Week4 VLN Literature Review

Rui Xu

November 2021

1 Intro

This review refers to a new discovery for the Minecraft Vision and Language Navigation: Revolutionary datasets(with filtering methods), create new autonomous embodied agent based on existing frameworks.

2 Some Terminology

• MineDOJO
• Video-text contrastive learning framework MineCLIP
• POMDP. We model the programmatic task as a partially observable • Markov decision process (POMDP).
• Content Filtering and Correlation Filtering

3 Direction

Give the partial observations, e.g., video snippet, and the language prompt that describes the task, the agent needs to figure out 3 non-trivial matters to better evaluate the current state.
First, whether the target entities are present within its field of vision? (Mine-CLIP has addressed this question)
Second, whether the agent has made the right action toward the right target? (Currently we are facing)
Third, What is the relationship between each video snippet and the degree of completion of the task?

4 Solved problems

We bring up three key issues we need to solve for learning a vision-language model as the universal reward function for open-ended tasks.
We build and release a neat vision-language dataset for Minecraft using the YouTube videos from MineDojo.
We propose an RL-friendly vision-language model, CLIP4MC, which aligns actions implicitly contained in the video and transcript clips in addition to entities.

5 Datasets

MineDojo provides an extensive suite of APIs to interact with the Minecraft environment. In MineDojo, a total of 640K video clips(8-16 seconds/clip)sourced from 730K+ YouTube videos and their corresponding transcripts are leveraged to train a vision-language model (MineCLIP) that provides auxiliary reward signals for agents to learn task-specific strategies efficiently.

It is quite large and contains substantial noise. Some videos are irrelevant to learning basic game concepts, while the alignment between transcripts and videos may not always be precise, leading to temporal or content discrepancies that could hinder the learning process for a vision-language model.

The goal is to capture crucial details within the Minecraft environment, such as important elements and fundamental semantic events, using a reduced amount of data. To attain this objective, a two-step filtering process was utilized to create a clean video-text dataset.

5.1 Content Filtering

In this process, we could manually identify entity keywords within the transcripts using a predefined keyword list. We then extract transcript clips of a fixed context length, denoted as L, in order to include as many of these keywords as possible. These extracted transcript clips form the textual component of our dataset and define the positions of the corresponding video clips.

5.2 Correlation Filtering

The video clip start and end times may not match the transcript clips, so we align the midpoint of each for potential semantic overlap. We also extract video clips with a fixed duration (D). However, this method doesn't ensure consistency between video and transcript content. To address this, we use the pre-trained MineCLIP model to obtain embeddings for selected video and transcript clips. Despite these efforts, we cannot ensure complete consistency between video and transcript content. To address this issue, we employ the pre-trained MineCLIP model to obtain embeddings for chosen video and transcript clips. Subsequently, we calculate the cosine similarity distribution of their embeddings to characterize their correlation.

TODO:
Find the predefined keyword list
Find the optimal D, the fixed duration
According to [Din+23],
??

References

[Din+23] Ziluo Ding et al. "CLIP4MC: An RL-Friendly Vision-Language Model for Minecraft". In: *arXiv preprint arXiv:2303.10571* (2023).