Le Robot Joueur de Basket

Introduction

Imagine un **robot joueur de basket** qui apprend à jouer dans une équipe. Son objectif: **marquer des points pour gagner des matchs.**

1 Les deux réseaux de neurones dans le contexte du basket

1. **Le réseau de politique π_θ : Le décideur stratégique** Ce réseau décide **quelle action prendre** à chaque instant, par exemple:
   - Passer le ballon,
   - Tirer,
   - Dribbler,
   - Reculer ou défendre.
2. **Le réseau de valeur V_ϕ : L'évaluateur de situation** Ce réseau regarde **la situation globale** sur le terrain et se demande:

   "Est-ce que cette position est avantageuse pour marquer ou défendre?"

Ces deux réseaux travaillent ensemble comme un joueur intelligent:

- Le réseau de politique agit en temps réel en prenant des décisions.
- Le réseau de valeur évalue si l'équipe est dans une bonne ou une mauvaise position.

2 Comment ça marche pendant un match?

Étape 1 : Observation (entrées des réseaux)

Le robot analyse l'état du jeu, par exemple:

- Où est le ballon ?
- Où sont les coéquipiers et adversaires?

- Quelle est la distance jusqu'au panier?
- Y a-t-il un défenseur proche?

Ces informations sont transformées en **données numériques** (par exemple, des coordonnées sur le terrain) qui sont données en entrée aux deux réseaux.

Étape 2 : Le réseau de politique décide quoi faire

Le réseau de politique, c'est **le cerveau qui agit**. Il prend l'état du jeu comme entrée et propose des actions avec des probabilités:

- Passer le ballon : 60 %,
- Tirer : 30 %,
- Dribbler : 10 %.

Exemple concret : Si le robot est à 3 mètres du panier avec un défenseur devant lui, le réseau pourrait dire :

- Passer à un coéquipier libre est l'option la plus sûre (probabilité : 80 %).
- Tirer directement est risqué mais possible (probabilité : 20 %).

Le robot choisit ensuite **l'action avec la probabilité la plus élevée** (passer, ici).

Étape 3 : Le réseau de valeur juge la situation

Le réseau de valeur, c'est **le conseiller stratégique**. Il regarde l'état global et attribue une note à la situation:

- "Cet état est prometteur, car tu es proche du panier avec un bon alignement (score : 8/10)."
- "Cet état est mauvais, car ton équipe est désorganisée (score : 2/10)."

Exemple concret :

- Si le robot a le ballon près du panier et qu'il est seul, le réseau pourrait dire : *"Situation très favorable (note : +10)."*
- Si le robot est encerclé par trois défenseurs : *"Situation très défavorable (note : -5)."*

3 Comment les réseaux apprennent-ils ?

Au début, le robot est un joueur débutant.

Le robot peut dribbler sans réfléchir, tirer quand il est trop loin du panier ou oublier ses coéquipiers. Mais il **apprend de ses erreurs** grâce à PPO.

1. **Apprentissage du réseau de politique π_θ :** Si passer mène à un panier, le réseau apprend que:

   "Passer est une bonne action, augmente sa probabilité."

   Si tirer mène à un échec, il apprend que:

   "Réduis la probabilité de tirer dans cette situation."

2. **Apprentissage du réseau de valeur V_ϕ :** Si le réseau dit que la situation est favorable mais que l'équipe perd, il apprend qu'il a **surestimé** l'état. Il ajuste sa note pour être plus précis à l'avenir.

4 Pourquoi PPO est parfait pour un joueur de basket?

Le clipping pour des ajustements progressifs

Imagine que le robot décide soudainement de tout changer et commence à tirer tout le temps sans raison. Cela pourrait le rendre instable. PPO utilise le **clipping** pour dire:

"OK, teste une nouvelle stratégie, mais fais-le doucement. Continue de jouer de façon cohérente."

L'entropie pour l'exploration

Parfois, le robot doit tester des actions qu'il ne prend pas souvent (par exemple, dribbler au lieu de passer). PPO ajoute une **pénalité d'entropie** pour encourager ce genre d'exploration.

"Et si je dribblais cette fois? Peut-être que ça ouvrirait une nouvelle opportunité."

5 Résumé de la métaphore du robot joueur de basket

1. **Réseau de politique π_θ :** Décide **quelle action prendre** (passer, tirer, dribbler, etc.) en fonction de l'état actuel et propose les actions avec des probabilités.

2. **Réseau de valeur V_ϕ :** Juge si la situation globale est favorable ou non, aidant le robot à comprendre si ses actions mènent à une victoire ou un échec.

3. **Pourquoi PPO est puissant?**

   - Il aide le robot à apprendre **progressivement** (grâce au clipping).
   - Il pousse le robot à **explorer** des actions moins courantes (grâce à l'entropie).
   - Il combine les deux réseaux pour prendre de meilleures décisions tout en comprenant le contexte du jeu.