Guía de Escalado de Parámetros para Diferentes Tamaños de Grafo

1. Introducción

Este documento explica cómo ajustar los parámetros del entrenamiento PPO (*Proximal Policy Optimization*) al cambiar el tamaño del grafo, asegurando un aprendizaje estable y eficiente del agente.

2. Tabla de Referencia Rápida

Tamaño Grafo	Nodos	max_steps	total_timesteps	eval_freq	learning_rate	ent_coef
5x5	25	30	200,000	5,000	3e-4	0.05
10x10	100	80	500,000	10,000	3e-4	0.05
20x20	400	200	1,000,000	20,000	1e-4	0.07
25x25	625	250	1,500,000	25,000	1e-4	0.07
50x50	2,500	600	3,000,000	50,000	5e-5	0.1

3. Parámetros Clave y Cómo Escalarlos

3.1. max_steps (Límite de pasos por episodio)

$$\text{max\_steps} \approx n\_nodes \times (0{,}3 \text{ a } 0{,}4)$$

Más nodos implican que el agente necesita más pasos para alcanzar los waypoints y evitar ciclos. **Ejemplos:**

- 5x5 (25 nodos): $25 \times 1{,}2 = 30$
- 20x20 (400 nodos): $400 \times 0{,}5 = 200$
- 50x50 (2500 nodos): $2500 \times 0{,}24 = 600$

3.2. total_timesteps (Duración del entrenamiento)

$$\text{total\_timesteps} \approx n\_nodes \times (1000 \text{ a } 2000)$$

Más nodos = espacio de estados más grande, por lo que se necesita más experiencia para generalizar. **Ejemplos:**

- 5x5: $25 \times 8,000 = 200,000$
- 20x20: $400 \times 2,500 = 1,000,000$
- 50x50: $2,500 \times 1,200 = 3,000,000$

3.3. eval_freq (Frecuencia de evaluación)

$$\text{eval\_freq} \approx \frac{\text{total\_timesteps}}{40}$$

Se recomienda realizar aproximadamente 40 evaluaciones durante el entrenamiento. **Ejemplos:**

- 5x5: $200,000/40 = 5,000$
- 20x20: $1,000,000/40 = 25,000$
- 50x50: $3,000,000/40 = 75,000$

3.4. learning_rate (Tasa de aprendizaje)

- Grafos pequeños (menos de 100 nodos): 3×10^{-4}
- Grafos medianos (100-500 nodos): 1×10^{-4} a 3×10^{-4}
- Grafos grandes (mas de 500 nodos): 5×10^{-5} a 1×10^{-4}

Comenzar con 3e-4 y reducir si el entrenamiento es inestable.

3.5. ent_coef (Coeficiente de entropía / exploración)

- Grafos pequeños: 0.05
- Grafos medianos: 0.05 - 0.07
- Grafos grandes: 0.07 - 0.1

Aumentar si el agente se queda en ciclos o reducir si explora demasiado.

3.6. gamma y gae_lambda (Factores de descuento)

```
gamma = 0.99
gae_lambda = 0.95
```

Valores que funcionan bien para la mayoría de los casos.

3.7. max_no_improvement_evals (Early stopping)

$$\text{max\_no\_improvement\_evals} = 10 \text{ a } 15$$

4. Señales de Que Necesitas Ajustar

4.1. Problemas comunes

- **El agente no aprende nada:** recompensa negativa constante. Aumentar `max_steps` o `ent_coef`.
- **Aprende pero muy lento:** aumentar `total_timesteps` o `learning_rate`.
- **Entrenamiento inestable:** reducir `learning_rate` o `clip_range`.
- **Se queda en ciclos:** aumentar `ent_coef` o reducir `max_steps`.

Señales de buen entrenamiento:

- Recompensa aumenta consistentemente.
- Early stopping actúa tras 50–100k pasos.
- El agente completa los waypoints en ¡30 % de los pasos máximos.

5. Configuración del Modelo PPO

Listing 1: Configuración del modelo PPO

```
model = PPO(
    "MlpPolicy",
    env,
    learning_rate=3e-4,       # velocidad de aprendizaje
    clip_range=0.2,           # l mite de cambios dr sticos
    ent_coef=0.05,            # exploraci n
    gamma=0.99,               # descuento de recompensas futuras
    gae_lambda=0.95,          # ventaja estimada
    verbose=1
)
```

6. Explicación de los parámetros

- **"MlpPolicy"**: Red neuronal multicapa que recibe el estado y devuelve probabilidades de acciones.
- **env**: Entorno donde el agente interactúa.
- **learning_rate**: Velocidad de actualización de los pesos.
- **clip_range**: Limita cambios bruscos en la política.
- **ent_coef**: Controla exploración vs explotación.
- **gamma**: Factor de descuento de recompensas futuras.
- **gae_lambda**: Cálculo de ventaja para evaluar acciones.
- **verbose**: Nivel de información impresa durante el entrenamiento.

7. Checklist de Escalado

- Ajustar `grid_size`
- Calcular nuevo `max_steps`
- Ajustar `total_timesteps`
- Actualizar `eval_freq`
- Revisar `learning_rate` y `ent_coef`
- Verificar distribución de waypoints
- Probar con pocos episodios antes del entrenamiento completo