Beyond the Image: A Multimodal Approach to Automatic Text Extraction and Sentiment Analysis in Advertisements

Name: Ravi Teja Seera

Capstone Project (M.S Project 69500)

Advisor: Prof. Ming Jang

Table of Contents

Abstract:

This project explores the implementation of the LLaVA (Large Language and Vision Assistant) model for enhanced advertisement content analysis, a technology initially developed by Liu et al. (2023) for general-purpose visual and language understanding. The LLaVA model, which integrates advanced machine learning techniques for multimodal data processing, is adapted to automatically extract text and analyze sentiment from image-based advertisements. By leveraging the model's capabilities in following visual instructions and interpreting multimodal inputs, the project aims to uncover the underlying persuasive strategies used in advertisements, facilitating a deeper understanding of their impact on consumer behavior. The adaptation involves fine-tuning the pre-trained model on a dataset of image advertisements, focusing on the extraction of textual content and sentiment analysis, thus extending the model's application from theoretical frameworks to practical advertisement analysis. This approach not only demonstrates the flexibility and effectiveness of LLaVA in a specialized domain but also contributes to the ongoing evolution of multimodal analysis in the field of advertising.

Introduction:

Analyzing and interpreting the complex interactions between textual and visual material in advertisements is made possible by the combination of machine learning and advertisement analysis. The LLaVA (Large Language and Vision Assistant) model, which was developed by Liu et al. (2023) for general-purpose visual and language processing tasks, is the foundation of this project. The strong visual instruction tuning and multimodal data processing capabilities of LLaVA offer novel approaches to the analysis of advertising, which are intricate combinations intended to sway customer behavior.

Advertisements leverage both imagery and language to persuade and inform, making them ideal candidates for analysis using advanced multimodal learning models. Ads' ability to persuade consumers is dependent on a combination of explicit material and more subtly conveyed signs and emotions that affect their views. Through this project, LLaVA—which was originally created to improve zero-shot learning capabilities through instruction tuning—is being applied to the more complex domain of advertisement analysis. In order to gain a better understanding of how advertisements influence customer decision-making, it attempts to decipher the clever use of both verbal and visual aspects.

The relevance of applying such advanced machine learning tools in advertising is underscored by Hussain et al. (2017), whose work on automatic content analysis of advertisements sets a foundational precedent for this project. By integrating the multimodal, instruction-following capabilities of LLaVA, this study extends Hussain's methodologies to not only recognize but also interpret the complex interplay of elements within ads. This project, therefore, not only tests the efficacy of LLaVA in a new domain but also aims to contribute to the broader field of computational advertising by providing deeper insights into the effectiveness and mechanics of ad content.

Furthermore, biases in the field of artificial intelligence (AI) can affect perceptions and results, especially in applications like sentiment analysis. It is essential to identify and resolve these biases in order to create just and efficient AI systems. To ensure that the insights obtained from ad content are fair and accurate, special attention is paid in this project to locating and removing any potential biases in sentiment analysis algorithms. This entails improving the model's sensitivity to the various cultural and societal circumstances that affect consumer mood in addition to honing its capacity to read complicated and varied advertisement datasets. By addressing these issues, the project adds insightful viewpoints to the continuing conversation on moral AI applications, especially with regard to their responsible use in business environments.

Literature Review:

- “Automatic Understanding of Image and Video Advertisements”:

  The pioneering work by Hussain et al. (2017) marks a significant milestone in the field of advertisement analysis through the lens of computer vision and machine learning. Their research introduced one of the first comprehensive datasets specifically designed for the automatic understanding of advertisements. This dataset includes thousands of annotated image advertisements, categorized by the emotions they evoke and the intended messages they convey. The authors developed models that could predict the sentiment and thematic content of advertisements based on visual and textual cues.

  Hussain et al. (2017) researched several techniques for combining textual and picture analysis to forecast the content and efficacy of an advertisement. They showed that integrating textual and visual data features results in predictions that are more accurate than utilizing only one modality. This integrated strategy fits in with the larger AI research trend of creating models that can comprehend and analyze multimodal inputs.

