Visual Question Answering for Peruvian Cuisine in Regional Spanish

Mariana Risco Cosavalente

Universidad Nacional de Trujillo
mrisco@unitru.edu.pe

Abstract

This project leverages Visual Question Answering (VQA) to promote Peruvian gastronomy by utilizing a culturally rich dataset and advanced models such as LLaVA-1.5 and GPT-2 Large. The evaluation will comprise both automated metrics and culinary expert assessments. This system addresses regional variations in dish names, promotes inclusivity by involving Peruvians from diverse regions in dataset construction, and enhances cultural representation.

Introduction

Visual Question Answering (VQA) represents an intriguing area of artificial intelligence (AI) that integrates computer vision with natural language understanding (Barra et al. 2021). It involves formulating questions about an image and providing corresponding answers. In Figure 1, the left side shows an input image of a Peruvian dish (likely "lomo saltado") with a question in Spanish. The VQA model processes this input, and the output appears on the right.

I am interested in studying the application of Visual Question Answering (VQA) in the context of Peruvian culture and gastronomy. In particular, I will adapt this technology to Spanish, which is the native language of over 500 million people worldwide. The majority of VQA systems and datasets have been developed in English (Romero et al. 2024), which makes it difficult to apply them in different contexts such as Peru.

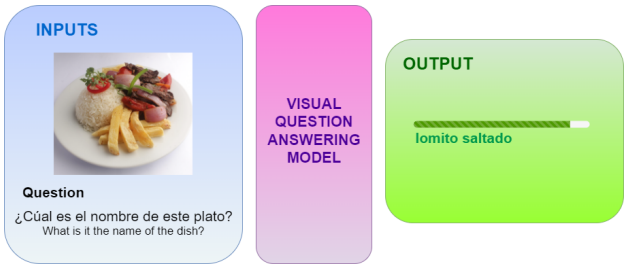

Figure 1: Example of VQA system for Peruvian cuisine in regional Spanish to know the dish's name.

Peru's cultural and linguistic diversity is reflected in everyday aspects, such as its food. A key challenge is collecting data that captures this richness, including regional language variations. This project aims to create a dataset that reflects the regional linguistic peculiarities, focusing on Peru's gastronomy, a top 10 global culinary destination (TasteAtlas 2024).

The goal is to develop a system where tourists can photograph Peruvian dishes and ask questions about their names, histories, and ingredients. This not only highlights Peru's gastronomic culture but also addresses regional naming differences, fostering inclusivity by involving Peruvians from all regions, especially those underrepresented.

Background

There are some approaches, one of them adapted VQA system to multiple languages, including Spanish in Latin American countries, by creating a dataset comprising images of food, sport, and everyday situations (Romero et al. 2024). Researchers compared several multimodal vision-language MLLMs, one of them had the best result which is LLaVA-1.5 (Liu et al. 2024), the accuracy achieved in food topic was 32.5% (Romero et al. 2024). Another related work is MarIA, a family of Spanish language models like RoBERTa and GPT-2-large, trained on 570GB of Spanish text, outperformed other models in tasks such as Question Answering (Gutiérrez-Fandiño et al. 2022). These studies are aligned with my project, where I aim to adapt vision-language models (VLLMs), such as LLaVA-1.5 (Liu et al. 2024) and GPT-2-large (Radford et al. 2019), which achieved good accuracy in food-related tasks, to Peruvian gastronomy.

In a previous work on plant pest detection, I collected images and employed data augmentation techniques to robustly enhance the dataset (Risco, Chang, and Cortegana 2024). Additionally, I conducted a Question Answering project in Spanish, utilizing pre-trained BERT models. I aim to integrate these prior experiences into my future work towards the promotion of Peruvian gastronomic culture.

Approach

In this study, I will employ state-of-the-art multimodal language models (VLLMs), namely LLaVA-1.5 (Liu et al. 2024) and GPT-2-large (Radford et al. 2019), which are proficient in vision-language tasks.

Dataset Construction and Annotation

The dataset will consist of 4,000 images of traditional Peruvian dishes from PeruFoodNet (Arzola 2024), featuring 40 dish categories. Data enrichment will leverage:

- **Social Networks:** Images and textual data will be obtained using the Selenium web scraping tool by querying hashtags such as #ceviche and #anticucho (Selenium 2024).
- **Google Maps:** The Google Maps API will extract reviews, ratings, and photos from popular regional restaurants, highlighting linguistic and cultural variations in dish descriptions.
- **Culinary Books:** Reference books will provide historical and cultural context for traditional Peruvian dishes (Guardia 2016).

Annotations will include dish labels, related questions, and their answers. Collaboration with linguistics students from Peru's Coast, Sierra, and Selva regions will ensure high-quality, context-specific annotations, addressing questions such as: "What is the name of this dish?", "What are its ingredients?", and "What is its origin?".

