Playing Minecraft Game with Human Feedback

SDSC8006 Presentation

Zichuan FU, Jinze LI, Lin LI, Wenlin ZHANG

April 19, 2024

Content

- Intro to the game and competition
- Methodology
- Experiment Results
- Conclusion

Minecraft

- The Game: A famous open-world game with high freedom

Figure: Minecraft

Competition

The Competition: NeurIPS 2022 MineRL BASALT [1]

- "Towards Solving Fuzzy Tasks with Human Feedback"

Challenge

- Complex Environments
- Hundreds Actions
- Sparse Reward
- Human Feedback

Tasks

- Data:
  Gameplay recordings for each task
- Four Tasks:
  Find-Cave/Make-Waterfall/Create-AnimalPen/Build-Village House

Methodology

- Random (bottom)
- Behavior Cloning (baseline—)
- KABasalt (RL)
- Human Expert (by Zichuan)

Behavior Cloning

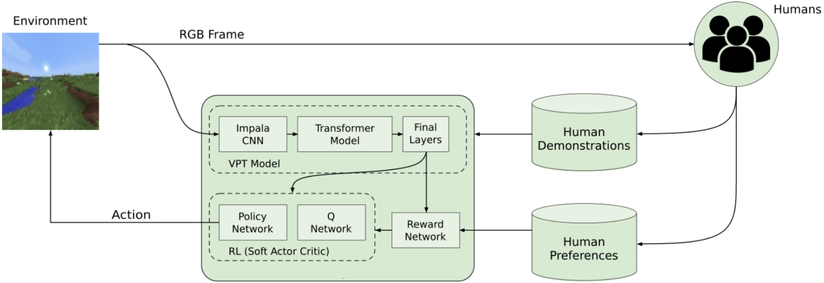
Behavior Cloning(BC): An imitation learning technique[2] that learns a policy to mimic the expert.[3]

- Expert Demonstrations: $D = \{(s_1, a_1), (s_2, a_2), ..., (s_N, a_N)\}$
- Learn the policy from the images
- Compare the two minimize the loss by update the policy
- Move to next step and loop the above

KAB

- Value Function: output of the OpenAI VPT model [4]
- State: 128 × 128 × 3 images
- Reward: learned from the Human preference
- Action: restricted to 16, eg. ”Forward”,”Left”,”Jump”

KAB Framework



Video Results

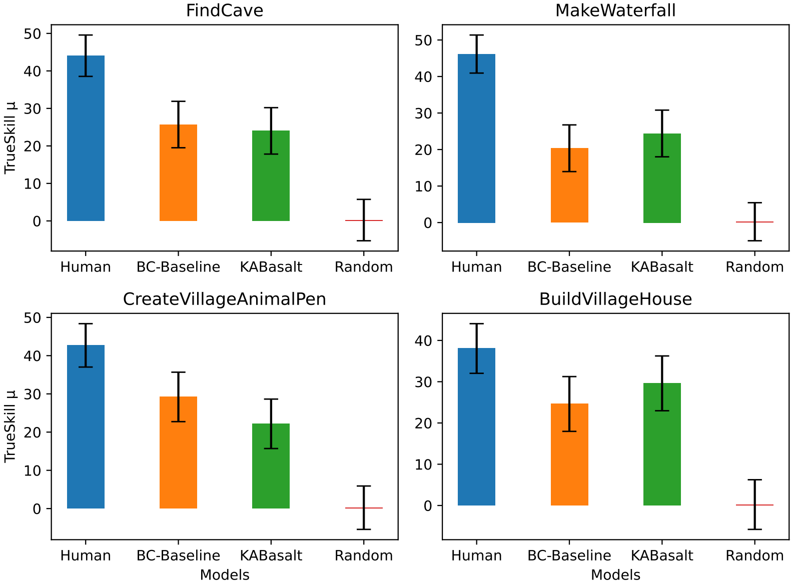
Find-Cave video:
video link

Evaluation

"TrueSkill" is a probabilistic rating system developed by Microsoft, primarily used to rate and match players in competitive gaming environments.[5]

- View each method as a kind of player
- The skill of each player follows a Normal Distribution with unknown parameters
- Rank players by pairs for each task in each round
- Update our belief about players by Bayesian approach, and get the posterior skill estimation

Performance Comparison



Performance Comparison

Model	FindCave	MakeWaterfall	CreateVillageAnimalPen	BuildVillageHouse	Average
Human	1.479	1.431	1.267	1.116	**1.323**
KABasalt	-0.108	0.021	-0.105	0.412	**0.055**
BC-Baseline	0.061	-0.283	0.446	-0.062	**0.041**
Random	-1.432	-1.169	-1.608	-1.466	-1.419

Table: Normalized TrueSkill scores for each model across the four tasks

Discussion

- The game is tough. Human's Feedback is helpful.
- Behavior Cloning simplifies the learning by supervised learning. Efficiency in the early stage. Challenges with distribution shift, Mix of expert demonstrations.
- RL method learns the environment.
  Partially observable, Hard to define the reward.
- Further: Learn the reasoning, Better reward design,...

References

[1] Stephanie Milani et al. "Towards solving fuzzy tasks with human feedback: A retrospective of the minerl basalt 2022 competition". In: *arXiv preprint arXiv:2303.13512* (2023).

[2] Adam Gleave et al. *imitation: Clean Imitation Learning Implementations*. 2022. arXiv: `2211.11972 [cs.LG]`.

[3] Anssi Kanervisto, Janne Karttunen, and Ville Hautamäki. "Playing minecraft with behavioural cloning". In: *NeurIPS 2019 Competition and Demonstration Track*. PMLR. 2020, pp. 56–66.

[4] Antonin Raffin et al. "Stable-Baselines3: Reliable Reinforcement Learning Implementations". In: *Journal of Machine Learning Research* 22.268 (2021), pp. 1–8. URL: `http://jmlr.org/papers/v22/20-1364.html`.

[5] B Schölkopf, J Platt, and T TrueSkill Hofmann. "A Bayesian Skill Rating System". In: *Advances in Neural Information Processing Systems* 20 (2006), pp. 569–576.