Team members:

Victor Wang zw2374

Daisy Huynh anh422

Marco Li yl9315

NLP Project Proposal: Sci-fi-zer

Introduction and Motivation:

Our project is to implement an interactive sci-fi themed text game in which a user can interact by inputting arbitrary text, or image inputs. Our model will then generate a storyline based on the user input, adapting them to a sci-fi theme. The model should have the capacity to continue projecting the storyline based on further prompts until the user ends the game.

We will acquire text by data scraping from human-written sci-fi stories available online or use available dataset obtained through similar sources. We will use the data to finetune a causal language model such as GPT2. To accept image inputs, we will employ a vision language model capable of few-shot learning, such as CLIP. Since we aim to process arbitrary inputs and “sci-fi-ze” them, we will use open-source image datasets for image captioning such as ImageNet. The challenge is to obtain sci-fi theme labels for these images which will be an important step for our project.

We are inspired by existing text-based games such as AI Dungeon. While AI Dungeon currently only allows users to interact by providing their actions through text descriptions, we want to upgrade the idea by allowing users to also input images while always maintaining the theme of the story.

Data Description:

For this project, there are two types of datasets that we will use, one for image input and the other for text input.

For text generation, we want to generate Sci-Fi stories. We will use Sci-Fi stories to fine-tune our model in the hope of thematic outputs. If we are to scrape data ourselves, we found a high-quality source called chooseyourstory.com containing original sci-fi stories. We’ll web scrape the text using Selenium. The website provides stories in the form of games, which means the plot will change based on the player’s response, and the storylines will develop in a tree structure. The data we get will be a group of stories and it might be good if we can store each story game in a tree. For pre-processing, we will use the tokenizer consistent with the causal language model we plan to use, such as gpt2 tokenizer.

If we plan to use an off-the-shelf dataset, we found a related source on Kaggle containing a single string of length 149326361. This string consists of several Sci-Fi books and contains copyright information and comments on each book. If we are going to use it, we will try to eliminate those irrelevant texts and split the string into different stories before passing it to the gpt2 tokenizer. One way to do this might be to track the special signs like “#” that appears in copyright information, which can be taken as the beginning of a new book.

For image input, currently we treat our task as an image captioning task where the model needs to generate themed captions for the image. However, dataset with regular images with creatively-altered captions are nowhere to be found. Therefore, we aim to perform data augmentation on the existing captioning datasets, altering the labels to a sci-fi setting. For the dataset, we might use Flickr30k, which contains 31783 images each labeled with 5 sentences, resulting in a total of around 150,000 captions. We will preprocess the data to remove irrelevant or low-quality images and captions. We can find Flickr30k on Kaggle. Besides, we can also use the COCO captioning dataset, which is 10 times larger. It depends on whether we would need such a big dataset. To perform data augmentation, one feasible way is to utilize large language models such as ChatGPT. To do that, we need prompt engineering to generate high-quality alternative captions efficiently. We are in discussion of whether to convert text prompts as well.

Algorithms:

Our approach is tentative. We are going to use the Huggingface library for our model implementations, which means our NLP algorithms are mostly transformer-based.

Our core model is a causal language model fine-tuned on sci-fi datasets simply with next-token generation. Since our dataset can have a tree structure, we can potentially adopt a more sophisticated way of keeping track of the game state, e.g. adding cross attention blocks to take in history context or game states.

To deal with text and image inputs, we utilize pretrained image and text encoders such as CLIP to obtain the embeddings. We can then train a supervised caption generator using the augmented caption labels generated by GPT-3. Using the resulting caption as the prompt, we can generate a paragraph of story using our causal language model. Architecture is flexible here and subject to change.

What we expected to gain from this project:

Through the progression of this project, we want to get hands-on experience of the knowledge that we learned in Natural Language Processing class on this project. Specifically, we want to understand the process of training a deep learning model with real data and learn how to combine state-of-the-art architectures to tackle a unique problem. Besides, we are really interested in how we can incorporate image inputs achieving multi-modality in our application, thus we would want to excel on this path for this final project for our class.