PPO - Cas réels

Introduction

Prenons un exemple **réel et sérieux** où PPO et ses réseaux de neurones sont utilisés dans l'industrie. PPO est largement adopté dans plusieurs domaines comme les **véhicules autonomes**, la **robotique**, et même l'optimisation dans les **data centers**. Voici un cas concret qui met en valeur son application.

1 Exemple réel : PPO dans les véhicules autonomes

Imagine une entreprise qui développe un **véhicule autonome**, une voiture capable de conduire sans intervention humaine. Le rôle de PPO est d'entraîner le système de la voiture pour qu'elle apprenne à conduire **en toute sécurité** et **efficacement** dans un environnement complexe.

1.1 Les deux réseaux dans ce contexte

1. **Le réseau de politique π_θ : Le conducteur**
   Il décide **quelle action effectuer** à chaque instant, par exemple :
   - Tourner le volant à gauche ou à droite,
   - Freiner ou accélérer,
   - Changer de voie ou rester dans la même.
2. **Le réseau de valeur V_ϕ : L'analyste de situation**
   Il évalue **si la situation actuelle est favorable ou non**. Par exemple :
   - Est-ce que la voiture est dans une bonne position par rapport aux autres véhicules ?
   - Est-elle à une distance sécuritaire des obstacles ?

Ces deux réseaux travaillent ensemble pour permettre à la voiture de conduire intelligemment.

1.2 Comment PPO fonctionne dans une voiture autonome ?

Étape 1 : Observation

La voiture analyse son environnement grâce à ses capteurs (caméras, lidar, radar). Ces capteurs fournissent des informations comme:

- La position des autres voitures,
- La distance jusqu'à l'obstacle le plus proche,
- La vitesse actuelle et la limite de vitesse,
- La courbure de la route.

Toutes ces données sont converties en **entrées numériques** pour les deux réseaux.

Étape 2 : Décision via le réseau de politique

Le réseau de politique π_θ utilise ces observations pour proposer des actions. Par exemple :

- **Tourner le volant** : 40 %,
- **Freiner légèrement** : 30 %,
- **Accélérer doucement** : 30 %.

La voiture exécute ensuite l'action avec la probabilité la plus élevée (par exemple, tourner le volant).

Étape 3 : Évaluation via le réseau de valeur

Le réseau de valeur V_ϕ regarde l'état global et évalue si la situation est bonne ou mauvaise. Par exemple :

- Si la voiture est bien centrée dans sa voie avec une distance sécuritaire des autres véhicules : **"Situation favorable, score : +8."**
- Si la voiture est trop proche d'un obstacle ou dévie de sa trajectoire : **"Situation dangereuse, score : -5."**

2 Comment PPO entraîne les réseaux pour s'améliorer ?

2.1 Au début : L'apprentissage par essais et erreurs

La voiture commence en mode "débutant" et fait beaucoup d'erreurs :

- Elle accélère trop près d'un virage serré,

- Elle freine brusquement quand ce n'est pas nécessaire,
- Elle dévie de sa trajectoire dans un rond-point.

Avec PPO, la voiture **apprend de ses erreurs** :

- Si une action (comme tourner brusquement) provoque une collision virtuelle, le réseau de politique apprend à **réduire la probabilité de cette action**.
- Si une action (comme ralentir doucement avant un virage) améliore la sécurité, le réseau apprend à **favoriser cette action**.

3 Pourquoi PPO est idéal pour les véhicules autonomes ?

3.1 Clipping pour éviter des changements brutaux

Imagine que la voiture apprend soudainement une nouvelle stratégie et commence à freiner brusquement tout le temps. PPO empêche cela en limitant les ajustements grâce au **clipping** :

> **"OK, teste une nouvelle approche, mais garde une conduite fluide."**

3.2 Encourager l'exploration

PPO pousse la voiture à tester des actions qu'elle n'aurait pas essayées autrement :

> **"Et si je changeais de voie plus tôt dans un embouteillage pour voir si cela accélère le trajet ?"**

Cette exploration permet à la voiture de découvrir de **nouvelles stratégies optimales**.

4 Un scénario réel : Conduire dans un embouteillage

Imagine une voiture autonome qui doit naviguer dans un embouteillage. Voici comment PPO gère la situation :

Le réseau de politique π_θ :

- **Entrée :**
  - Position des voitures proches,
  - Vitesse actuelle,
  - Distance jusqu'au véhicule devant.
- **Sortie :**

  - Ralentir : 70 %,
  - Changer de voie : 20 %,
  - Accélérer : 10 %.

Le réseau de valeur V_ϕ :

- **Entrée :**
  - Situation actuelle (embouteillage, distance sécuritaire, etc.).
- **Sortie :**
  - Une évaluation : **"Cet état vaut +5, car il est sécurisé et stable."**

5 Où PPO est-il réellement utilisé dans l'industrie?

- **Waymo (Google) :** PPO est utilisé pour entraîner des voitures autonomes à gérer des situations complexes.
- **Tesla Autopilot :** PPO ou des algorithmes similaires aident à affiner les décisions en temps réel.
- **Optimisation des data centers (Google) :** PPO optimise la consommation d'énergie dans les data centers.
- **Robotique industrielle :** PPO entraîne des bras robotisés à effectuer des tâches complexes.

Résumé

- PPO est idéal pour les systèmes dynamiques et complexes nécessitant des décisions rapides et précises.
- Les réseaux π_θ et V_ϕ travaillent ensemble pour garantir des décisions intelligentes et sécurisées.
- PPO garantit fluidité, stabilité et exploration pour découvrir des stratégies optimales.