Training Reinforcement Learning Agents for Ultimate Mortal Kombat 3

Nilay Verma
IIT Gandhinagar
ID: 23110219
nilay.verma@iitgn.ac.in

Soham Ashish Gaonkar
IIT Gandhinagar
ID: 23110314
soham.gaonkar@iitgn.ac.in

Rudra Pratap Singh
IIT Gandhinagar
ID: 23110281
rudra.pratap@iitgn.ac.in

Romit Mohane
IIT Gandhinagar
ID: 23110279
romit.mohane@iitgn.ac.in

***Abstract*—This report presents a comparative study of three deep reinforcement learning (RL) algorithms Advantage Actor Critic (A2C), Proximal Policy Optimization (PPO), and a value based Deep Q Network (DQN) applied to the arcade fighting game *Ultimate Mortal Kombat 3* (UMK3) using the DIAMBRA Arena environment. All agents share a dual input neural architecture that combines convolutional processing of visual observations with multi layer perceptron (MLP) processing of game state features. We analyze the impact of algorithmic choices and reward shaping on learning dynamics, aggressive behavior, and susceptibility to local optima. In our experiments, the DQN agent matches the A2C baseline with consistently losing rewards around -332, whereas PPO augmented with an aggression oriented reward function yields substantially higher episodic returns and more proactive combat behavior, at the cost of increased sensitivity to local minima.**

I. Introduction

Deep reinforcement learning has demonstrated strong performance on high-dimensional control tasks, particularly in domains involving visual perception and discrete action spaces. Fighting games such as *Ultimate Mortal Kombat 3* (UMK3) pose additional challenges: sparse and delayed rewards, highly non-linear combat dynamics, and a large space of temporally extended strategies.

This work investigates the training and behavior of two policy gradient methods (A2C and PPO) and a value based Deep Q Network (DQN) on UMK3. The primary objectives are:

- To evaluate whether a shared dual stream visual and state based architecture can support competent control in this domain.
- To compare the learning stability and sample efficiency of A2C, PPO, and DQN under a common architecture and environment configuration.
- To study how reward shaping tailored toward aggressive play influences emergent strategies, particularly in the PPO setting, and how a value based method such as DQN copes with the same reward structure.

The analysis emphasizes methodology, model architecture, and interpretation of learning dynamics, rather than implementation or usage details.

II. Methods

A. Environment Configuration

All experiments are conducted in the DIAMBRA Arena implementation of UMK3 under a fixed configuration:

- **Game and character:** UMK3, Kitana.
- **Difficulty:** Level 2 AI opponent.
- **Action space:** Discrete set of 14 actions, including movement, blocking, and a subset of offensive actions (e.g., high punch, low kick).

Episodes consist of a sequence of combat interactions against the built in game AI. The environment exposes both raw visual observations and structured game state features (e.g., remaining health, timer, and aggressor gauge).

B. Observations and State Representation

Each time step provides two complementary modalities:

1) **Visual input:**
   - Raw RGB frames are converted to grayscale.
   - Frames are resized to 128×128 pixels.
   - A stack of the 4 most recent frames is maintained to encode short term temporal dynamics (e.g., projectile travel, jump arcs, and follow up attacks).
2) **Game state scalars:**
   - Normalized character identifier.
   - Health in the range $[0, 1]$.
   - Aggressor bar in the range $[0, 1]$.
   - Normalized timer value.

This design allows the agent to leverage high level state features that are difficult to infer reliably from vision alone,

while still benefiting from the rich spatial structure of raw frames.

