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

Bowen Baker[*†] bowen@openai.com
Ilge Akkaya[*†] ilge@openai.com
Peter Zhokhov[*†] peterz@openai.com
Joost Huizinga[*†] joost@openai.com
Jie Tang[*†] jietang@openai.com
Adrien Ecoffet[*†] adrien@openai.com
Brandon Houghton[*†] brandon@openai.com
Raul Sampedro[*†] raulsamg@gmail.com
Jeff Clune[*†‡] jclune@gmail.com

Abstract

Pretraining on noisy, internet-scale datasets has been heavily studied as a technique for training models with broad, general capabilities for text, images, and other modalities.[1–6] However, for many sequential decision domains such as robotics, video games, and computer use, publicly available data does not contain the labels required to train behavioral priors in the same way. We extend the internet-scale pretraining paradigm to sequential decision domains through semi-supervised imitation learning wherein agents learn to act by watching online unlabeled videos. Specifically, we show that with a small amount of labeled data we can train an inverse dynamics model accurate enough to label a huge unlabeled source of online data – here, online videos of people playing Minecraft – from which we can then train a general behavioral prior. Despite using the native human interface (mouse and keyboard at 20Hz), we show that this behavioral prior has nontrivial zero-shot capabilities and that it can be fine-tuned, with both imitation learning and reinforcement learning, to hard-exploration tasks that are impossible to learn from scratch via reinforcement learning. For many tasks our models exhibit human-level performance, and we are the first to report computer agents that can craft diamond tools, which can take proficient humans upwards of 20 minutes (24,000 environment actions) of gameplay to accomplish.

1 Introduction

Work in recent years has demonstrated the efficacy of pretraining large and general foundation models[7] on noisy internet-scale datasets for use in downstream tasks in natural language[1–4] and computer vision.[5,6,8] For sequential decision domains (e.g. robotics, game playing, and computer usage) where agents must repeatedly act within an environment, a wealth of data also exists on the web; however, most of this data is in the form of *unlabeled* video (i.e. without the actions taken at each frame), making it much less straightforward to train a behavioral prior in these domains than it is in e.g. natural language. In a few rare settings, such as Chess, Go, and StarCraft, there

[*]This was a large effort by a dedicated team. Each author made huge contributions on many fronts over long time periods. All members were full time on the project for over six months. BB, IA, PZ, and JC were on the original VPT project team and were thus involved for even longer (over a year). Aside from those original team members, author order is random. It was also randomized between IA and PZ.

[†]OpenAI

[‡]University of British Columbia

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

Bowen Baker[*†] bowen@openai.com
Ilge Akkaya[*†] ilge@openai.com
Peter Zhokhov[*†] peterz@openai.com
Joost Huizinga[*†] joost@openai.com
Jie Tang[*†] jietang@openai.com
Adrien Ecoffet[*†] adrien@openai.com
Brandon Houghton[*†] brandon@openai.com
Raul Sampedro[*†] raulsamg@gmail.com
Jeff Clune[*†‡] jclune@gmail.com

Abstract

Pretraining on noisy, internet-scale datasets has been heavily studied as a technique for training models with broad, general capabilities for text, images, and other modalities.[1–6] However, for many sequential decision domains such as robotics, video games, and computer use, publicly available data does not contain the labels required to train behavioral priors in the same way. We extend the internet-scale pretraining paradigm to sequential decision domains through semi-supervised imitation learning wherein agents learn to act by watching online unlabeled videos. Specifically, we show that with a small amount of labeled data we can train an inverse dynamics model accurate enough to label a huge unlabeled source of online data – here, online videos of people playing Minecraft – from which we can then train a general behavioral prior. Despite using the native human interface (mouse and keyboard at 20Hz), we show that this behavioral prior has nontrivial zero-shot capabilities and that it can be fine-tuned, with both imitation learning and reinforcement learning, to hard-exploration tasks that are impossible to learn from scratch via reinforcement learning. For many tasks our models exhibit human-level performance, and we are the first to report computer agents that can craft diamond tools, which can take proficient humans upwards of 20 minutes (24,000 environment actions) of gameplay to accomplish.

1 Introduction

Work in recent years has demonstrated the efficacy of pretraining large and general foundation models[7] on noisy internet-scale datasets for use in downstream tasks in natural language[1–4] and computer vision.[5,6,8] For sequential decision domains (e.g. robotics, game playing, and computer usage) where agents must repeatedly act within an environment, a wealth of data also exists on the web; however, most of this data is in the form of *unlabeled* video (i.e. without the actions taken at each frame), making it much less straightforward to train a behavioral prior in these domains than it is in e.g. natural language. In a few rare settings, such as Chess, Go, and StarCraft, there

[*]This was a large effort by a dedicated team. Each author made huge contributions on many fronts over long time periods. All members were full time on the project for over six months. BB, IA, PZ, and JC were on the original VPT project team and were thus involved for even longer (over a year). Aside from those original team members, author order is random. It was also randomized between IA and PZ.

[†]OpenAI

[‡]University of British Columbia