Video PreTraining (VPT): Learning to Act by Watching Unlabeled Online Videos

Rui Xu

Solved questions

Many researchers have extensively explored the concept of pretraining using noisy, large-scale internet datasets as a method to develop models with wide-ranging, versatile capabilities in handling various data types like text, images, and more. Nevertheless, in the case of sequential decision-oriented domains like robotics, video games, and computer applications, publicly accessible data often lacks the necessary annotations needed for training behavioral priors in a similar fashion.

The key contributions of this study encompass the following:

1. Pioneering Application of Semi-Supervised Imitation Learning: This research marks the first instance of demonstrating promising outcomes by applying semi-supervised imitation learning to exceptionally extensive, noisy, and publicly accessible video datasets within the context of sequential decision domains.

2. Empowerment of Agents for Previously Insurmountable Tasks: The investigation reveals that the combination of pretraining and fine-tuning empowers agents to tackle tasks that were hitherto considered insurmountable through conventional learning methods.

3. Efficient Utilization of Labeled Contractor Data within the VPT Method: The study illustrates that the labeled contractor data is significantly more efficiently harnessed within the VPT (Video Pretraining) framework as opposed to direct training of a foundational model using this data.

4. Contribution to Research Community: The authors have generously open-sourced their contractor data, trained model weights, and the Minecraft environment. This invaluable resource is made available for the benefit of future

research endeavors focused on the study of learning to make decisions through semi-supervised imitation learning at a large scale.

Direction

The researchers extended the internet-scale pretraining paradigm to encompass sequential decision domains through the utilization of a semi-supervised imitation learning approach. In this approach, agents acquired the capability to make informed decisions by observing unlabeled online videos. To elaborate further, their study demonstrated that even with a limited quantity of labeled data, it was possible to effectively train an inverse dynamics model of sufficient accuracy. This model, in turn, facilitated the annotation of a substantial repository of unlabeled online data. In their case, this technique was applied to online videos capturing individuals engaged in Minecraft gameplay, ultimately facilitating the establishment of a comprehensive behavioral prior.

Technology

- The IDM is a type of model that predicts the action taken at each time step in a video.It is trained using a small amount of labeled data.The labeled data is used to teach the IDM how to predict the actions taken in a video.Once trained, the IDM can be used to predict the actions taken in a large but unlabeled dataset.
- Pseudo-labels are labels that are generated using a model rather than being manually assigned. These pseudo-labels can then be used to train other models or to evaluate the performance of existing models.
- Behavioral Cloning (BC) is a technique used in machine learning where a model learns to imitate a behavior by observing examples of that behavior. BC requires a large amount of data because the model must

learn to infer intent and the distribution over future behaviors from only past observations.

- Inverse dynamics modeling is simpler than behavioral cloning because it does not require the model to learn the intent behind the actions taken by the agent. Instead, the model only needs to learn the relationship between the current and future states of the environment and the actions taken by the agent. This makes the task of inverse dynamics modeling simpler and requires less data than behavioral cloning.

Methods

- Inverse Dynamics Models (IDM) VPT involves training a model IDM using labeled contractor data to predict actions at each timestep, considering both past and future events. This approach is easier and more data efficient compared to behavioral cloning, as shown in sections 4.1 and 4.6.
- IDM can be trained with as little as 100 hours of labeled data and can be used to label online videos, providing data for the task of behavioral cloning. Detailed training and data collection information can be found in appendices D and B.
- The foundation model is trained using standard behavioral cloning to minimize the negative log-likelihood of actions predicted by the IDM on clean data. This model can exhibit nontrivial zero-shot behavior and can be further improved with imitation learning and RL.

Performance

DM trains on only 1962 hours of data (compared to the ~70k hours of clean data we collected from the internet) and achieves 90.6% keypress accuracy and a 0.97 R2 for mouse movements evaluated on a held-out validation set of contractor-labeled data

Idea

Agents are rewarded for each item obtained in the sequence, with lower rewards for items that have to be collected in bulk and higher rewards for items near the end of the sequence.

A significant challenge encountered during the fine-tuning process using Reinforcement Learning (RL) pertains to the issue of catastrophic forgetting. This issue arises when previously acquired skills are lost before their inherent value can be effectively leveraged. For instance, although our Video Pretraining (VPT) foundation model does not explicitly exhibit the complete sequence of actions necessary for zero-shot iron smelting, it has been trained on instances of players engaging in the smelting process using furnaces. Consequently, the model may possess latent capabilities related to iron smelting once all the requisite prerequisites have been met.

To address the problem of catastrophic forgetting and harness latent skills for ongoing exploration and improvement throughout the RL fine-tuning process, we incorporate an auxiliary loss component based on the Kullback-Leibler (KL) divergence. This auxiliary loss function is applied to minimize the discrepancy between the RL model and the preserved pretrained policy, thus mitigating the potential loss of valuable latent skills during the fine-tuning procedure.

Conclusion

The findings presented in this paper significantly contribute to the prospective utilization of the abundant reserves of unlabeled web data within the realm of sequential decision domains. In contrast to generative video modeling or contrastive methods, which primarily yield representational priors, Video Pretraining (VPT) introduces a compelling prospect: the direct acquisition of actionable knowledge during the pretraining phase. The models in the study have demonstrated remarkable proficiency in zero-shot behavioral capabilities. Furthermore, VPT enables the utilization of these acquired behavioral priors as exceptionally potent exploration priors for Reinforcement Learning (RL).

Learning with the human keyboard and mouse interface is highly general and allows loss-lessly modeling the entire distribution of human behavior.

Importantly, VPT may also emerge as a superior general representation learning approach, even when the ultimate task is not directly related to action-based learning within the same domain. For instance, it could excel in fine-tuning to elucidate the content of a video. This is attributable to the argument that the most critical information within any given scene is inherently embedded in features that have been trained to accurately predict the distribution of future human actions.