Université Tunis-elManar

Faculté des Sciences de Tunis

Projet d'Apprentissage par Renforcement

Réalisateurs :
Bouzidi Mariam
Bouhouch Eya
Lajmi Sahar

Encadrant :
Prof. Samir Toumi

Date de Réalisation : April 21,2024

Sommaire

7. **Défis et Complexités**
   (a) L'exploration-exploitation
   (b) La gestion des etats continus
8. **Avenir de l'AR**
   (a) Tendances Emergentes
   (b) Défis Actuels et Futurs
   (c) Ethique et Interprétabilité
   (d) Adaptation a l'Adversité et a la Dynamique
   (e) Applications Sociales et Economiques
9. **Conclusion**

Introduction

Definition

L'apprentissage par renforcement (RL), désigné comme apprentissage par récompense, se situe au cœur de l'intelligence artificielle, définissant la manière dont un agent intelligent doit prendre des décisions dans le but de maximiser une récompense cumulative au sein d'un environnement complexe et souvent changeant. Cette discipline repose sur un processus d'interaction dynamique entre l'agent, doté de capacités décisionnelles, et son environnement. À travers cette interaction continue, l'agent explore différentes actions, évalue les conséquences de ses choix et ajuste son comportement afin d'optimiser la réception de récompenses.

Distinction entre les types d'apprentissage

Bienque l'apprentissage supervisé implique l'utilisation d'un ensemble de données étiquetées pour entraîner un modèle à effectuer des prédictions ou à prendre des décisions et que l'apprentissage non supervisé consiste à découvrir des structures ou des motifs inhérents à un ensemble de données non étiquetées.l'apprentissage par renforcement repose sur un processus itératif d'interaction entre l'agent et l'environnement. Le caractère distinctif de l'apprentissage par renforcement réside dans l'absence de supervision explicite.Au contraire,l'agent apprend progressivement par essais et erreurs, en ajustant ses stratégies en fonction des récompenses obtenues au fil du temps.

Historique et évolution de dommaine

L'apprentissage par renforcement a connu une trajectoire fascinante, évoluant depuis ses fondements conceptuels dans les années 1950 jusqu'à devenir une force motrice majeure dans le domaine de l'intelligence artificielle (IA) aujourd'hui. Des avancées clés, telles que la formalisation théorique dans les années 1980 et l'intégration réussie de l'apprentissage profond dans les années 2010, ont propulsé l'AR vers des sommets impressionnants, comme en témoigne la victoire d'AlphaGo en 2016.

L'Ecosysteme de l'AR

Acteurs et Piliers de l'AR

- **Agent** : La partie autonome qui prend des décisions dans un environnement donné. L'agent cherche à apprendre une politique optimale pour maximiser les récompenses.
- **Environnement** : Le contexte dans lequel l'agent opère. C'est l'espace où les actions de l'agent ont des conséquences, et où les récompenses ou punitions sont générées.
- **Récompense** : Le signal fourni à l'agent pour évaluer la qualité de ses actions. Les récompenses guident l'agent vers des comportements souhaitables.

ondements Conceptuels de l'AR

- **État (State)** : La représentation du contexte dans lequel se trouve l'agent. Les états peuvent être observables ou non observables, et l'agent prend des décisions en fonction de ces états.
- **Action** : Les différentes actions que l'agent peut entreprendre dans un état donné. L'ensemble des actions possibles constitue l'espace d'action de l'agent.
- **Politique** : La stratégie que l'agent suit pour prendre des décisions. La politique peut être déterministe ou probabiliste, guidant les actions de l'agent en fonction de l'état actuel.

Processus Décisionnel de Markov

Définition

Un Processus de Décision de Markov (MDP) est un cadre mathématique utilisé pour modéliser et résoudre des problèmes de prise de décision séquentielle dans des environnements incertains. Les MDP sont largement utilisés en intelligence artificielle, en apprentissage automatique, en robotique et dans d'autres domaines où des agents autonomes doivent prendre des décisions dans des situations dynamiques. Un MDP est défini comme un tuple $\langle S, A, T, R, \gamma \rangle$ où :

- S : ensemble fini d'états.
- A : ensemble fini d'actions.
- T : fonction de transition.
- R : fonction de récompense.
- γ : facteur de rabais.

