-Text Model: Zixiang Qu
ViT-GPT2-image-captioning: Sofia Marin-Quiros
Conclusion: Zixiang Qu

4. REFERENCES

[1] Google Research, Perception Labs. *Nutrition5k: A Comprehensive Nutrition Dataset*, 2021. Available at: `https://github.com/google-research-datasets/Nutrition5k`, accessed on December 6, 2024.

[2] Quin Thames et al. *Nutrition5k: Towards Automatic Nutritional Understanding of Generic Food.* 2021.

[3] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021.

[4] Hamidreza Mahyar Taraneh Ghandi, Hamidreza Pourreza. *DEEP LEARNING APPROACHES ON IMAGE CAPTIONING: A REVIEW*. 2023.

[5] Hugging Face. *ALIGN: Contrastive Learning of Image and Language Representations*. Available at: https://huggingface.co/docs/transformers/en/model_doc/align, accessed on December 2, 2024.

[6] Mostafa Ibrahim. *The Basics of ResNet50*. Available at: https://wandb.ai/mostafaibrahim17/ml-articles/reports/The-Basics-of-ResNet50---Vmlldzo2NDkwNDE2, accessed on December 2, 2024.

[7] Analytics Vidhya. *T5: A Detailed Explanation*. Available at: https://medium.com/analytics-vidhya/t5-a-detailed-explanation-a0ac9bc53e51, accessed on December 2, 2024.

[8] Ankur Kumar, Hugging Face. *nlpconnect/vit-gpt2-image-captioning*. Available at: https://huggingface.co/nlpconnect/vit-gpt2-image-captioning, accessed on December 2, 2024.

[9] Ankur Kumar. *The Illustrated Image Captioning using transformers*. Available at: https://ankur3107.github.io/blogs/the-illustrated-image-captioning-using-transformers/, accessed on December 2, 2024.

[10] Indrani Vasireddy et al. *Transformative Fusion: Vision Transformers and GPT-2 Unleashing New Frontiers in Image Captioning within Image Processing Authors*. 2023.