Data Augmentation and Preprocessing

Data augmentation techniques, including rotation, scaling, brightness adjustment, and noise injection, will enhance robustness to variations in image presentation, ensuring improved model generalization.

Named Entity Recognition (NER)

NER techniques will extract critical entities, including dish names, ingredients, and regional terms, from textual data. This ensures that the model delivers contextually relevant responses aligned with the linguistic and cultural nuances of Peruvian cuisine (Pakhale 2023).

Zero-Shot Prompting

Zero-shot prompting will be employed to enable the model to infer answers for dishes absent in the training data. This approach leverages the model's pre-trained understanding of general linguistic patterns to ensure adaptability across unseen dishes and regional culinary variations (Brown et al. 2020).

Evaluation

Evaluation Method

In order to evaluate the two fine-tuning models, L LLaVA-1.5 (Liu et al. 2024) and GPT-2 large (Radford et al. 2019), a zero-shot evaluation will be conducted. As the models have not been trained for this particular task, it will be necessary to adapt them to the context of this research project.

Evaluation Metric

The primary metrics for evaluating the responses during the training and testing phases will be accuracy, perplexity, and F1 score, which will help to assess the models' ability to provide accurate responses.

Human Evaluation

In addition to metrics, human evaluation by people with greater Peruvian gastronomic knowledge, such as senior citizens, and chefs from all locations in Peru, provides a robust evaluation of the project.

Discussion

In this project, I hope to enhance the performance of multimodal language models (VLLMs), such as LaVA-1.5 (Liu et al. 2024) and GPT-2 large (Radford et al. 2019), by aligning them with Peruvian Spanish through integration with a gastronomy-specific dataset. A comparative evaluation of the models will identify the best-performing approach for dish recognition and answering questions about preparation and ingredients.

The creation of a benchmark from this dataset will provide a valuable reference for researchers, advancing natural language processing and computer vision in culturally specific contexts. Engaging university students in data collection and annotation fosters research participation, with opportunities for academic contributions, particularly in linguistics, gastronomy, and artificial intelligence.

Additionally, the project encourages user contributions from diverse Peruvian regions, enriching collective knowledge of Peruvian cuisine. Its findings may enhance the experiences of local and foreign tourists, highlighting gastronomy as a key element of Peru's cultural identity.

Conclusion

This project addresses the challenge of adapting Visual Question Answering (VQA) technology to Peruvian Spanish in the context of gastronomy. By curating a diverse dataset from social networks, Google Maps, and culinary books, and fine-tuning state-of-the-art vision language models (VLLMs) such as LaVA-1.5 (Liu et al. 2024) and GPT-2-large (Radford et al. 2019), the objective is to enhance cultural accessibility in AI. Evaluation will involve both automated metrics and assessments from culinary experts. Ultimately, this project has the potential to create a valuable benchmark for VQA systems, promote Peru's cultural diversity, and enhance the tourism experience by providing tailored gastronomic information to locals and visitors alike.

Acknowledgments

I would like to express my gratitude to the AAAI UC team for their invaluable mentorship throughout this project. I am especially grateful to my mentor, Nils Murrugarra Llerena, a fellow Peruvian, whose guidance and insightful advice have been key in making this project more achievable.

References

Arzola, F. e. a. 2024. PeruFoodNet - A Traditional Peruvian Food Dataset.

Barra, S.; Bisogni, C.; De Marsico, M.; and Ricciardi, S. 2021. Visual question answering: Which investigated applications? *Pattern Recognition Letters*, 151: 325–331.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Guardia, S. B., ed. 2016. *Cocina peruana : Historia, cultura y sabores*. Edit.Fond. Usmp. ISBN 9786124221453.

Gutiérrez-Fandiño, A.; Armengol-Estapé, J.; Pàmies, M.; Llop-Palao, J.; et al. 2022. MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, 39–60.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved Baselines with Visual Instruction Tuning. arXiv:2310.03744.

Pakhale, K. 2023. Comprehensive Overview of Named Entity Recognition: Models, Domain-Specific Applications and Challenges. arXiv:2309.14084.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Risco, M.; Chang, S. J.; and Cortegana, C. A. 2024. Benchmarking CNN-Based Systems for Corn Leaf Pest Detection using Fine-Tuning. In *Latinx in AI @ NeurIPS 2024*.

Romero, D.; et al. 2024. CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Selenium. 2024. Selenium automates browsers. https://www.selenium.dev/. Accessed: 2024-12-17.

TasteAtlas. 2024. TasteAtlas Awards 23/24: These are the 100 Best Cuisines and Dishes of the World. https://www.tasteatlas.com/tasteatlas-awards-23-24.