Fonction de Transition

La probabilité d'arriver dans l'état s' à l'instant $t + 1$ en prenant l'action a dans l'état s à l'instant t est définie par $P(s'|s, a)$. Cette propriété markovienne signifie que la probabilité dépend uniquement de l'état actuel et de l'action entreprise, pas de l'historique complet des états et actions précédents.

Fonction de Récompense

$R(s, a, s')$ attribue une récompense numérique à chaque transition état-action-état . L'objectif dans un MDP est de trouver une politique (π*) qui maximise la somme des récompenses attendues au fil du temps. Cela peut être formulé comme la maximisation de la fonction de valeur (Vπ*) ou de la fonction de valeur d'action (Qπ*).

Politique

Une politique π est une stratégie qui spécifie quelle action prendre dans chaque état. La politique peut être déterministe (une action spécifique dans chaque état) ou stochastique (une distribution d'actions dans chaque état).

La maximisation de la somme des récompenses

La maximisation de la somme des récompenses attendues au fil du temps est représentée par la formule suivante :

$$G_t = \sum_{k=1}^{\infty} \gamma^k r_{t+k+1}$$

Cela représente la somme des récompenses à chaque instant t, pondérée par le facteur de rabais γ élevé à la puissance k, où k varie de 1 à l'infini. Le facteur de rabais modélise la préférence de l'agent pour les récompenses immédiates par rapport aux récompenses futures. Il s'agit d'un nombre réel compris entre 0 et 1. Cette formule prend en compte les récompenses obtenues à chaque instant futur $t + k + 1$ et les réduit en fonction du facteur de rabais.

Fonctions de Valeur

Fonction de Valeur d'État (V)

Représente la valeur d'être dans un certain état sous une certaine politique.

$$V^{\pi}(s) = E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(s_t), s_{t+1}) \mid s_0 = s\right]$$

Elle est définie comme l'espérance des récompenses futures à partir d'un état s sous une politique donnée

Fonction de Valeur d'Action (Q)

Représente la valeur d'entreprendre une certaine action dans un certain état sous une certaine politique.

$$Q^{\pi}(s, a) = E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(s_t), s_{t+1}) \mid s_0 = s, a_0 = a\right]$$

Elle est définie comme l'espérance des récompenses futures à partir d'un état s en prenant l'action a sous une politique donnée

Fonctions Optimal (V^* et Q^*)

Définition

Les fonctions de valeur optimales (V^* et Q^*) représentent les valeurs optimales d'être dans un état ou d'entreprendre une action, respectivement, sous la politique optimale.

Notation

$V^*(s)$ et $Q^*(s, a)$.

Calcul

Elles sont définies de manière similaire aux fonctions de valeur, mais en considérant la politique optimale π^*.

Résolution

Pour trouver la politique optimale, on utilise la programmation dynamique pour résoudre les MDP discrets, où l'ensemble d'états et d'actions est fini. La résolution implique la résolution itérative des équations de Bellman.

Équations de Bellman

$$V^*(s) = \max_a Q^*(s, a)$$

indique que la valeur optimale d'un état est égale à la valeur maximale parmi toutes les actions possibles dans cet état

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} P(s'|s, a) V^*(s')$$

Pour déterminer la politique optimale π^*, vous pouvez choisir l'action a dans chaque état s de manière à maximiser $Q^*(s, a)$.

Algorithmes d'Apprentissage par Renforcement

1) Le Q-learning

Le Q-learning est développé pour résoudre des problèmes où un agent prend des décisions séquentielles pour maximiser la récompense cumulée. Il est basé sur la notion de Q-value, représentant la "valeur" d'une paire état-action. Formellement, le Q-value d'une paire (état, action), noté $Q(s, a)$, est défini comme :

$$Q(s, a) = R(s, a) + \gamma \cdot \max_{a'} Q(s', a')$$

Objectif du Q-learning

L'objectif est de trouver la fonction d'action optimale, où chaque paire état-action a la valeur maximale possible. L'algorithme explore l'environnement, ajuste ses estimations de Q-values en fonction des récompenses observées et converge vers la fonction d'action optimale.

Algorithme de Q-learning

L'algorithme utilise la règle de mise à jour basée sur la différence temporelle (TD) :

$$Q(s, a)_{new} = Q(s, a)_{old} + \alpha \cdot [R(s, a) + \gamma \cdot \max_{a'} Q(s', a') - Q(s, a)_{old}]$$

Où α est le taux d'apprentissage.