C. Neural Network Architecture

All three agents share an identical dual input, shared backbone architecture:

1) **CNN encoder (visual branch):** A three layer convolutional network processes the 4 frame grayscale stack. Convolutions with progressively increasing receptive fields and channels extract spatial and motion sensitive features (e.g., opponent position, proximity, projectiles).
2) **MLP encoder (state branch):** A small MLP processes the scalar game state features (character ID, health, aggressor, timer), embedding them into a feature vector capturing coarse global context (e.g., winning/losing status, urgency based on time).
3) **Fusion and heads:** The outputs of the CNN and MLP encoders are concatenated and passed through shared fully connected layers to obtain a joint latent representation. This representation feeds two task specific heads:
   - *Actor head:* Outputs policy logits $\pi_\theta(a \mid s)$ over the discrete action set.
   - *Critic head:* Outputs a scalar value estimate $V_\phi(s)$, representing the expected return from the current state.

This shared representation encourages the model to learn features that are simultaneously useful for both action selection and value estimation, which is standard in modern actor critic methods.

D. Advantage Actor Critic (A2C)

The A2C agent performs synchronous updates over collected trajectories.

1) Advantage Estimation: The advantage is computed as

$$A(s_t, a_t) = R_t - V(s_t), \tag{1}$$

where R_t is the discounted return from time step t.

2) Loss Function: The total loss combines policy, value, and entropy terms as

$$L_{\text{total}} = L_{\text{policy}} + c_1 L_{\text{value}} - c_2 H(\pi), \tag{2}$$

where $H(\pi)$ is the policy entropy and c_1, c_2 control the relative weighting of value error and exploration, respectively.

3) Optimization Details: The A2C agent uses the Adam optimizer with gradient norm clipping at 0.5. An entropy coefficient of 0.1 encourages continued exploration, and temperature scaling (factor 1.0) on the logits maintains a well conditioned action distribution. The emphasis in this configuration is on stable, low variance learning dynamics rather than aggressively improving short term performance.

E. Proximal Policy Optimization (PPO)

The PPO agent uses a clipped surrogate objective to constrain policy updates and improve stability over vanilla policy gradient methods.

1) Advantage Estimation and Updates: PPO employs Generalized Advantage Estimation (GAE) to reduce variance while retaining sufficient bias control. Rollouts are collected and then reused for several epochs of mini batch stochastic gradient descent on the PPO objective.

The clipped objective takes the form

$$L_{\text{PPO}} = \mathbb{E}_t\big[\min(r_t A_t, \text{clip}(r_t, 1-\epsilon, 1+\epsilon)A_t)\big], \tag{3}$$

where r_t is the probability ratio between new and old policies, A_t is the estimated advantage, and ϵ is the clipping parameter.

2) Reward Shaping for Aggressive Behavior: Initial PPO experiments relying solely on the default DIAMBRA reward signal produced agents that exhibited overly cautious or passive strategies. To promote more aggressive play, a domain informed reward shaping scheme was introduced. The action set is partitioned into movement and attack categories, and the reward is modified as follows:

```
# Reward shaping logic for PPO
movement_ids = [0, 1, 2, 3, 4, 5, 6, 7, 8, 13, 14]
attack_ids = [9, 10, 11, 12]

# Amplify positive rewards (successful hits)
if reward > 0:
   reward *= 2.0

# Incentivize attempting attacks (aggression bonus)
if action.cpu().item() in attack_ids:
   reward += 2.8
```

Listing 1. Reward shaping logic for PPO

This scheme doubles positive rewards associated with successful damage and adds a fixed aggression bonus whenever an attack action is issued. The intent is to bias exploration toward offensive maneuvers while retaining the original environment feedback as the primary learning signal.

F. Deep Q-Network (DQN)

In addition to the actor critic methods, we implemented a value based Deep Q-Network (DQN) agent using the same dual input architecture. Instead of separate actor and critic heads, the DQN head outputs a vector of Q values $Q_\theta(s, a)$, one for each discrete action in UMK3.

The network consists of a three layer convolutional encoder for the stacked grayscale frames and a small MLP encoder for the normalized game state features. The two embeddings are concatenated and passed through a fully connected "combined feature" MLP whose output is fed into the Q head. An overview of this architecture is shown in Fig. 1.