  This work is significant because it applies deep learning techniques to a new domain, laying the groundwork for future research into the autonomous analysis of advertisements to comprehend consumer behavior and response. According to their findings, analyzing advertisements effectively necessitates having a sophisticated grasp of the explicit and implicit signals that are communicated through various media.

  Because it offers a conceptual and methodological foundation for utilizing innovative machine-learning techniques to analyze advertisements, this research is very important to the current undertaking. This study intends to extend the work of Hussain et al. by incorporating more recent advances in AI, specifically the LLaVA model, by utilizing comparable multimodal methodologies. The goal is to gain a greater understanding of the persuasive strategies used in commercials.

- ## Large Language and Vision Assistant (LLaVA):

The integration of language and vision models has advanced significantly in recent machine learning advances, opening new possibilities for multimodal data processing and comprehension. The LLaVA (Large Language and Vision Assistant) model created by Liu et al. (2023) is a noteworthy contribution in this field. By fusing complex vision encoders with the reliable processing powers of large language models (LLMs), this model constitutes a breakthrough in the field.

The architecture of LLaVA is designed to enhance the model's ability to follow and interpret multimodal instructions, which is crucial for tasks that involve complex interactions between visual and textual data. The model consists of two primary components:

Vision Encoder: This component is crucial for capturing and encoding visual data. The encoder is typically pre-trained on a vast array of images to capture a wide variety of visual features. It transforms raw images into a structured format that can be seamlessly integrated with textual data, allowing for a comprehensive understanding of visual content.

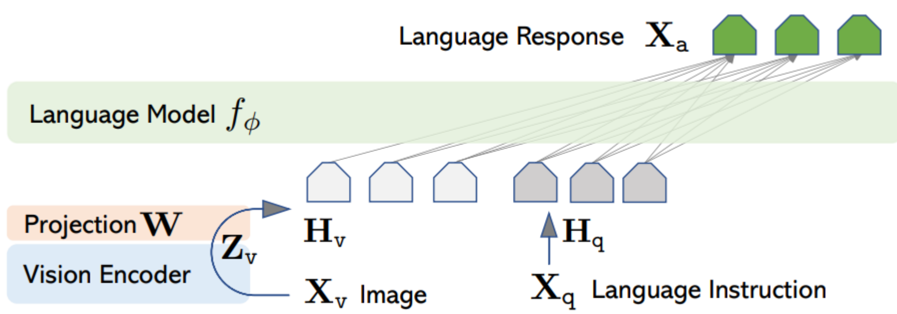


Figure 1Architecture of the LLaVA model, illustrating the integration of the vision encoder and the language model. Adapted from Liu et al. (2023).

Language Model: The second component is a sophisticated language model that processes the encoded visual data alongside textual inputs. This model is fine-tuned with instruction-following data, enabling it to generate linguistic outputs that are not only relevant but also contextually aligned with the visual inputs. The language model's training involves a technique known as instruction tuning, which enhances its ability to execute tasks in a zero-shot manner performing tasks without prior explicit example-based training on those specific tasks.

The integration of these two components allows LLaVA to perform a variety of complex tasks that require an understanding of both text and images. For example, it can generate accurate descriptions for images, answer questions about the content within an image, and engage in dialogue where visual understanding is necessary.

In the context of advertisement analysis, the LLaVA model's capabilities are particularly relevant. Advertisements often combine visual elements and text to influence and persuade viewers. By leveraging LLaVA's ability to interpret these multimodal elements, this project aims to dissect and analyze advertisements more effectively. The model's proficiency in understanding the interplay between images and accompanying text allows for a deeper analysis of how advertisements use these elements to affect consumer behavior.

Methodology:

This project utilizes the Large Language and Vision Assistant (LLaVA) model to analyze the content and sentiment of image-based advertisements. The methodology involves setting up a machine learning pipeline that leverages image-to-text capabilities to extract and interpret visual and textual elements from advertisements.

