Adobe Behaviour Simulation Challenge

Team 41

Code Report

Abstract

In the contemporary landscape of artificial intelligence, the integration of multimodal learning has emerged as a groundbreaking avenue for understanding, predicting, and generating content. This report explores the synergistic potential of Video Llava(Lin et al., 2023) and Behavior Tokens in modeling behavioral patterns from visual stimuli, ultimately enabling to understand what would be behaviour of receiver on specific content and generate content which would get specific behaviour from receiver.

Introduction

The problems faced in Communication can be studied at three levels:

- Technical problem
- Semantic problem
- Effectiveness problem

In this report, we are going to understand one way to can tackle the effectiveness problem.

With advancement in Large Language Models (LLMs) like BERT, GPT-3, GPT-4 etc. which give a huge boost to Natural Language Processing (NLP) task. The Model which has been pre-trained on huge unspecified dataset can be fine-tune to perform specific task like question-answering, sentiment analysis which help us to solve the semantic problem very well. To understand the effect of message on receiver we need to understand the content. The objective can be, for example, to persuade the receiver to buy a product, to convince the receiver to vote for a candidate, or to make the student understand a concept or something else.

Literature Survey

VideoChat

VideoChat is an end-to-end chat-centric video understanding system. It integrates video foundation models and large language models via a learnable neural interface, excelling in spatiotemporal reasoning, event localization, and causal relationship inference. To instructively tune this system, we propose a video-centric instruction dataset, composed of thousands of videos matched with detailed descriptions and conversations. This dataset emphasizes spatiotemporal reasoning and causal relationships, providing a valuable asset for training chat-centric video understanding systems.

VideoChatGPT

VideoChat is a multifunctional video question answering tool that combines the functions of Action Recognition, Visual Captioning and ChatGPT. Our solution generates dense, descriptive captions for any object and action in a video, offering a range of language styles to suit different user preferences. It supports users to have conversations in different lengths, emotions, authenticity of language.

- Video-Text Generation
- Chat about uploaded video
- Interactive demo

Video-LLaVA

Video-LLaVA a Large Multimodal Model (LMM) design to align video and audio representations with a Large Language Model (LLM). Main feature of Video-LLaVA that is proficiently manage both video and audio data in conversational contexts and have additional feature of effective video grounding.

Approach

Effectiveness problem deals with how the given communication have effect on the receiver. To understand the the effect on the receiver we have to understand what message(like Images, GIF, Videos) has been send through the communication, to whom it was sent means what category of people and through which channel.

So we are going to use the Video-LLaVA to understand the Videos, Images and GIF. As we know Video-LLaVA has been designed for video in which it consider every frame as it features i.e $V_i \in R^{T \times H \times W \times C}$ where T represents the frame count of the video (in case of image T=1) and H, W, C so the height, width, color system of a frame. This leads to generating frame-level embedding from which we are able to gather temporal and spatial features and use the combination of them as single feature for a frame according to Video-LLaVA framework.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDEr
fully fine-tuned model	0.079	0.036	0.019	0.013	0.121	0.0.286
10k Finetuned Model	0.077	0.030	0.016	0.010	0.092	0.089
no fine-tuned	0.042	0.014	0.006	0.003	0.072	0.027

Table 1: Evaluation Metrics for Content Simulation.

Video-LLaVA have a learn able Multi-Layer Preceptron (MLP) which serve as our modal connector basically helps to convert the spatio-temportal feature such that it is compatibly with language decoder used in the Video-LLaVA.

The content feature which is provided in the data set which represent what "text" communicator have provided with Video, Image and GIF. This text is than used by the Grounding Module. Grounding Module extract key noun phrase from the text and try to associate it with the spatial feature got from the frames. Simultaneously, an image tagging model, Recognize Anything Model(RAM), tags visual elements in each frame, creating a detailed map of the video content.

We generate custom prompts for both behaviour and content simulation. Fine-tuning the model on this data set helps it understand the structure of tweets much better.

Model	RMSE
10k Finetuned Model	3427.902603303069
Full Finetuned Model	3412.89696154207

Table 2: Evaluation Metrics for Behaviour Simulation

For behaviour simulation training we use the content, media, username to understand to generate a custom prompt. From these features we try to predict how many likes we will we get from the social media post. This is compared against the actual number of likes.

For content simulation we use media (image or video) of the tweet, username and number of likes to generate the prompt. We ask the model to write the text content for such tweet. This is compared against the actual tweet content.

Social media content is highly diverse and subjective. People have varied interests, and predicting how a specific audience will respond to a particular post can be quite difficult. We think that better predictions can be made through preparing a richer data set which contains information about popular trends, popular accounts for mentions etc.

Results

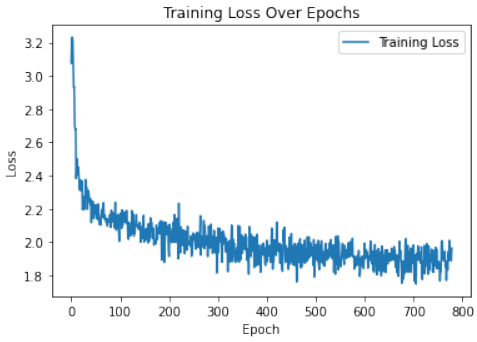

Figure 1: Loss

The plot above is loss variation during the fine-tuning of the model. The model was fine-tuned on a data set with both behaviour and content simulation prompts. We can see there is a steady decline in the loss indicating that the model learns to simulate contents of a tweet and predict the amount of likes on a tweet.
From table 1, which is for content simulation, we see that all evaluation metric saw an improvement with more fine-tuning data.
This shows that model has a greater understanding of the media, likes and its relation with the tweet.
From table 2 we observe that likes are much harder to predict from the given dataset as fine-tuning on 10k or 200k samples doesn't make much of a difference.
We think that reason for this might be likes are quite harder to predict given just the media and text content. More information such as number of followers, tags and metions used could potentially increase the predictability of likes

References

Ashmit Khandelwal, Aditya Agrawal, Aanisha Bhattacharyya, Yaman K Singla, Somesh Singh, Uttaran Bhattacharya, Ishita Dasgupta, Stefano Petrangeli, Rajiv Ratn Shah, Changyou Chen, and Balaji Krishnamurthy. 2023. Large content and behavior models to understand, simulate, and optimize content and behavior.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection.

(Khandelwal et al., 2023)