2) SARSA

SARSA (State-Action-Reward-State-Action) apprend une politique optimale dans un environnement séquentiel. Il se concentre sur l'apprentissage des valeurs Q pour chaque paire État-action.

Concepts de SARSA

- **State-Action Pair (État-Action)** : Représente la récompense attendue en effectuant une action à un état.
- **Politique ϵ-Greedy** : Utilise une politique ϵ-greedy pour l'exploration.

Algorithme SARSA

La mise à jour de SARSA est basée sur la différence temporelle :

$$Q(s, a) = Q(s, a) + \alpha \cdot [R(s, a) + \gamma \cdot Q(s', a') - Q(s, a)]$$

3) Policy Gradient

Policy Gradient est une classe d'algorithmes pour apprendre des politiques dans des environnements séquentiels.

Concepts de Policy Gradient

- **Politique (Policy)** : Définit la probabilité de choisir chaque action à partir d'un état.
- **Fonction Objectif** : Vise à maximiser la récompense cumulative.
- **Gradient de la Politique** : Calcul du gradient de la fonction objectif par rapport aux paramètres de la politique.

Les méthodes Policy-Gradients proposent de s'intéresser directement à la politique :

$$\pi_\theta(a|s) = P[a|s, \theta]$$

Dans ce cadre, la probabilité d'une trajectoire $= (s1, a1, s2, a2, \ldots, s\|\tau\|) est donnée par$:

$$\pi_\theta(\tau) = P(s_1) \prod_{t=1}^{\|\tau\|-1} \pi_\theta(a_t|s_t) P(s_{t+1}|s_t, a_t)$$

Algorithme de Policy Gradient

La mise à jour des paramètres θ est effectuée de manière itérative :

$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$$

Politique Déterministe ou Stochastique

La politique peut être déterministe ou stochastique.

Baseline

Certains algorithmes utilisent une baseline pour réduire la variance des estimations de gradient.

Les méthodes de Policy Gradient sont utilisées dans des environnements à espace d'action continu ou lorsque la politique optimale est complexe.

Domaines d'application d'AR

L'apprentissage par renforcement (AR) a des applications diverses et prometteuses dans plusieurs domaines du monde réel, démontrant son potentiel à résoudre des problèmes complexes. Quelques exemples concrets d'applications incluent :

1) Systèmes de Recommandation Personnalisée

Les algorithmes d'AR sont largement utilisés dans les systèmes de recommandation tels qu'Amazon, Netflix, et Spotify. Ces systèmes apprennent les préférences individuelles des utilisateurs au fil du temps, ajustant les recommandations en fonction des actions passées pour améliorer l'expérience utilisateur et augmenter la pertinence des suggestions.

2) Gestion de la Chaîne d'Approvisionnement

Dans la logistique et la gestion de la chaîne d'approvisionnement, l'AR peut optimiser les décisions sur le stockage, la distribution, et la gestion des stocks. Cela conduit à une utilisation plus efficace des ressources et à une réduction des coûts opérationnels.

3) Marketing en Ligne et Publicité

Les entreprises utilisent l'AR pour optimiser leurs stratégies marketing en ligne, ajustant les campagnes publicitaires en temps réel en fonction du comportement des utilisateurs pour maximiser l'efficacité et le retour sur investissement.

4) Gestion de l'Énergie et Bâtiments Intelligents

Dans le secteur de l'énergie, l'AR peut être appliqué à la gestion intelligente de l'énergie dans les bâtiments. Les systèmes apprennent à réguler automatiquement l'éclairage, le chauffage, et la climatisation en fonction des habitudes d'utilisation, contribuant à des économies d'énergie significatives.

5) Robotique Industrielle

Les robots autonomes utilisent l'AR pour apprendre à accomplir des tâches complexes dans des environnements industriels, y compris la navigation sécurisée, la manipulation d'objets délicats, et l'optimisation des processus de fabrication.

6) Jeux Vidéo et Simulation

Outre les jeux stratégiques comme AlphaGo, l'AR est couramment utilisé dans le développement de jeux vidéo pour créer des agents non joueurs (NPC) plus intelligents et adaptatifs. Il est également utilisé dans des simulations pour former des agents virtuels dans des environnements contrôlés.

Ces applications illustrent la polyvalence de l'AR dans la résolution de problèmes du monde réel, ouvrant la voie à des innovations significatives dans divers secteurs.

