2025 ML

XMEN

X-Ray Multi-Input Engagement Navigators: A Recurrent D3QN for Minecraft Zombie Combat

Team1

Overview

We aim to train a reinforcement learning agent to perform better in a mini-game called “Zombies” in Minecraft

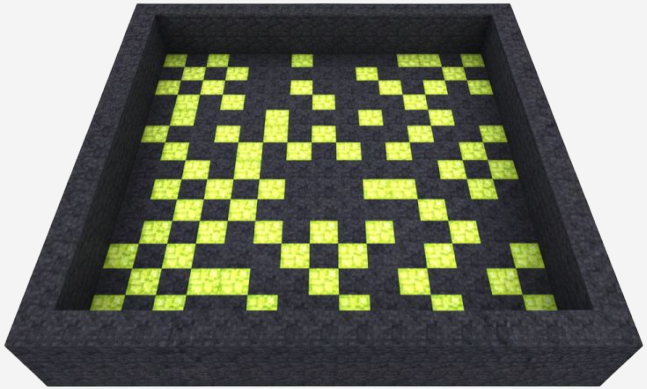
GOAL

Kill as many zombies as we can !

Environment Setting

Zombie
(Max. 2, respawn every 2 secs.)

30x30



Players
(Damage tolerance: 10s)

Platform: Malmo (Minecraft)
Tick Rate: 50 ms per tick (20 ticks/sec)
Regeneration: Natural regeneration is disabled
Difficulty: Normal (difficulty 1)

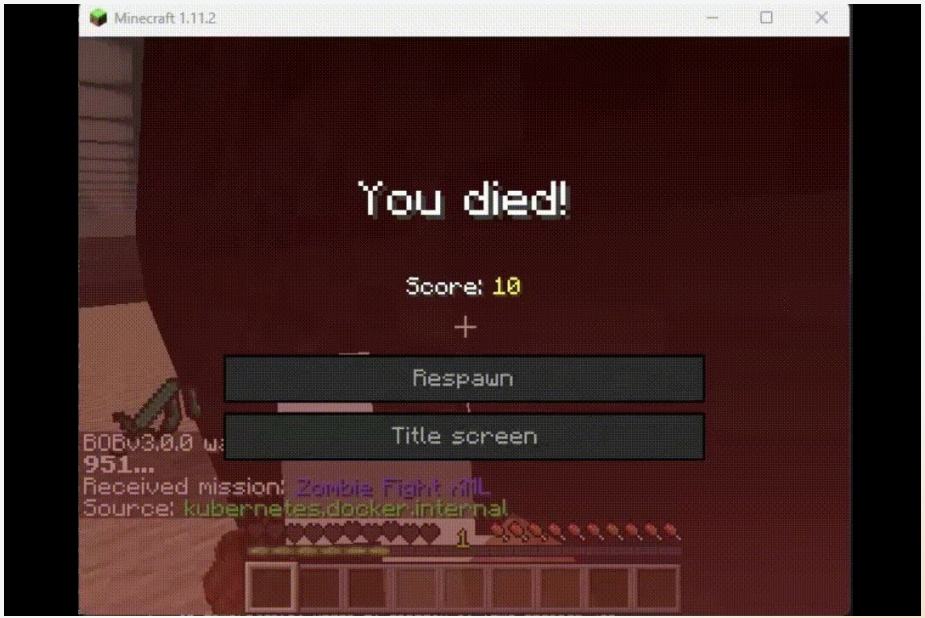
Related Work

Input

Visual Frame

Model

Resnet50 + DQN



Reward Function

Event	Original (Baseline)
Death	-100
Damage attack	+15
Damage Taken	-4
Per Action Taken	+0.05
Zombie Killed	+150

Problems

1. Unstable Training Process
2. Model can not correctly locate the position of zombies
3. Accumulate rewards most of the time are negative

Problems

1. Unstable Training Process
   - Need more stable model - Double Dueling DQN
2. Model can not correctly locate the position of zombies
3. Accumulate rewards most of the time are negative

Disadvantage of DQN

$$C(\Theta) = \sum_i \left[R^{(i)} + \gamma \max_{a'} f_{Q^*}\left(s^{(i+1)}, a'; \Theta^-\right) - f_{Q^*}\left(s^{(i)}, a^{(i)}; \Theta\right) \right]^2$$

Use **the exact same network** to choose action and evaluate it inside the **max** term