1. Setup and Configuration:

1.1 Environment Setup:

Python is the programming language used with several specialized libraries. Key installations include the transformers and bits and bytes libraries, essential for loading and optimizing the LLaVA model.

```
!pip install -q -U transformers==4.37.2
!pip install -q bitsandbytes==0.41.3 accelerate==0.25.0
```

1.2. Model Configuration:

The LLaVA model is configured to optimize performance without compromising on processing speed or accuracy, essential for real-time advertisement analysis. The BitsAndBytesConfig is set to enable 4-bit loading of model weights to reduce the memory footprint. This configuration allows the model to run on hardware with limited resources while maintaining a high level of precision needed for text extraction from complex ad images. The specific settings chosen reflect a balance between computational efficiency and the fidelity of the model's output, crucial for accurately interpreting advertisement content.

```
import torch
from transformers import BitsAndBytesConfig

quantization_config = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_compute_dtype=torch.float16
)
```

```
from transformers import pipeline

model_id = "llava-hf/llava-1.5-7b-hf"

pipe = pipeline("image-to-text", model=model_id, model_kwargs={"quantization_config": quantization_config})
```

2. Image Processing and Text Extraction

2.1 Data Loading and Image Processing:

Images of advertisements are loaded from a specified directory, and each image is processed individually. The pipeline uses a custom function, process_image, to convert images into textual descriptions using the LLaVA model.

```
# Function to process a single image file
def process_image(image_path):
    try:
        image = Image.open(image_path)
        outputs = pipe(image, prompt=prompt, generate_kwargs={"max_new_tokens": 200})
        return outputs[0]["generated_text"]
    except Exception as e:
        print(f"Error processing image {image_path}: {e}")
        return None
```

2.2 Interactive User Interface:

To facilitate the selection and display of images, an interactive dropdown menu is created using IPython widgets. Users can select an advertisement image from the menu, and the selected image is displayed alongside the model's textual interpretation.

```python
import os
from PIL import Image
from IPython.display import display
from ipywidgets import Dropdown, Button, Output
```

```python
# Create a dropdown menu for image selection
dropdown = Dropdown(
    options=[(filename, filename) for filename in sorted(images.keys())],
    description='Select Image:',
    disabled=False,
)

# Display area for the selected image
output_area = Output()

# Button to display the image
button = Button(description="Display Image")

# Global variable to store the reference image
ref_image = None

# Function to update the output area with the selected image
def on_button_clicked(b):
    global ref_image
    ref_image = images[dropdown.value]
    with output_area:
        output_area.clear_output()
        display(ref_image)
        print(f"Selected image: {dropdown.value}")

button.on_click(on_button_clicked)

# Display the widget and output
display(dropdown, button, output_area)
```

3. Sentiment Analysis

3.1. Sentiment Extraction:

The textual output from the image-to-text conversion is further analyzed to extract sentiments. This step involves using natural language processing techniques to determine the emotional tone conveyed by the text, which is indicative of the advertisement's intended impact on viewers.

Result:

Background context:

The dataset utilized in this project includes not only images of advertisements but also JSON files containing suggested actions flagged by users after viewing each advertisement. These suggestions are essential benchmarks for assessing the model's performance in understanding and suggesting appropriate actions based on ad content.

Detailed Analysis of Images in tabular form:

Image Input	Question: What should I do?	
	Real sentiment	Predicted Sentiment
	"I should wear Ray Ban." "I should buy ray Bans", "I should buy Ray Bans. "	After seeing this advertisement featuring an older man wearing a red and white scarf, you should consider visiting the Ray-Ban store to explore their eyewear collection. The advertisement aims to promote their products and may encourage potential customers to visit the store and make a purchase.
ANALYSIS:	The LLaVA model gives you a more generalized answer which is closely related to the User answer.	
	"I should buy an Audi.", "I should buy an Audi.", "I should watch car racing.", "I should buy an Audi e-tron. ", "I should get a fast car."	The advertisement showcases the Audi R18 e-tron racing car, which is an innovative and advanced vehicle. By visiting the website or contacting the company, you can gather more information about the car, its performance, and its potential benefits. This could help you make an informed decision about whether to purchase or invest in the Audi R18 e-tron or explore other options in the market.
ANALYSIS:	The output of the model is more focused on persuading prospective purchasers to look for additional information, which is consistent with the advertisement's goal of educating and captivating viewers about the features of the Audi R18 e-tron. The Audi purchase is suggested more directly by the user annotations, but the model's response offers an additional level of customer interaction by recommending more research and well-informed decision-making. This points to a deeper comprehension of the advertisement's strategic objective, which goes beyond simple sales and aims to increase consumer interest and trust in the brand.	