Training follows the standard DQN recipe with several stabilizing components. A replay buffer stores transitions $(s, a, r, s', \text{done})$ and is sampled uniformly to break temporal correlations. A separate target network Q_{θ^-} is updated periodically with the online parameters θ and is used to compute the temporal difference target

$$L(\theta) = \mathbb{E}\big[(r + \gamma \max_{a'} Q_{\theta^-}(s', a') - Q_\theta(s, a))^2\big]. \tag{4}$$

Action selection during training uses an ϵ greedy policy, with ϵ linearly annealed from near 1.0 to a small value (approximately 0.01). Default hyperparameters include a learning

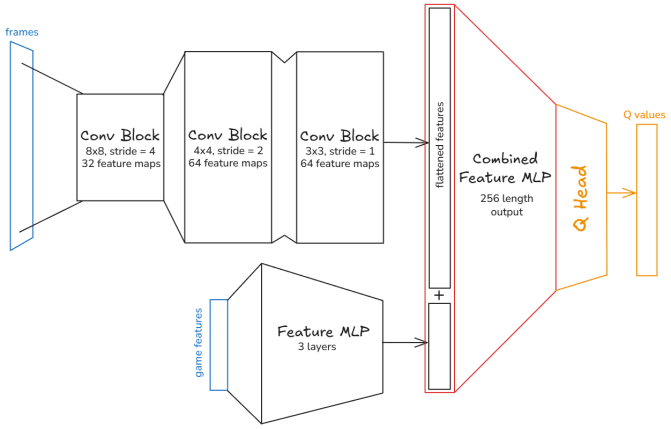

Fig. 1. Dual input DQN architecture. A three layer CNN processes stacked visual frames while a small MLP embeds normalized game state variables. The concatenated representation is mapped to per action Q values through a fully connected head.

rate of 10^{-4}, discount factor $\gamma = 0.99$, replay capacity of 10^5 transitions, batch size 64, and target network updates every 1000 environment steps.

III. Experimental Setup

Training runs are organized to highlight the distinct behaviors of the three algorithms (A2C, DQN, and PPO) and the effect of reward shaping:

- **A2C:** Trained with the standard environment rewards and the dual input architecture described above.
- **DQN:** Trained with the same dual input architecture but a value based Q head, using replay buffer, target network, and ϵ greedy exploration with a linear decay schedule.
- **PPO Phase A (baseline):** Trained with the standard environment rewards, without custom shaping.
- **PPO Phase B (aggressive shaping):** Trained with the reward shaping scheme of Section II-D.
- **PPO Phase C (local minima regime):** Continued training beyond the onset of high performance to observe convergence behavior and potential collapse into sub optimal deterministic strategies.

We did not implement a tabular Q learning baseline because the observation space is high dimensional and continuous (raw image frames and normalized state scalars), making classical discrete state Q learning impractical without heavy, arbitrary discretization.

Performance is monitored via average episodic reward and qualitative inspection of gameplay behavior (e.g., action diversity, spacing, and pressure strategies against the AI opponent).

IV. Results

A. A2C Performance

The A2C agent exhibits gradual but consistent improvement over training when evaluated under the default reward signal. Training curves show a smooth upward trend in average episodic reward, with relatively low variance. Synchronous updates and normalized advantages contribute to stable learning,

