GEN AI PRINCIPLES Course Project -II:

Multimodal Conversational AI for E-commerce: A Vision-Language Approach

Project Context

In the e-commerce domain, providing effective customer support is crucial to maintaining customer satisfaction and loyalty. Customers frequently seek detailed information about products, which can include textual descriptions, visual images, and user-generated content such as reviews and questions. Traditional customer support systems often rely solely on text, limiting their ability to provide comprehensive answers, especially when queries involve visual elements.

A multimodal conversational AI system can bridge this gap by integrating both language and vision capabilities. Such a system can interpret and respond to customer queries that include both text and images, enhancing the user experience by providing accurate, context-aware responses. This approach is particularly valuable for e-commerce platforms, where users may upload images of products and ask about features, specifications, or compatibility. By leveraging multimodal data, the system can reduce response times, improve accuracy, and decrease the workload on human customer support agents.

Scope of the Project

The project aims to develop a multimodal conversational chatbot capable of answering product-related questions using both text and images. The system will utilize the Amazon Product Dataset 2020, which includes product images and descriptions, and will integrate RAG patterns and Vision-Language models powered by LLMs and foundational Vision-Language models for effective multimodal interaction. The project is divided into the following components:

1. **Understanding Multimodal Data**
   - **Objective**: Analyze and preprocess multimodal data, including both images and text descriptions from the Amazon Product Dataset 2020.
   - **Tasks**:
     - Define the optimal combination of product attributes (such as title, brand, price, features, and images) for creating comprehensive product descriptions.
     - Ensure data consistency and quality to optimize the performance of the AI models in subsequent steps.

2. **Implementing Vision-Language Retrieval-Augmented Generation (RAG)**
   - **Objective**: Create a retrieval-augmented generation framework using state-of-the-art multimodal model.

- **Tasks**:
    - **Embedding Generation**: Utilize the CLIP model as a foundational multimodal encoder to generate embeddings for both text descriptions and images. CLIP’s ability to align visual and textual data in a shared embedding space makes it ideal for this purpose.
    - **Embedding Storage**: Store these embeddings in a vector database (e.g., Google Vertex AI Vector Search) to facilitate efficient retrieval of relevant data.
- **Outcome**: A robust retrieval system that provides accurate contextual information to a language model, enhancing its ability to generate relevant responses.

- **Accuracy of Retrieval**: Assess the correctness of the retrieved items.

- **Recall at Different Cutoffs**: Calculate Recall@1, Recall@5, and Recall@10 to evaluate the proportion of relevant items retrieved at each cutoff level.

3. **Integrating with a Large Language Model (LLM)**
    - **Objective**: Enable conversational interactions using an LLM.
    - **Tasks**:
        - Integrate the retrieval mechanism with an open-source LLM (e.g., Meta-Llama-3.1 or Mixtral) for dynamic question-and-answer capabilities.
        - Leverage your prior learning on crafting template prompts with zero-shot, few-shot, and multiple-shot techniques to generate relevant conversational interactions.
        - Ensure that the LLM can utilize the retrieved embeddings to generate accurate and context-aware responses.
    - **Outcome**: A conversational interface capable of understanding and responding to both text and image-based queries effectively.

4. **Developing a User Interface**
    - **Objective**: Build an intuitive user interface to provide a seamless chatbot experience.
    - **Tasks**:
        - Develop the interface using the Streamlit framework, allowing users to input queries via text or upload images.
        - Design the interface to display responses in a user-friendly format, showcasing the chatbot's multimodal capabilities.
    - **Outcome**: An engaging platform that demonstrates the full range of the chatbot's abilities.

