Pommerman

Game Description

* Has been held at NeurIPS@2018 and NeurIPS@2019

* Objectives: Use bomb to wipe out enemies

* 11x 11 grid world with two agents and two rule based enemies

* Wood and walls;

* Bomb: blast range is 3; 10 timestep life; blast then gain a new one

—https://www.pommerman.com/

Game Description

Power-Ups: Half of the wooden walls have hidden power-ups that are revealed when the wall is destroyed. These are:

- *Extra Bomb*: Picking this up increases the agent's ammo by one.
- *Increase Range*: Picking this up increases the agent's blast strength by one.
- *Can Kick*: Picking this up permanently allows an agent to kick bombs by moving into them. The bombs travel in the direction that the agent was moving at one unit per time step until they are impeded either by a player, a bomb, or a wall.

Observations

- *Board*: 121 Ints. The flattened board. In partially observed variants, all squares outside of the 5x5 purview around the agent's position will be covered with the value for fog (5).
- *Position*: 2 Ints, each in [0, 10]. The agent's (x, y) position in the grid.
- *Ammo*: 1 Int. The agent's current ammo.
- *Blast Strength*: 1 Int. The agent's current blast strength.
- *Can Kick*: 1 Int, 0 or 1. Whether the agent can kick or not.
- *Teammate*: 1 Int in [-1, 3]. Which agent is this agent's teammate. In non-team variants, this is -1.
- *Enemies*: 3 Ints in [-1, 3]. Which agents are this agent's enemies. In team variants, the third int is -1.
- *Bomb Blast Strength*: List of Ints. The bomb blast strengths for each of the bombs in the agent's purview.
- *Bomb Life*: List of Ints. The remaining life for each of the bombs in the agent's purview.

Actions

On every turn, agents choose from one of six actions:

1. *Stop*: This action is a pass.
2. *Up*: Move up on the board.
3. *Left*: Move left on the board.
4. *Down*: Move down on the board.
5. *Right*: Move right on the board.
6. *Bomb*: Lay a bomb.

Reward

* Win: 1
* Loss: 0

Challanges

Sparse and deceptive rewards: the former refers to the fact that the only non-zero reward is obtained at the end of an episode. The latter refers to the fact that quite often a winning reward is due to the opponents' involuntary suicide, which makes reinforcing an agent's action based on such a reward *deceptive*. Note that suicide happens frequently during learning since an agent has to place bombs to explode wood to move around on the board, while due to terrain constraints, in some cases, performing non-suicidal bomb placement requires complicated, long-term, and accurate planing.

Delayed action effects: the only way to make a change to the environment (e.g., bomb wood or kill an agent) is by means of bomb placement, but the effect of such an action is only observed when the bomb's timer decreases to 0; more complications are added when a placed bomb is kicked to another position by some other agent.

Uninformative multiagent credit assignment: In the team environment, the same episodic reward is given to two members of the team. It may not be clear how to assign credit to individual agents. For example, consider an episode where an agent eliminates an opponent but then commits suicide, and its teammate eliminates the remaining opponent. Under this scenario, both team members get a positive reward from the environment, but this could reinforce the suicidal behaviour of the first agent. Similarly, one agent could eliminate both opponents whereas its teammate just camps; both agents would get positive rewards, reinforcing a *lazy* agent [7].