- **Cause positively overestimate Q-value**

Double DQN

- Use two different networks to choose actions and evaluate choosing actions separately

$$C(\Theta) = \sum_i \left[R^{(i)} + \gamma f_{Q^*}\left(s^{(i+1)}, \arg\max_{a'} f_{Q^*}\left(s^{(i+1)}, a'; \Theta\right); \Theta^-\right) - f_{Q^*}\left(s^{(i)}, a^{(i)}; \Theta\right) \right]^2$$

Disadvantage of DQN

$$C(\Theta) = \sum_i \left[R^{(i)} + \gamma \max_{a'} f_{Q^*}\left(s^{(i+1)}, a'; \Theta^-\right) - f_{Q^*}\left(s^{(i)}, a^{(i)}; \Theta\right) \right]^2$$

Only update specific action based on the state

- **Can we find a more efficient and stable way that updates actions relative to the state?**

Dueling DQN

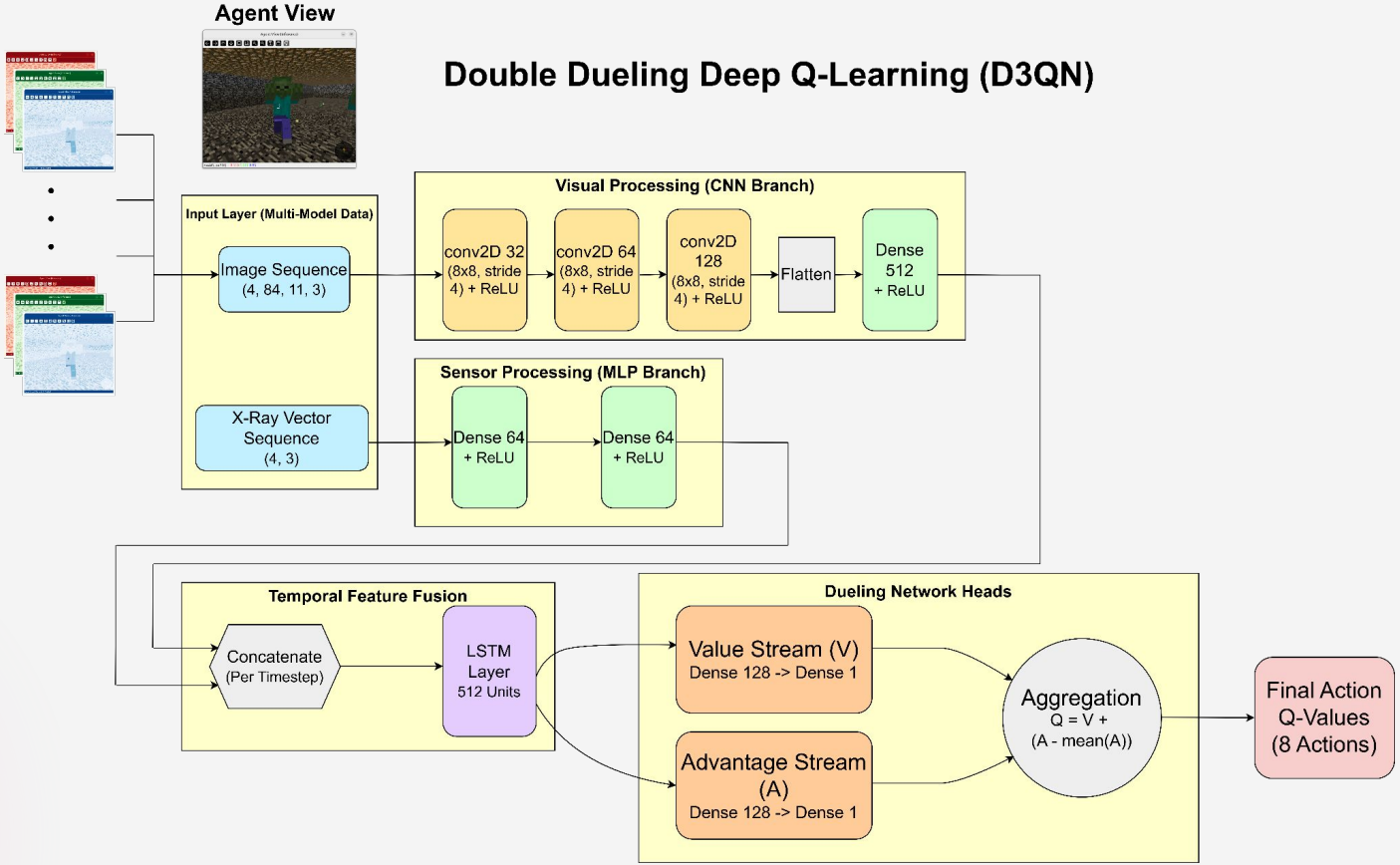
- Use V(s) + A(s,a) = Q(s,a) where V(s) represent how good a state is and A(s,a) represent how much better or worse each action is

$$f_{Q^*}(s,a) = f_{V^*}(s) + \left(f_{A^*}(s,a) - \frac{1}{|\mathcal{A}|} \sum_{a'} f_{A^*}(s,a') \right)$$

