Visual Instruction Tuning Enhanced Recipe Generation

Daoting Chen, Tianxu Jiang, Wensong Hu, Yichen Lu

EECS 545 Machine Learning

MICHIGAN ENGINEERING
UNIVERSITY OF MICHIGAN

INTRODUCTION

Motivation

Models with multi-modality that including the Large Language Models (LLMs) have increasingly penetrated various fields, since it is capable of generating really reasonable outputs in a lot of scenarios. However, the application of these models remains to be developed, especially in domains that requires detailed and accurate procedures, notably the culinary domain. Large Language and Vision Assistant (LLaVA) is an extensively trained multi-modal model that has demonstrated excellent performances in many multimodal communicative tasks.

We perform visual instruction tuning on LLaVA to enhance its ability to generate coherent and applicable cooking instructions from image-text pairs.

DATASET

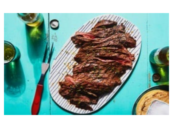
Recipe1M+

Recipe1M+ is an extensive dataset collected by Amaia Salvador et. al. It is a large-scale, structured corpus of over a million recipes and 13 million food images. Originally proposed for a cross-modal retrieval task, it can be easily adapted to visual instruction tuning since the ingredients are written clearly and the recipes are recorded accurately and elaborately.

Food Ingredients and Recipes with Images

Due to the large RAM demand for training or fine-tuning LLaVA on Recipe1M+, we started visual instruction tuning on a much smaller dataset: Food Ingredients and Recipes with Images. Similar to the structure of Recipe1M+, the dataset contains 13500 recipes together with the corresponding ingredients and the food images. We split the dataset to 3 subsets (pretraining: 70%, fine-tuning:20%, evaluation: 10%) and transform the data into conversation data with image-text pairs.

{
"Title": ...,
"Ingredients": [...],
"Instructions": ... ,
"Image_Name": ...,
}

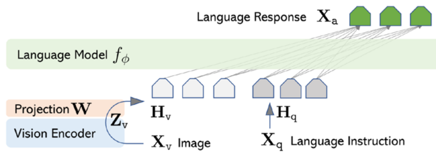


User: <image>\nHow can I cook it?
Model: <Ingredients>\n<Instructions>

METHODOLOGY

Baseline

The baseline model is the original LLaVA-v1.5-7B with Clip ViT as the vision encoder and Vicuna v1.5 as the base LLM.



Visual Instruction Tuning

The entire visual instruction tuning procedure can be boiled down into 2 stages:

1. **Pre-training for feature alignment**: perform Concept Alignment (if possible) and only adjust the projection layer

2. **End-to-end LoRA fine-tuning**: Both the projection matrix and LLM are updated

We optimize LLaVA's performance through this 2 stage tuning procedure which is adapted from the original visual instruction tuning method proposed in the LLaVA paper. The two stages are performed on the corresponding subset of Food Ingredients and Recipes with Images split in advance.

Evaluation

We utilize Gemini API to simulate human evaluation of the output of models. Since Recipe1M+ is no generated by LLM, we can utilize the recipes from Recipe1M+ as ground truth. We carefully design our prompts to obtain the evaluation from Gemini in the form of numerical scores. The average scores can be used as an evaluation metric of the models.

Question: What are the instructions to replicate the dish shown in the picture?

LLaVA

Preheat broiler. Toss plums with sugar, vanilla...

User: You are a helpful and precise assistant for checking the quality of the answer. Please tell me the score from 0 to 100. The true image is <image>. The true answer is <ground truth>. The model output is <model output>. Please only return the score from 0 to 100 and do not add any other words or characters.
Gemini: 75

EXPERIMENT

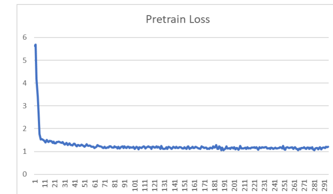
Experiment Results

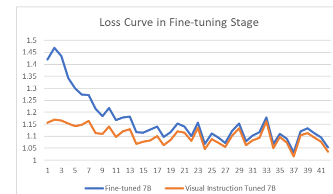
The experiment involves 3 models: original LLaVA (baseline model), fine-tuned LLaVA (no pre-training) and LLaVA after visual instruction tuning.

Pre-training reduced the final training loss on the LLaVA 7B model by more than 5%, which indicates the essence of both two stages of visual instruction tuning. However, the decrease is not significant on the LLaVA 13B model.

From the curves of the loss function in fine-tuning stage of the fine-tuned model and the model after visual instruction tuning, we can see that fine-tuning makes a difference, since it helps decrease the loss for both models, especially for the model without being pre-trained.

In the example below, we gave the results generated by both visual instruction tuned LLaVA and the original LLaVA. Compared with the original LLaVA, the model after visual instruction fine-tuning can better recognize the ingredients needed for the dish in the picture. In addition, there are fewer redundant contents and fewer incorrect or irrelevant instructions in results generated by LLaVA after visual instruction fine-tuning.





User: What are the instructions to replicate the dish shown in the picture?

LLaVA 13B (Visual Instruction Tuning): Preheat oven to 350\u00b0F. Butter 9-inch-diameter springform pan.\nWhisk together flour, baking powder, and salt in medium bowl. Using electric mixer, beat butter and sugar in large bowl until light and fluffy. Add eggs 1 at a time, beating well after each addition. Beat in vanilla. Add flour mixture and milk alternately to butter mixture, beginning and ending with flour mixture, and beat until just combined.\nSpoon batter into prepared pan. Bake cake until tester inserted into center comes out clean, about 45 minutes. Cool cake in pan on rack 10 minutes. Remove side of pan. Cool cake completely.\nPlace cake on platter. Spoon 1/4 cup raspberry sauce over top of cake. Top with 1/4 cup whipped cream. Repeat with remaining raspberry sauce and whipped cream. Sprinkle with confectioners sugar.

LLaVA 13B: To replicate the dish shown in the picture, follow these steps:\n\n1. Gather the necessary ingredients: You will need a bowl, a spoon, and a cup.\n2. Prepare the ingredients: You will need a base for the soup, such as vegetable or chicken broth, and some additional ingredients like herbs, spices, and seasonings.\n3. Cook the ingredients: Heat the broth in a pot on the stove, and add the additional ingredients like herbs, spices, and seasonings. Cook the ingredients until they are well combined and heated through.\n4. Pour the soup into the bowl: Carefully pour the cooked soup into the bowl.\n5. Serve the soup: Use the spoon to serve the soup into individual bowls, and enjoy the warm and flavorful dish.\n\nRemember that the specific ingredients and cooking methods may vary depending on the type of soup you are making. The image shows a bowl of soup with a spoon and a cup, but the exact ingredients and cooking process will depend on the recipe you choose to follow."

DISCUSSION

Current Issues

The dataset for visual instruction tuning is small compared with the Recipe1M+ dataset. And the evaluation method relies on Gemini API, which may be affected by the instability of the results generated by API.

Future Directions

1. Perform visual instruction tuning on a larger dataset

2. Carry out evaluation on a larger scale and perform human evaluation to obtain more accurate results

3. Perform data augmentation on images