**1. Metriche di Valutazione (eval/):**

Queste metriche sono calcolate utilizzando l’EvalCallback e forniscono informazioni sulle prestazioni dell’agente durante le fasi di valutazione.

• **mean\_ep\_length (Lunghezza media dell’episodio):** Indica la durata media degli episodi durante la valutazione. Una lunghezza costante o decrescente, in combinazione con ricompense crescenti, può suggerire che l’agente sta diventando più efficiente nel raggiungere gli obiettivi prefissati.

• **mean\_reward (Ricompensa media per episodio):** Rappresenta la ricompensa media ottenuta per episodio durante la valutazione. Un trend ascendente in questa metrica è un indicatore chiave che l’agente sta migliorando le sue prestazioni complessive.

• **success\_rate (Tasso di successo):** Misura la percentuale di episodi in cui l’agente ha raggiunto con successo l’obiettivo. Un valore di 1.0 indica un successo del 100%. Per calcolare questa metrica, il dizionario info dell’ambiente deve contenere una chiave is\_success.

**2. Metriche di Rollout (rollout/):**

Queste metriche forniscono informazioni sulle prestazioni dell’agente durante l’interazione con l’ambiente nel processo di addestramento.

• **ep\_len\_mean (Lunghezza media dell’episodio):** Indica la lunghezza media degli episodi durante l’addestramento, calcolata su una finestra di episodi recenti (100 per impostazione predefinita). Fluttuazioni in questa metrica possono riflettere cambiamenti nel comportamento dell’agente, come una maggiore efficienza nel raggiungere obiettivi o una tendenza a fallire più rapidamente.

• **ep\_rew\_mean (Ricompensa media per episodio):** Rappresenta la ricompensa media ottenuta per episodio durante l’addestramento, calcolata su una finestra di episodi recenti. Un aumento costante suggerisce che l’agente sta apprendendo una politica efficace. È necessario utilizzare un wrapper Monitor per calcolare questa metrica, che viene aggiunto automaticamente da funzioni come make\_vec\_env.

• **exploration\_rate (Tasso di esplorazione):** Indica il valore corrente del tasso di esplorazione quando si utilizza DQN. Corrisponde alla frazione di azioni eseguite in modo casuale (epsilon dell’esplorazione “epsilon-greedy”). Un tasso di esplorazione decrescente nel tempo è tipico, poiché l’agente passa gradualmente dall’esplorazione allo sfruttamento delle conoscenze acquisite.

**3. Metriche Temporali (time/):**

Queste metriche forniscono informazioni sul tempo e sul progresso dell’addestramento.

• **episodes (Numero totale di episodi):** Indica il numero totale di episodi completati dall’inizio dell’addestramento. Un numero crescente riflette il progresso dell’addestramento.

• **fps (Frame per secondo):** Misura il numero di frame elaborati al secondo, includendo il tempo impiegato per l’aggiornamento dei gradienti. Un valore stabile suggerisce che l’addestramento procede senza colli di bottiglia significativi.

• **iterations (Numero di iterazioni):** Indica il numero di iterazioni completate, dove un’iterazione tipicamente comprende la raccolta dei dati e l’aggiornamento della politica (ad esempio, per algoritmi come A2C/PPO).

• **time\_elapsed (Tempo trascorso):** Mostra il tempo totale trascorso dall’inizio dell’addestramento, espresso in secondi.

• **total\_timesteps (Numero totale di timestep):** Indica il numero totale di timestep eseguiti nell’ambiente. Un valore crescente riflette l’avanzamento dell’addestramento.

**4. Metriche di Addestramento (train/):**

1. **approx\_kl (Divergenza KL approssimativa):**

• **Descrizione:** Misura la divergenza media di Kullback-Leibler tra la vecchia e la nuova politica.

• **Interpretazione:** Valori troppo alti possono indicare che la nuova politica si discosta eccessivamente dalla precedente, mentre valori troppo bassi potrebbero suggerire aggiornamenti insufficienti.

• **Valori Tipici:** Per PPO, un valore intorno a 0.01-0.05 è generalmente considerato appropriato.