<table>
<tr>
<td>
</td>
<td>"I should visit the beauty bar.",
"I should visit the beauty bar ",
"I should go to the Beauty Bar",
"I should go to the beauty bar.",
"CALL THEM"</td>
<td>you should consider contacting The Beauty Bar at 699 East Broadway in South Boston, MA. The advertisement is promoting hair styling, coloring, and skin care services. If you are interested in any of these services, you can call the provided phone number (617-236-2276) to make an appointment or learn more about the services offered by the salon.</td>
</tr>
<tr>
<td>ANALYSIS:</td>
<td colspan="2">The model's output aligns well with the user annotations by encouraging contact with The Beauty Bar. While the user annotations suggest visiting the salon directly, the model extends this suggestion by also recommending calling for appointments or further information, which reflects a comprehensive understanding of the ad's purpose to not just attract walk-ins but also facilitate bookings.

The negative as you can see is that the phone number provided in the model's output does not match the number shown in the advertisement.</td>
</tr>
</table>

CONCLUSION:

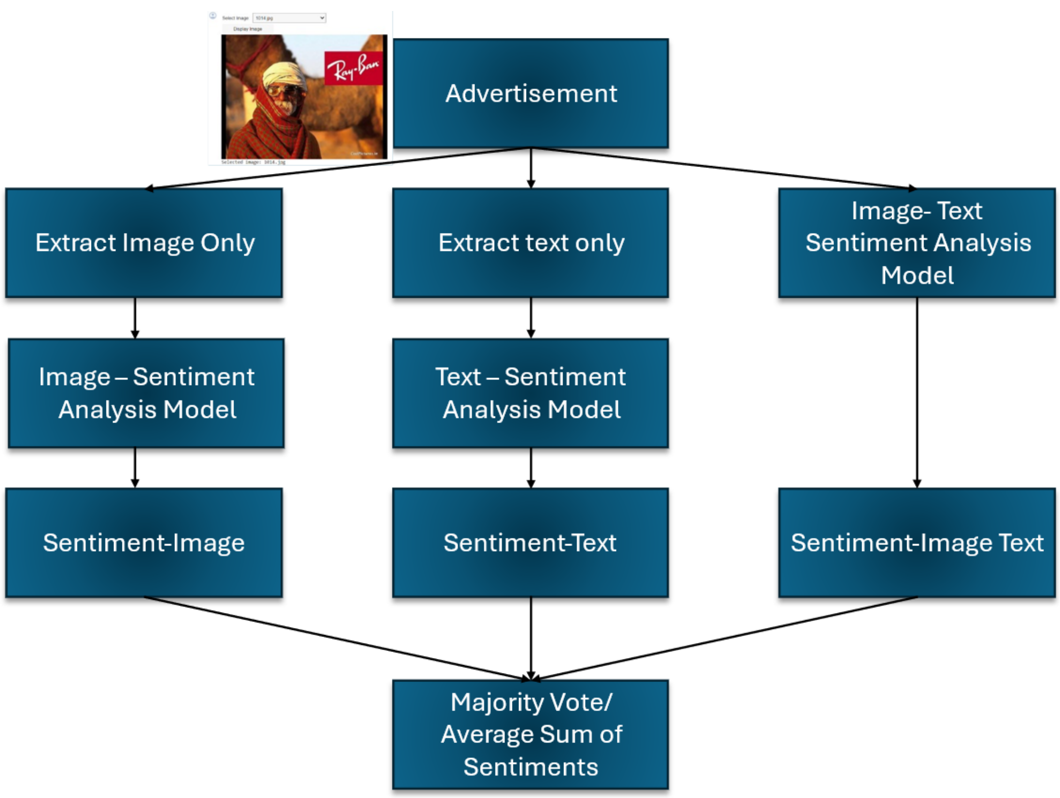
1. The model exhibits a strong capacity to comprehend and decipher the main idea and message of many commercials, whether they market services like The Beauty Bar's or cars like Audi. Generally speaking, it matches its recommendations with the kind of concrete actions that prospective clients are expected to take, like going to a store or getting in touch with a service provider.
2. Every time, the model makes suitable suggestions that are generally in line with user annotations, indicating that it can comprehend and react to the main components of an advertisement. This involves knowing when to recommend a visit, a buy, or make contact—all of which are essential for turning ad viewers into concrete actions.
3. Although the model works well in general, its specificity and accuracy should be enhanced. To improve the usefulness of the model's output, for example, contact information accuracy (as seen by the Beauty Bar example's wrong phone number) and more tailored responses depending on the advertisement content should be provided. Furthermore, promoting greater hands-on involvement that is, going somewhere in person as opposed to merely contacting someone might better serve corporate objectives.
4. According to the model's responses, it might improve consumer engagement by offering more context or action items for interaction, such as recommending more research into products or setting up meetings. This talent can be very helpful for digital marketing techniques since it can increase conversion rates by effectively engaging customers.
5. It could be advantageous to integrate the model with real-time data verification tools to prevent problems such as the provision of inaccurate information. This would increase the