Problems

1. Unstable Training Process
2. Model can not correctly locate the position of zombies
   - Require additional features to locate zombies
3. Accumulate rewards most of the time are negative

Model Architecture



Model

Input data

- visual sequence (4, 84, 112, 3).
- x-ray sequence (4, 3) Vector: [local_x, local_z, distance]
- 4 consecutive frames allow the model to infer trajectory information

Feature Extraction

- Visual Processing (CNN Branch)
    - conv2D * 3 kernel(3,3), strides =1, channels=32, 64, 128.
    - flatten + Dense 128
- Sensor Processing (MLP Branch)
    - Dense 64 * 2

Model

Temporal Features Fusion

- Concatenate visual features and position features for each timestep
- LSTM 512 as Memory Unit

Dueling DQN Head

- Use Dense 128 -> Dense 1 as Value function
- Use Dense 128 -> Dense |A| as Advantage function
- Output = Value + (Advantage - mean(Advantage))

Problems

1. Unstable Training Process
2. Model can not correctly locate the position of zombies
3. Accumulate rewards most of the time are negative
   - Poor design of reward functions

Reward Function

Event	Original (Baseline)	Ours	Description
Death	-100	-400	Terminal penalty
Damage attack	+15	+30	Reward for dealing damage
Damage Taken	-4	-10	Per HP lost (Reduced from -50 to encourage trading blows)
Per Action Taken	+0.05	-	
Zombie Killed	+150	+100	Primary objective
Survive (Per tick)	-	+0.1	If no damage taken
Aiming	-	+5	Zombie <= 15 degrees of the crosshair and <= 2.5 blocks)
Kiting Bonus	-	+0.5	If aiming at zombie while distance is b/w 3 and 6 blocks
Missed Opportunity	-	-2	< 2.5 blocks and attack is ready, but does NOT attack
Inactivity	-	-0.5	If agent hasn't attacked for > 5 seconds
Wall Penalty	-	-5	If facing a wall < 1.5 blocks
Safe Attack Bonus	-	+10	If agent attacks and doesn’t lose HP this step

Reward Mining Process – 1

Death penalty >> Multiple kills reward

Player will always tend to run away from zombies

Single kill reward >> Death penalty + Damage taken penalty

player will always attack zombies with maximum aggression without retreating

How To Solve?

Expectation of kills * single kill reward = - (Death penalty + Damage taken penalty)