Problèmes Typiques Résolus par l'AR

L'AR peut résoudre divers problèmes dans divers domaines :

1.Jeux Stratégiques

L'AR a été utilisé pour créer des agents capables de jouer à des jeux stratégiques tels que les échecs, le jeu de go et les jeux vidéo. Des algorithmes comme AlphaGo utilisent l'AR pour apprendre des stratégies gagnantes.

AlphaGo

AlphaGo, développé par DeepMind, a marqué l'histoire en 2016 en battant le champion du monde de go, Lee Sedol. Ce programme d'intelligence artificielle utilise des techniques d'apprentissage profond, notamment des réseaux de neurones, pour évaluer les positions et les mouvements. Grâce à l'auto-jeu, AlphaGo a affiné ses stratégies en explorant des millions de parties contre lui-même. La victoire d'AlphaGo contre Lee Sedol illustre la puissance de l'apprentissage par renforcement et de l'apprentissage profond dans la résolution de problèmes complexes.

2. Robotique

Les robots autonomes peuvent apprendre à effectuer des tâches complexes en utilisant l'AR. Équipés de capteurs pour percevoir leur environnement et d'actuateurs pour effectuer des actions physiques, les robots autonomes peuvent apprendre à accomplir des tâches complexes grâce à l'AR. Des algorithmes comme le Q-learning ajustent progressivement leurs stratégies en fonction des récompenses reçues, permettant aux robots de s'adapter à des environnements changeants. Un exemple concret de robot autonome utilisant l'apprentissage par renforcement est le robot aspirateur Roomba d'iRobot. Les modèles avancés de Roomba, tels que le Roomba i7+ et le Roomba s9+, intègrent des capacités d'apprentissage pour optimiser leur performance dans différents environnements domestiques.

Robots de netoyage Roomba

Ces robots autonomes sont équipés de capteurs pour détecter les obstacles, les changements de surface et les zones à nettoyer. Ils utilisent des algorithmes d'apprentissage par renforcement pour améliorer leur navigation au fil du temps. Par exemple, en utilisant le Q-learning, le Roomba peut ajuster ses trajectoires en fonction des récompenses associées à certaines actions, telles que la réussite de la couverture d'une zone spécifique ou l'évitement efficace des obstacles. Ainsi, le Roomba apprend de manière autonome à optimiser son nettoyage en s'adaptant aux caractéristiques spécifiques de la maison dans laquelle il évolue, offrant un exemple pratique de l'application de l'apprentissage par renforcement dans le domaine des robots domestiques.

3. Gestion de Ressources

L'AR peut être appliqué à des problèmes de gestion de ressources, tels que la gestion de l'énergie, du trafic, ou la distribution de ressources limitées.

Gestion de l'Énergie dans les Bâtiments

Dans la gestion de l'énergie, un système AR peut ajuster automatiquement les paramètres du chauffage, de la ventilation, et de la climatisation pour minimiser les coûts tout en maintenant le confort. L'algorithme AR sélectionné s'adapte aux caractéristiques spécifiques du problème, permettant une optimisation itérative où le système apprend de ses actions passées. L'AR offre également une capacité d'adaptation continue aux changements dynamiques dans l'environnement, assurant une gestion efficace des ressources en temps réel.

Défis et Complexités

1) L'exploration-exploitation

L'exploration-exploitation est un concept fondamental dans le contexte de l'apprentissage par renforcement (RL) et il fait référence au dilemme auquel un agent est confronté lorsqu'il prend des décisions séquentielles dans un environnement incertain. L'agent doit choisir entre deux stratégies concurrentes : l'exploration et l'exploitation.

Exploration

- **Définition :** La recherche d'informations nouvelles en essayant des actions inconnues ou sous-optimales.
- **Objectif :** Acquérir une connaissance plus complète de l'environnement et découvrir des actions potentiellement plus rentables.

Exploitation

- **Définition :** Tirer profit des actions connues ou estimées comme étant les meilleures en fonction des connaissances actuelles.
- **Objectif :** Maximiser les récompenses à court terme en choisissant les actions les plus rentables selon les informations actuelles.

Dilemme Exploration-Exploitation

L'agent doit décider comment allouer ses ressources entre l'exploration et l'exploitation, confronté au risque de manquer des opportunités ou de gaspiller des ressources.