dependability of AI-driven marketing advice by guaranteeing that all recommended activities or supplied details are correct and current.

In conclusion, the model demonstrates an optimistic performance in comprehending and reacting to commercials. Its outputs could be improved to be more precise, and accurate, and even provide more direct interaction recommendations, which would make it an even more effective tool for customer engagement and advertising.

Future Scope:

Future research focuses on a thorough sentiment analysis of the ad's constituent parts to determine which ones—text, image, or composition as a whole have the biggest effects on viewers' perceptions and feelings. The approach for carrying out this study is delineated in the flowchart that goes with it (see Figure), and it illustrates the suggested methodology.



Methodology explained in above flowchart:

1. Image Component Separation:
   - Text Only: Extract all the text from the image(advertisements) to analyze the sentiment conveyed through words alone.
   - Image Only: Remove all the text from the image and analyze the sentiment conveyed only through image.

- Combined Text and Image (Original Advertisement): Assess the sentiment of the entire advertisement.

2. Compare the results to determine which component most effectively communicates the intended sentiment.
3. Make use of the results to determine which method produces the most accurate and compelling sentiment analysis. We can also employ here majority voting mechanism or employ LLM text model which can average out individual sentiments. This will make it easier to comprehend how each element affects the audience's emotional response, both alone and collectively.

Apart from better-analyzing sentiments intended from the advertisements, such framework analysis will help designers of advertisements create more successful designs by enabling them to highlight the elements that most effectively communicate the intended emotion. Furthermore, this research may aid in the creation of more advanced AI-driven sentiment analysis systems that take into consideration the complex nature of advertisements.

References:

- Automatic Understanding of Image and Video Advertisements Zaeem Hussain Mingda Zhang Xiaozhong Zhang Keren Ye Christopher Thomas Zuha Agha Nathan Ong Adriana Kovashka Department of Computer Science University of Pittsburgh {zaeem, mzhang, xiaozhong, yekeren, chris, zua2, nro5, kovashka}@cs.pitt.edu
- Visual Instruction Tuning Haotian Liu1∗ , Chunyuan Li2∗ , Qingyang Wu3 , Yong Jae Lee1 1University of Wisconsin–Madison 2Microsoft Research 3Columbia University https://llava-vl.github.io