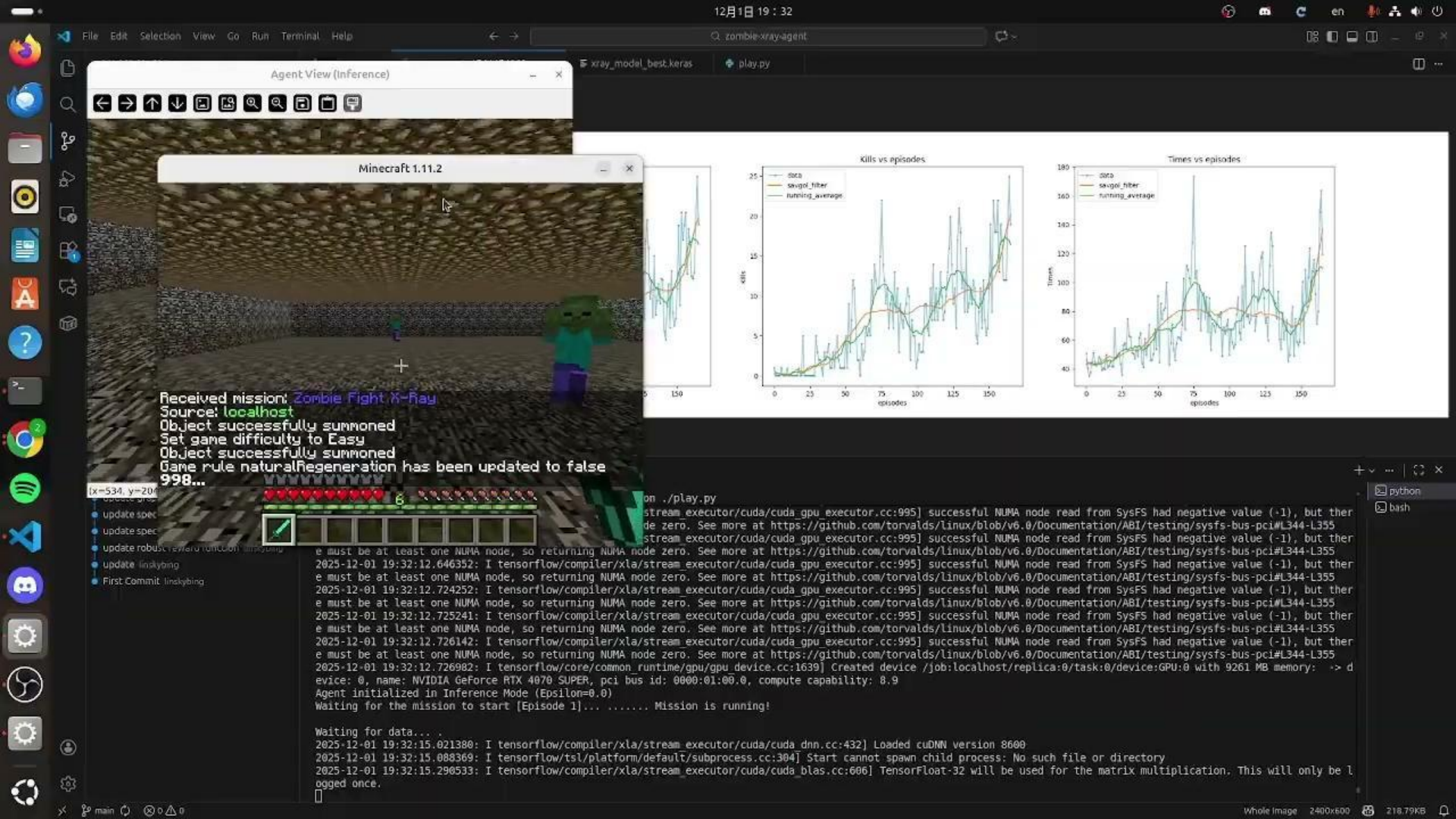
Reward Mining Process – 2

The player cannot aim at zombies within the valid angle and attack them within range

How To Solve?