2. **clip\_fraction (Frazione di clipping):**

• **Descrizione:** Indica la percentuale di aggiornamenti della perdita surrogata che sono stati limitati (clipped) durante l’addestramento.

• **Interpretazione:** Una frazione di clipping elevata potrebbe suggerire che molti aggiornamenti stanno subendo il clipping, il che potrebbe limitare l’apprendimento.

• **Valori Tipici:** Un valore intorno al 0.1-0.3 è spesso osservato durante un addestramento stabile.

3. **clip\_range (Intervallo di clipping):**

• **Descrizione:** Rappresenta il valore corrente dell’intervallo di clipping utilizzato per la perdita surrogata.

• **Interpretazione:** Questo valore è un iperparametro impostato prima dell’addestramento e determina quanto la politica può deviare durante gli aggiornamenti.

• **Valori Tipici:** Per PPO, un valore comune è 0.2.

4. **entropy\_loss (Perdita di entropia):**

• **Descrizione:** Misura l’entropia media della politica; valori più alti indicano maggiore casualità nelle azioni dell’agente.

• **Interpretazione:** Un’entropia elevata favorisce l’esplorazione, mentre un’entropia bassa indica che l’agente sta sfruttando maggiormente le conoscenze acquisite.

• **Valori Tipici:** I valori variano in base all’ambiente e all’algoritmo, ma una diminuzione graduale dell’entropia durante l’addestramento è comune, poiché l’agente passa dall’esplorazione allo sfruttamento.

5. **explained\_variance (Varianza spiegata):**

• **Descrizione:** Indica la frazione della varianza del ritorno spiegata dalla funzione di valore; un valore di 1.0 indica una perfetta predizione.

• **Interpretazione:** Valori vicini a 1.0 suggeriscono che la funzione di valore sta prevedendo accuratamente le ricompense future, mentre valori negativi o vicini a zero indicano scarse prestazioni predittive.

• **Valori Tipici:** Un valore superiore a 0.5 è generalmente desiderabile, ma questo può variare in base all’ambiente.

6. **learning\_rate (Tasso di apprendimento):**

• **Descrizione:** Il tasso al quale il modello aggiorna i suoi parametri durante l’addestramento.

• **Interpretazione:** Un tasso di apprendimento troppo alto può causare instabilità, mentre uno troppo basso può rallentare l’apprendimento.

• **Valori Tipici:** Per PPO, valori comuni sono nell’intervallo 2.5e-4 - 3e-4.

7. **loss (Perdita totale):**

• **Descrizione:** Rappresenta la somma delle diverse componenti di perdita che l’algoritmo sta cercando di minimizzare.

• **Interpretazione:** Una diminuzione costante della perdita totale suggerisce che l’algoritmo sta convergendo. Tuttavia, fluttuazioni sono comuni nel reinforcement learning.

• **Valori Tipici:** I valori variano ampiamente; è importante osservare la tendenza generale piuttosto che un valore specifico.

8. **n\_updates (Numero di aggiornamenti):**

• **Descrizione:** Indica il numero di aggiornamenti dei gradienti effettuati fino a quel momento.

• **Interpretazione:** Un numero crescente riflette il progresso dell’addestramento.

• **Valori Tipici:** Dipende dalla configurazione dell’addestramento, inclusi il numero totale di timestep e la frequenza degli aggiornamenti.

9. **policy\_gradient\_loss (Perdita del gradiente della politica):**

• **Descrizione:** Misura l’errore associato all’aggiornamento della politica dell’agente.

• **Interpretazione:** Valori negativi indicano che l’aggiornamento sta procedendo nella direzione corretta per massimizzare la ricompensa attesa.

• **Valori Tipici:** I valori variano; l’importante è monitorare la tendenza generale piuttosto che un valore specifico.

10. **std (Deviazione standard):**

• **Descrizione:** Indica la deviazione standard delle azioni campionate dalla politica, riflettendo il livello di esplorazione.

• **Interpretazione:** Valori più alti indicano maggiore esplorazione, mentre valori più bassi suggeriscono che l’agente sta diventando più deterministico.