Stratégies d'Exploration-Exploitation : (Stratégie -greedy)

Choisir l'action optimale avec une probabilité $1 - \epsilon$ et explorer une action aléatoire avec une probabilité ϵ.

- ϵ est un hyperparamètre pour ajuster le niveau d'exploration(Plus est élevé, plus l'agent est enclin à explorer).
- Équilibre entre exploration et exploitation.
- Ajustement de ϵ en fonction de la connaissance de l'agent.

Avenir de l'AR

1) Tendances Émergentes

Les tendances actuelles révèlent une convergence accrue entre l'apprentissage par renforcement (AR) et d'autres domaines de l'intelligence artificielle, notamment l'apprentissage profond (DRL) et l'apprentissage automatique hors ligne. L'émergence de modèles plus sophistiqués et la gestion efficace des états continus ont conduit à des avancées telles que l'AR profond (Deep RL). L'exploration de politiques distribuées et d'apprentissage par imitation promet d'étendre les frontières de l'AR.

2) Défis Actuels et Futurs

Des défis subsistent, parmi lesquels la délicate équilibre entre exploration et exploitation, demeure fondamental pour l'AR. La gestion des états continus représente également un défi majeur, exigeant des approches novatrices pour les environnements réels plus complexes. L'AR profond doit surmonter des obstacles liés à la stabilité de l'entraînement et à la nécessité de vastes ensembles de données. Les recherches en cours explorent des solutions hybrides et des méthodes d'apprentissage plus efficientes.

3) Éthique et Interprétabilité

À mesure que la puissance de l'AR s'accroît, des préoccupations éthiques émergent quant à la prise de décision autonome dans des domaines sensibles. La transparence et l'interprétabilité des modèles AR deviennent cruciales pour garantir une utilisation responsable de cette technologie. Les recherches actuelles se concentrent sur le développement de méthodes permettant de comprendre et d'expliquer les décisions prises par les agents AR.

4) Adaptation à l'Adversité et à la Dynamique

L'AR est confronté au défi de s'adapter à des environnements changeants et adverses. Les recherches en cours visent à développer des algorithmes capables de gérer des adversaires intelligents et de s'ajuster à des situations dynamiques, assurant ainsi une performance robuste dans des contextes variés.

5)Applications Sociales et Économiques

L'impact attendu de l'AR sur la société et l'économie suscite des études exploratoires dans des domaines tels que l'éducation, la santé et la prise de décision politique. Ces investigations soulèvent des questions complexes liées à la responsabilité, à la confidentialité des données et à l'équité.

En résumé, les perspectives d'avenir pour l'AR sont riches en innovations technologiques, mais elles nécessitent également une attention particulière pour résoudre les défis émergents, garantir une utilisation éthique et responsable, et étendre son impact dans des domaines sociaux et économiques clés.

Conclusion

Dans le sillage de cette étude approfondie sur l'Apprentissage par Renforcement (AR), il est manifeste que ce domaine constitue un pilier central de l'intelligence artificielle, suscitant un intérêt croissant tant dans le milieu académique que dans l'industrie. L'analyse des fondements théoriques, des algorithmes fondamentaux, des applications sectorielles variées, et des défis émergents a jeté une lumière précieuse sur la sophistication croissante de l'AR.

Les tendances émergentes mettent en avant l'intégration avec d'autres domaines de l'IA, offrant des perspectives prometteuses mais exigeant également des solutions novatrices pour des défis tels que la gestion d'états continus et la stabilité des modèles. Les défis actuels, notamment la délicate balance entre exploration et exploitation, et la nécessité de l'interprétabilité éthique, soulignent la nécessité d'une recherche continue pour assurer le développement responsable de l'AR.

L'avenir de l'AR réside non seulement dans ses avancées technologiques, mais également dans son impact potentiel sur des secteurs clés de la société et de l'économie. Les questions éthiques et sociales, la capacité à s'adapter à l'adversité, et l'extension des applications dans des contextes toujours plus variés préfigurent un paysage dynamique pour l'AR.

En définitive, cette exploration scientifique de l'AR a fourni un cadre complet pour comprendre son écosystème complexe. Alors que l'AR continue d'évoluer, les avenues de recherche et les applications concrètes promettent de contribuer de manière significative à la manière dont nous conceptualisons, interagissons, et résolvons des problèmes complexes au sein de la société contemporaine.