Adding Aiming reward and Wall penalty to help model better aim at zombies

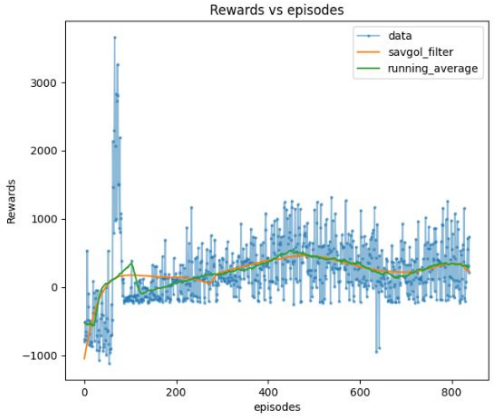
DEMO

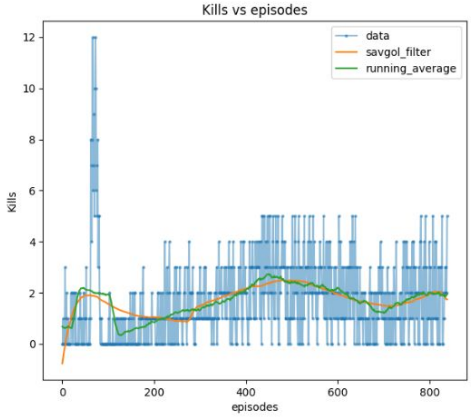

12月1日 19：32
File Edit Selection View Go Run Terminal Help
zombie-xray-agent
Agent View (Inference)
Minecraft 1.11.2
Received mission: Zombie Fight X-Ray
Source: localhost
Object successfully summoned
Set game difficulty to Easy
Object successfully summoned
Game rule naturalRegeneration has been updated to false
998...
xray_modal_best.keras
play.py
Kills vs episodes
Times vs episodes
data
savgol_filter
running_average
kills
Times
episodes
update linskybing
First Commit linskybing
on ./play.py
stream_executor/cuda/cuda_gpu_executor.cc:995] successful NUMA node read from SysFS had negative value (-1), but there must be at least one NUMA node, so returning NUMA node zero. See more at https://github.com/torvalds/linux/blob/v6.0/Documentation/ABI/testing/sysfs-bus-pci#L344-L355
2025-12-01 19:32:12.646352: I tensorflow/compiler/xla/stream_executor/cuda/cuda_gpu_executor.cc:995] successful NUMA node read from SysFS had negative value (-1), but there must be at least one NUMA node, so returning NUMA node zero. See more at https://github.com/torvalds/linux/blob/v6.0/Documentation/ABI/testing/sysfs-bus-pci#L344-L355
2025-12-01 19:32:12.724252: I tensorflow/compiler/xla/stream_executor/cuda/cuda_gpu_executor.cc:995] successful NUMA node read from SysFS had negative value (-1), but there must be at least one NUMA node, so returning NUMA node zero. See more at https://github.com/torvalds/linux/blob/v6.0/Documentation/ABI/testing/sysfs-bus-pci#L344-L355
2025-12-01 19:32:12.725241: I tensorflow/compiler/xla/stream_executor/cuda/cuda_gpu_executor.cc:995] successful NUMA node read from SysFS had negative value (-1), but there must be at least one NUMA node, so returning NUMA node zero. See more at https://github.com/torvalds/linux/blob/v6.0/Documentation/ABI/testing/sysfs-bus-pci#L344-L355
2025-12-01 19:32:12.726142: I tensorflow/compiler/xla/stream_executor/cuda/cuda_gpu_executor.cc:995] successful NUMA node read from SysFS had negative value (-1), but there must be at least one NUMA node, so returning NUMA node zero. See more at https://github.com/torvalds/linux/blob/v6.0/Documentation/ABI/testing/sysfs-bus-pci#L344-L355
2025-12-01 19:32:12.726982: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1639] Created device /job:localhost/replica:0/task:0/device:GPU:0 with 9261 MB memory: -> device: 0, name: NVIDIA GeForce RTX 4070 SUPER, pci bus id: 0000:01:00.0, compute capability: 8.9
Agent initialized in Inference Mode (Epsilon=0.0)
Waiting for the mission to start [Episode 1]... ..... Mission is running!
Waiting for data... .
2025-12-01 19:32:15.021380: I tensorflow/compiler/xla/stream_executor/cuda/cuda_dnn.cc:432] Loaded cuDNN version 8600
2025-12-01 19:32:15.088369: I tensorflow/tsl/platform/default/subprocess.cc:304] Start cannot spawn child process: No such file or directory
2025-12-01 19:32:15.290533: I tensorflow/compiler/xla/stream_executor/cuda/cuda_blas.cc:606] TensorFloat-32 will be used for the matrix multiplication. This will only be logged once.
python
bash
main
Whole Image 2400x600 218.79KB

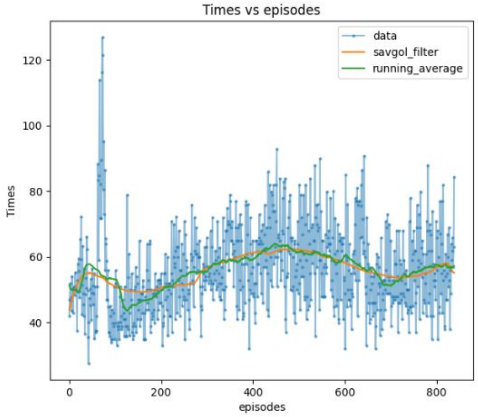
Performance Comparison

	Avg		Highest	
	Kill number	Survival time	Kill number	Survival time
Baseline	0.14	21.49s	2 (episode 259)	36.41s (episode 182)
Ours	**9.52**	**81.95s**	**22** (episode 21)	**166.77s** (episode 3)

Result


Rewards vs episodes
data
savgol_filter
running_average
Rewards
episodes
3000
2000
1000
0
−1000
0
200
400
600
800


Kills vs episodes
data
savgol_filter
running_average
Kills
episodes
12
10
8
6
4
2
0
0
200
400
600
800


Times vs episodes
data
savgol_filter
running_average
Times
episodes
120
100
80
60
40
0
200
400
600
800

Thank You!