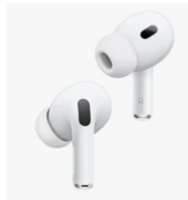
Example of Desired Interaction with the Multimodal Chatbot

- **Text-Based Questions**:
  1. **Question**: "What are the features of the Samsung Galaxy S21?"
     - **Answer**: "The Samsung Galaxy S21 comes with a 6.2-inch Dynamic AMOLED display, a triple-camera setup (12MP wide, 64MP telephoto, 12MP ultrawide), and a 4000mAh battery.
  2. **Question**: "Can you compare the Amazon Echo Dot with the Google Nest Mini?"
     - **Answer**: "The Amazon Echo Dot features Alexa voice assistant, a 1.6-inch speaker, and Bluetooth connectivity. The Google Nest Mini, on the other hand, comes with Google Assistant, a 40mm driver, and supports both Bluetooth and Wi-Fi. Both devices are designed for smart home control and music playback, but the choice depends on your preferred ecosystem (Amazon Alexa or Google Assistant)."
- **Image-Based Questions**:
  1. **Question**: "Can you identify the product in this image and describe its usage?"
     - **User uploads an image of a specific product**.

     - **Answer**: "This is a KitchenAid Artisan Stand Mixer. It is used for mixing, kneading, and whipping ingredients, making it ideal for baking and cooking tasks. It comes with multiple attachments for various culinary tasks, such as making pasta or grinding meat."
  2. **Question**: "What is the name of this product, and how do I use it?"
     - **User uploads an image of a specific product**.

     - **Answer**: "This is a Fitbit Charge 4 Fitness Tracker. It is used to monitor your heart rate, track steps, calories burned, and sleep patterns. It also has built-in GPS to track outdoor activities and workouts. To use it, wear it on your wrist, and sync it with the Fitbit app on your smartphone."
- **Request for a Specific Product Image**:
  1. **Question**: "Can you show me a picture of the Apple AirPods Pro?"

     - **Answer**: "Sure, here is an image of the Apple AirPods Pro:



The AirPods Pro features active noise cancellation, a customizable fit with silicone tips, and are sweat and water-resistant, making them ideal for workouts and daily use."

Data Section

The project will utilize the **Amazon Product Dataset 2020**, available on Kaggle, as the primary data source. This dataset includes a comprehensive collection of product images and textual descriptions, which are essential for developing the chatbot. The focus will be on defining the right combination of product attributes to form effective product descriptions and ensuring data consistency for optimal chatbot performance.

- **Dataset URL**: Amazon Product Dataset 2020

Expected Deliverables

- **A Fully Functional Multimodal Chatbot**: Capable of understanding and responding to product-related queries using both text and image inputs.
- **Detailed Documentation**: Covering preprocessing steps, model architectures, integration details, and system design.
- **User Interface**: A fully developed interface using Streamlit that allows users to interact with the chatbot and visualize its responses.
- **Research Report**: A comprehensive report detailing the implementation process, challenges faced, solutions developed, and suggestions for future improvements.
- **Evaluation Metrics and Results**: Metrics such as retrieval accuracy, response relevance, etc.

References

1. **CLIP (Contrastive Language–Image Pre-training)**: A model developed by OpenAI that learns visual concepts from natural language descriptions, used for encoding both text and images.
    - Radford, Alec, et al. "Learning Transferable Visual Models From Natural Language Supervision." *Proceedings of the 38th International Conference on Machine Learning (ICML)*. 2021. Link to Paper
2. **VLAVA (Vision-Language Alignment and Variance Adjustment)**: A model designed to enhance alignment between visual and textual modalities by adjusting variance differences.

- Liu, Z., Wang, Y., et al. "Vision-Language Alignment and Variance Adjustment." *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023. Link to Paper

3. **RAG Pattern (Retrieval-Augmented Generation)**: A framework that combines retrieval-based methods with generative models to enhance response quality.
   - Lewis, Patrick, et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." *Advances in Neural Information Processing Systems (NeurIPS)*. 2020. Link to Paper
4. **Vision-Language Foundation Models**: Models trained on large-scale multimodal datasets to support a variety of vision-language tasks.
   - Li, Xiangtai, et al. "Pre-trained Vision and Language Transformer for Multimodal Understanding and Generation." *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2022. Link to Paper