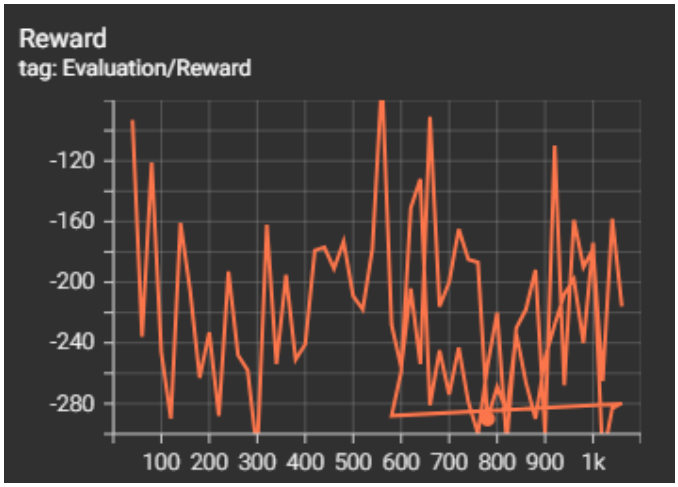

Fig. 2. A2C training rewards. The agent exhibits a gradual increase in performance, demonstrating that the dual input architecture and actor critic formulation can learn meaningful control policies in UMK3. (the redo plotting is because of checkpointing)

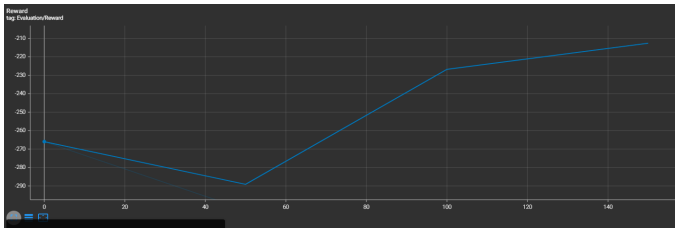

Fig. 3. DQN evaluation rewards over training. The curve exhibits substantial variance and remains in the negative reward regime, indicating that the value based agent struggles to consistently defeat the opponent under the current hyperparameters.

with no pronounced collapses or oscillations. The entropy term successfully prevents premature convergence to degenerate policies dominated by a single repetitive action.

Over a small held out evaluation set of ten episodes (recorded in the accompanying `results.csv`), the final A2C checkpoint achieves an average reward of -332. All evaluation episodes attain the same reward, indicating highly consistent but still losing performance against the built in opponent.

B. DQN Performance

The DQN agent uses the same dual input backbone as A2C and PPO, but optimizes a value based objective with replay buffer, target network, and ϵ greedy exploration as described in Section II-E. Training curves for DQN (Fig. 3) show high variance in returns across episodes, with large swings between comparatively good and poor rollouts.

Despite the more sample efficient off policy updates, the final DQN policy evaluated on ten episodes matches the A2C score with an average reward of -332, suggesting that this configuration is unable to substantially outperform the actor critic baseline in this domain.

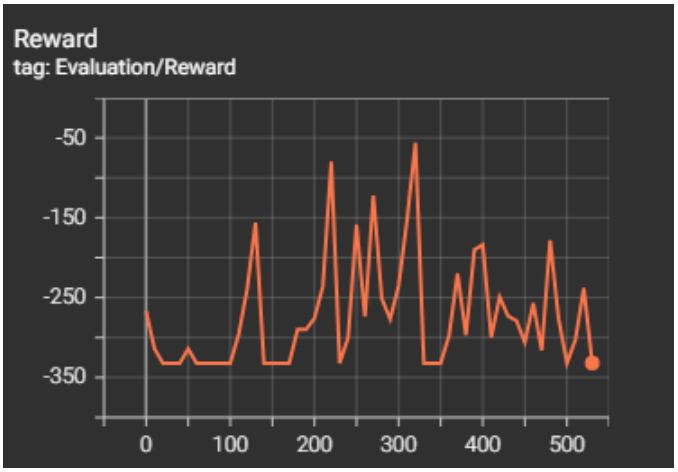

Fig. 4. PPO rewards using only default environment signals. The agent fails to reliably discover rewarding offensive strategies, resulting in stagnant performance.

