Introduction à PPO et aux Réseaux de Neurones

1 C'est quoi un réseau de neurones?

Un réseau de neurones, c'est comme un **cerveau simplifié** dans un ordinateur. Il est composé de **petites unités**, appelées **neurones** (inspirés du cerveau humain). Ces neurones travaillent ensemble pour **prendre des décisions** ou **faire des prédictions**.

Comment ça marche?

- **Entrée :** Tu lui donnes une information (par exemple, "Est-ce que le bâton penche à droite?").
- **Traitement :** Les neurones dans le réseau réfléchissent ensemble et discutent :
  - **Premier neurone :** "Ah, le bâton penche un peu."
  - **Deuxième neurone :** "Ça veut dire qu'on devrait peut-être aller à droite!"
- **Sortie :** Le réseau donne une réponse, par exemple :
  - "Va à droite!" avec 80 % de certitude.
  - "Va à gauche!" avec 20 % de certitude.

Un réseau de neurones est donc **une machine à prendre des décisions**, qui apprend en répétant et en s'améliorant avec le temps.

2 Pourquoi PPO a besoin de deux réseaux?

Dans PPO, on utilise **deux réseaux de neurones différents**. Imagine que ton robot est un **joueur de jeu vidéo**. Ces deux réseaux l'aident de différentes manières :

1. **Le réseau de politique (π_θ) :**
   - C'est le **cerveau stratégique** du robot.
   - Il décide **quelle action prendre** dans chaque situation.
   - Exemple : "Si le bâton penche à droite, je vais à droite!"
2. **Le réseau de valeur (V_ϕ) :**

- C'est le **cerveau de l'évaluation**.
- Il dit **à quel point la situation actuelle est bonne ou mauvaise**.
- Exemple : "Est-ce que je suis dans une position favorable pour garder le bâton en équilibre?"

Ces deux réseaux travaillent ensemble :

- Le réseau de politique **agit**.
- Le réseau de valeur **évalue** si l'action et la situation sont bonnes ou mauvaises.

3 Comment fonctionne le réseau de politique π_θ ?

Le réseau de politique est **comme un assistant qui prend des décisions**. Il utilise ce qu'il voit pour **proposer la meilleure action**.

Imagine un jeu vidéo comme CartPole :

1. **Entrée du réseau :** Le réseau regarde l'état actuel. Par exemple :
   - La position du chariot (gauche ou droite).
   - L'angle du bâton (penché ou droit).
   - La vitesse du chariot.
   - La vitesse du bâton. (C'est comme lui montrer une photo du jeu à un instant précis.)
2. **Neurones cachés :** À l'intérieur, des centaines de neurones réfléchissent :
   - "Si le bâton penche à droite et que le chariot va vite, il faudrait peut-être ralentir!"
   - Ils discutent et analysent pour comprendre ce qui se passe.
3. **Sortie du réseau :** Le réseau donne une réponse comme :
   - "Probabilité d'aller à gauche : 30 %."
   - "Probabilité d'aller à droite : 70 %."

Le robot choisit ensuite l'action avec la plus haute probabilité.

4 Comment fonctionne le réseau de valeur V_ϕ ?

Le réseau de valeur est **comme un conseiller**. Il dit au robot :

"Est-ce que la situation est bonne pour moi ou pas?"

Toujours dans le jeu CartPole :

1. **Entrée du réseau :** Comme pour le réseau de politique, il regarde l'état actuel (la position du chariot, l'angle du bâton, etc.).
2. **Neurones cachés :** Ces neurones réfléchissent et analysent :
   - "Si le bâton est droit, c'est une bonne position."
   - "Mais si le chariot est trop près du bord, ce n'est pas idéal."
3. **Sortie du réseau :** Le réseau donne une **valeur numérique unique**. Exemple :
   - "Cet état vaut 50 points (c'est une situation prometteuse)."
   - "Cet état vaut -10 points (c'est une situation dangereuse)."

Ce réseau aide le robot à comprendre si ses actions l'amènent dans une bonne direction.

5 Comment ces réseaux apprennent-ils ?

Ces deux réseaux sont comme des élèves dans une école : ils **apprennent en faisant des erreurs** et en corrigeant leurs réponses.

1. **L'apprentissage du réseau de politique :**
   - Si le robot prend une bonne action (par exemple, aller à droite pour rattraper le bâton), le réseau apprend à **augmenter la probabilité** de cette action.
   - Si l'action était mauvaise (par exemple, aller à gauche quand il fallait aller à droite), le réseau apprend à **réduire la probabilité** de cette action.
2. **L'apprentissage du réseau de valeur :**
   - Si le réseau dit qu'un état vaut "50 points", mais que le robot finit par perdre, il comprend qu'il a **surestimé** cet état.
   - Il ajuste ses prédictions pour être plus précis à l'avenir.

Les deux réseaux sont entraînés avec une méthode appelée **rétropropagation** (ou back-propagation). C'est comme leur donner un cours particulier chaque fois qu'ils se trompent.

6 Pourquoi ce sont des réseaux de neurones?

Ces deux réseaux sont des **réseaux de neurones** parce qu'ils doivent gérer des situations complexes. Par exemple :

- Dans le jeu CartPole, il y a des **milliers de combinaisons possibles** entre la position, l'angle et la vitesse.

- Les réseaux de neurones sont capables de :
  - **Comprendre des motifs** complexes (comme "Si le bâton penche ET le chariot va à droite, c'est dangereux").
  - **Généraliser** pour prendre des décisions dans des situations nouvelles qu'ils n'ont jamais vues.

Résumé pour débutants complets

1. PPO utilise **deux réseaux de neurones intégrés** :
   - **Le réseau de politique (π_θ)** : Il décide quelle action prendre dans chaque situation.
   - **Le réseau de valeur (V_ϕ)** : Il prédit à quel point la situation actuelle est bonne.
2. Ces réseaux sont comme des **élèves** :
   - Ils apprennent en jouant et en corrigeant leurs erreurs.
   - Plus ils jouent, plus ils deviennent intelligents.
3. **Pourquoi des réseaux de neurones ?**
   - Parce qu'ils peuvent comprendre des motifs complexes et prendre des décisions dans des environnements compliqués.