Techs may helpful

Action Filter

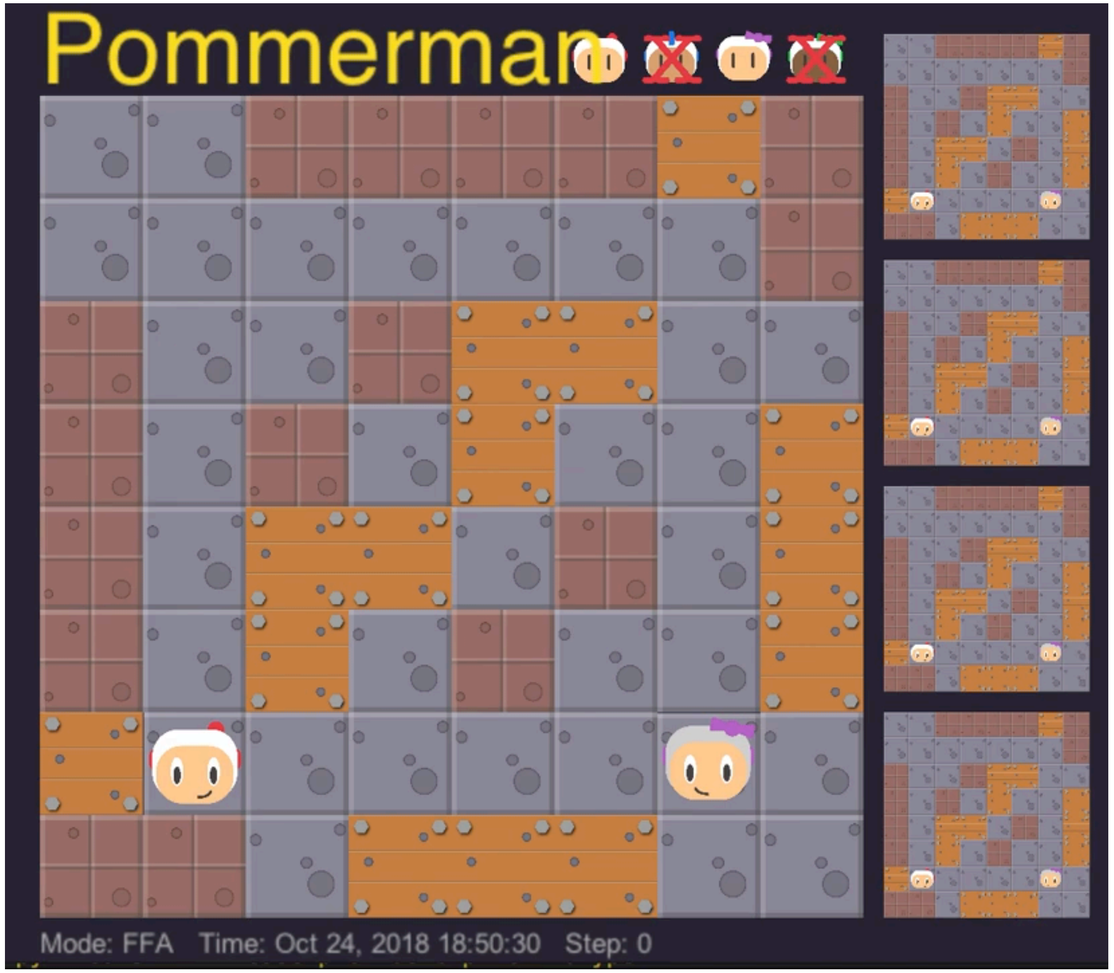
knowledge to the agent by telling the agent *what not to do* and let the agent discover *what to do* by trial-and-error. The benefit is twofold: 1) the learning problem is simplified since suicidal actions are removed and bomb placing becomes safe; and 2) superficial skills such as not moving into flames and evading bombs in simple cases are handled. Below we describe the main components of our team: the ActionFilter and the reinforcement learning aspect.

Table 1: ActionFilter rules

Avoiding Suicide	Not going to positions that are flames on the next step. Not going to *doomed* positions, i.e., positions where if the agent were to go there the agent would have no way to escape. For any bomb, doomed positions can be computed by referring to its `blast strength`, `blast range`, and `life`, together with the local terrain.
Bomb Placement	Not place bombs when teammate is close, i.e., when their Manhattan distance is less than their combined blast strength. Not place bombs when the agent's position is covered by the blast of any previously placed bomb.

—https://github.com/BorealisAI/pommerman-baseline

Techs may helpful



—https://github.com/BorealisAI/pommerman-baseline

Techs may helpful

Network Structure

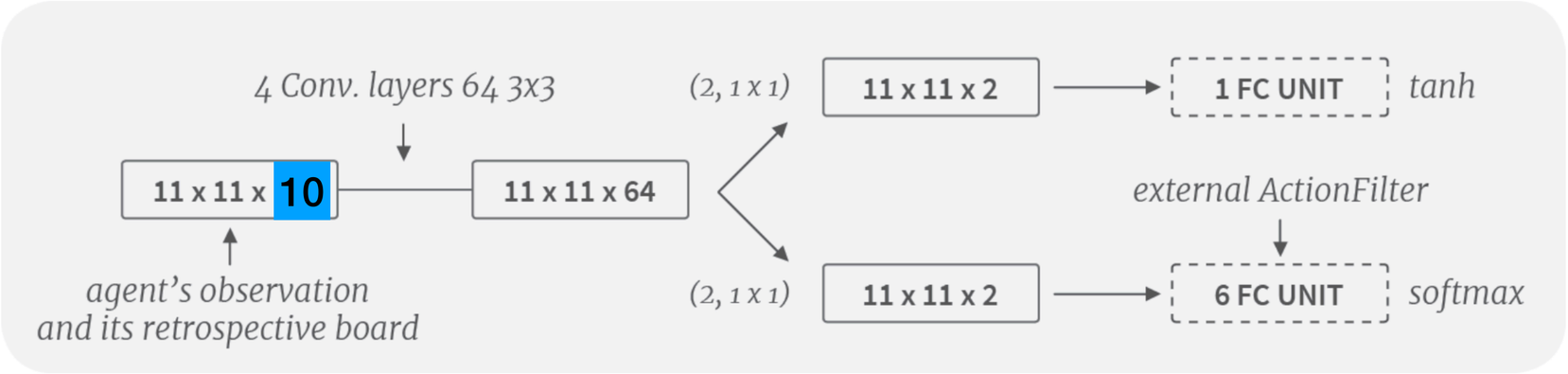


Figure 2: Architecture used for the `skynet955` agents

—https://github.com/BorealisAI/pommerman-baseline

Techs may helpful

Reward Shaping

Table 2: Reward Shaping for `skynet955` agents

Going to a cell not in a 121-length FIFO queue gets 0.001.	At the end of a game, dead agent in the winning team gets 0.5.
Picking up `kick` gets 0.02.	For draw games, all agents receive 0.0.
Picking up `ammo` gets 0.01.	On one enemy's death gets 0.5.
Picking up `blast strength` gets 0.01.	On a teammate's death gets -0.5.

—https://github.com/BorealisAI/pommerman-baseline