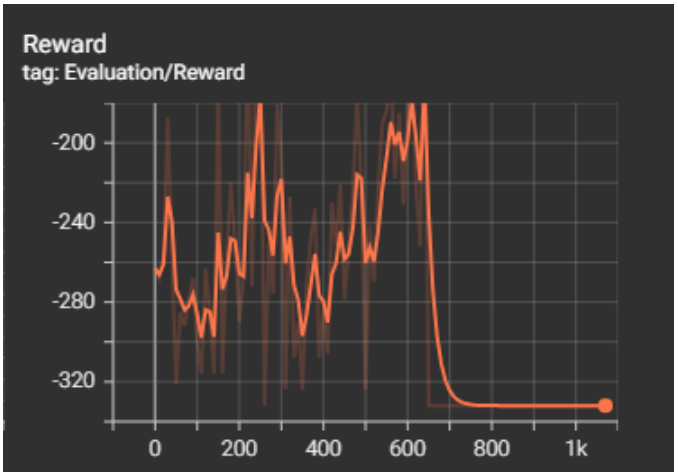

Fig. 5. PPO rewards with custom aggressive reward shaping. The curve shows a pronounced upward trend and increased variance, consistent with a more actively engaging combat style.

C. PPO Performance

1) Phase A: Baseline PPO with Default Rewards: With the unmodified environment reward, the PPO agent struggles to discover high reward strategies. The learning curve remains relatively flat, and no clear upward trend in episodic returns is observed.

Qualitative inspection indicates a prevalence of safe but unproductive behaviors, such as excessive movement or blocking without committing to attacks.

2) Phase B: PPO with Aggressive Reward Shaping: Introducing the aggression oriented reward shaping produces a marked behavioral shift. The agent learns to prioritize attack actions, exhibiting sustained offensive pressure rather than passive circling or disengagement. Average reward per episode increases substantially, and the variance of returns also rises, reflecting more frequent engagement and higher risk strategies.

The shaped rewards effectively communicate that taking offensive actions even when they do not immediately result in damage is preferable to inert behavior, thereby accelerating the discovery of successful attack patterns.

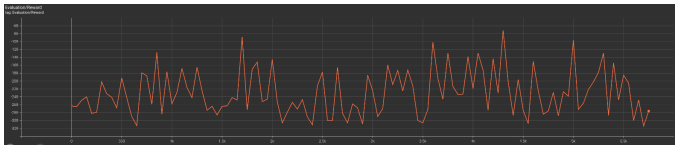

Fig. 6. Illustration of a local minimum in PPO training. After an initial improvement phase, the reward curve flattens, suggesting that the agent has converged to a sub optimal but stable strategy and ceased meaningful exploration.

TABLE I
EVALUATION PERFORMANCE OF A2C, DQN, AND PPO AGENTS ON THEIR RESPECTIVE HELD OUT RUNS. PPO ATTAINS A SUBSTANTIALLY HIGHER (LESS NEGATIVE) REWARD THAN THE A2C AND DQN BASELINES, WHICH REMAIN STUCK NEAR -332.

Algorithm	Episodes	Mean reward
A2C	10	-332.0
DQN	10	-332.0
PPO	5	-233.6

3) Phase C: Local Minima and Policy Collapse: Despite the substantial gains introduced by reward shaping, extended PPO training reveals a tendency to converge to local minima. In these regimes, the policy locks into a narrow subset of actions often a repetitive attack or simple exploit that yields modest but reliable rewards while avoiding risk.

This behavior highlights a trade off: aggressive shaping accelerates learning and improves peak performance, but if not combined with mechanisms that sustain exploration, it can encourage the consolidation of brittle, locally optimal tactics.

D. Quantitative Comparison Across Algorithms

To compare the three agents directly, we summarize the small held out evaluation runs stored in each implementation's `results.csv` file. Table I reports the mean reward over evaluation episodes; higher values correspond to better performance.

These results agree with the qualitative trends in the learning curves: A2C provides a stable but limited baseline, DQN in this configuration fails to improve beyond that baseline, and PPO with aggressive reward shaping achieves the best episodic returns at the cost of increased variance.

V. DISCUSSION

A. Architectural Considerations

The shared dual input architecture proves effective for both A2C and PPO. By fusing CNN based visual features with compact state descriptors, the agents can exploit explicit information about health, time, and aggressor levels while retaining the expressive power of visual perception for spatial reasoning. This design reduces the burden on the visual encoder to infer quantities that are already explicitly available from the game engine, which likely contributes to the stability observed in

A2C and the rapid behavioral changes under shaped rewards in PPO.

B. Algorithmic Trade offs: A2C, DQN, and PPO

The experiments reveal complementary strengths and weaknesses of the three agents in the UMK3 domain:

- **A2C:** Offers stable learning dynamics, lower sensitivity to hyperparameters, and reduced risk of catastrophic policy updates, but achieves slower performance gains and a lower asymptotic performance ceiling compared to the best PPO configuration.
- **DQN:** Reuses the same architecture with a value based objective and replay buffer, but in this sparse reward setting struggles to move beyond the A2C baseline, remaining stuck near a mean evaluation reward of -332 despite high variance during training.
- **PPO:** Provides higher potential performance and faster adaptation when combined with task specific reward shaping, but exhibits greater susceptibility to local minima and policy collapse in the absence of sustained exploration incentives.

These observations are consistent with the broader RL literature, where PPO is often favored for its performance but requires careful tuning and monitoring, while A2C (and simple DQN baselines) serve as robust but comparatively conservative reference points.

C. Role of Reward Shaping

Reward shaping is pivotal in this setting. The default DIAMBRA reward signal does not strongly penalize passivity nor sufficiently reward attempts at aggression, leading PPO to converge to conservative behaviors. The introduced shaping strategy directly encodes the design preference for aggressive play:

- Doubling positive rewards increases the learning signal associated with successful hits.
- The fixed aggression bonus for attack actions biases the policy toward exploring offensive action sequences.

However, this comes with caveats. If the aggression bonus is too large relative to other reward components, the agent may exploit it by spamming a narrow set of attacks, ignoring more nuanced tactics such as spacing, baiting, or defensive reads. The observed local minima are an instance of this phenomenon.

D. Mitigating Local Minima

The local minima encountered in PPO suggest several potential improvements:

- **Dynamic entropy scaling:** Increasing entropy regularization when the policy becomes too deterministic could encourage continued exploration and reduce premature convergence.
- **Curiosity or novelty bonuses:** Intrinsic motivation signals based on state visitation novelty or prediction error could incentivize the agent to seek new state action configurations rather than repeating a single exploit.
- **Action diversity regularization:** Explicit penalties for overly concentrated action distributions or repeated action patterns might encourage richer strategies.

These techniques could be especially valuable in complex fighting games where narrow exploitative behaviors are easy to discover but hard to escape once reinforced.

VI. Conclusion and Future Work

This report has examined the training of deep RL agents for UMK3 using A2C, DQN, and PPO with a shared dual input architecture. The main findings are:

1) The dual stream architecture that fuses visual frames with scalar game state features is effective for learning combat policies in UMK3.
2) A2C provides stable learning and a reliable baseline, but its performance ceiling is limited relative to PPO.
3) Under the same architecture and hyperparameters, the simple DQN agent fails to improve beyond the A2C baseline and remains at a mean evaluation reward of roughly -332, highlighting the difficulty of applying vanilla value based methods in this sparse reward, high variance domain.
4) PPO, when combined with aggression oriented reward shaping, yields substantially more proactive combat behavior and higher episodic rewards than either A2C or DQN.
5) The same reward shaping that accelerates PPO learning also increases the risk of convergence to local minima, where the agent exploits narrow, sub optimal strategies.

Future work should focus on integrating dynamic exploration mechanisms such as adaptive entropy, curiosity driven bonuses, or diversity aware regularization to preserve the benefits of aggressive reward shaping while mitigating policy collapse. Additional directions include training against a wider variety of opponents and difficulty levels, extending the action space to incorporate more complex combos, and conducting ablation studies on the relative contributions of visual versus scalar inputs in the shared architecture.