+33 (0)1 48 10 09 46
formation@ambient-it.net
140 Avenue Jean Lolive 93500 Pantin - France

Mis à jour le 23/09/2025

S'inscrire

Formation Introduction au Deep Reinforcement Learning

2 jours (14 heures)

Présentation

Le Reinforcement Learning met en œuvre un système large où un agent doit apprendre à résoudre un problème à partir de récompenses. Si ce domaine existe depuis un certain temps, l'arrivée du Deep Learning l'a bouleversé en mettant à disposition de nouveaux outils, approximant des outils (Q function, policy, etc.) par des réseaux de neurones. De nombreuses réussites ont démontré que malgré sa difficulté particulière, cette approche peut révolutionner certains problèmes : jeu vidéo, optimisation de process, jeu de go, contrôle continu ou robotique.

L'objectif ici est de présenter les bases du Reinforcement Learning, puis les principales avancées apparues ces dernières années : Deep Q Learning, Rainbow, Policy gradients (A3C, PPO), exploration (World models, Imagination augmented agents) jusqu'à une étude détaillée d'AlphaGo et AlphaGo Zero.

Objectifs

- Maîtrise des concepts du reinforcement learning et des approches "model-free" principales.
- Compréhension des approches basées sur l'exploration et étude des approches d'optimisation
- Étude de solutions "modelbased" : apprentissage du modèle ou utilisation directe
- Illustration des points abordée via les exemples d'application AlphaGo et AlphaGoZero

Public visé

Développeurs, Architectes, Big Data Data analyst / Data scientist & Engineer

Pré-requis

- Connaissance de Python

Pour aller plus loin

- Nous vous proposons en introduction un formation sur l'Intelligence Artificielle
- En complément les technologies
  - Pytorch de Facebook
  - TensorFlow de Google

Programme de notre formation sur le Deep Reinforcement Learning

[JOUR 1]

1. Introduction aux concepts du Reinforcement Learning

- Présentation du reinforcement learning : contrôle d'un agent dans un environnement défini par un état et des actions possibles. Modélisations fondamentales
- Modélisation en Markov Decision Processes, définition des Value Functions, équation de Bellman, dynamic programming. Distinction entre observation et état de l'environnement
- Approche par Value prediction : Temporal Difference & Monte Carlo. Mise en exemple de ces algorithmes
- Policy iteration & evaluation : algorithme fondamental de convergence d'une politique d'action.
- Q Learning

2. Model Free Deep Reinforcement Learning (deux exemples d'implémentation Tensorflow ou PyTorch sont étudiés selon les directions des élèves)

- Deep Q-Learning : Approche fondamentale, approximation de la fonction Q, Experience Replay, Double Q Learning. Étude des résultats en détail
- Deep Recurrent Q-Learning : Problématique d'un état partiellement observable. Comparaison avec le Deep Q Learning
- Rainbow : analyse des avancées et modifications d'architecture en Deep Q Learning: dueling networks, prioritized experience replay, approche distributionnelle, utilisation d'un bruit. Analyse des apports combinés et individuels de chaque approche

Références : - Playing Atari with Deep Reinforcement Learning, Mnih et al, 2013. - Deep Recurrent Q-Learning for Partially Observable MDPs, Hausknecht and Stone, 2015 - Rainbow: Combining Improvements in Deep Reinforcement Learning, Hessel et al, 2017.

- Policy Gradients : Architecture Actor Critic

- Approche Asynchrone A3C. Définition asynchrone du Deep Q Learning. Algorithme A3C, intérêt, performances et souplesse de l'approche asynchrone
- Évolution d'une policy par policy gradient : Trusted Policy Optimization et Proximal Policy Optimization. Avantages apportés par l'approche PPO. Étude des résultats et des conditions d'application.
- Soft actor critic : utilisation d'un paramètre d'entropie pour maximiser l'exploration. Détails d'architecture

Références : - Asynchronous Methods for Deep Reinforcement Learning, Mnih et al, 2016 - Proximal Policy Optimization Algorithms, Schulman et al, 2017.

- Approche distributionnelle : adaptation des équations et définitions fondamentales. Motivation de l'approche et résultats observés.
- Algorithmes à évolution : utilisation de Natural Evolution Strategies pour une convergence Deep Reinforcement Learning. Vision de l'optimisation et de la parallélisation possible de l'apprentissage. Analyse des résultats comparés.

Références : - Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor, Haarnoja et al, 2018 - Evolution Strategies as a Scalable Alternative to Reinforcement Learning, Salimans et al, 2017

[JOUR 2]

3. Exploration de l'environnement

- Exploration versus apprentissage : quelle pondération, quel intérêt ? Comment définit-on l'exploration ?
- Étude des explorations basées sur un décompte des états/actions.
- Analyse des modélisations possibles de l'état par Hash. Apprentissage du hash par Variational Autoencoder (rappel des principes du VAE)
- Concepts de "curiosité"
- Approche basée uniquement sur l'exploration sans récompense directe. Résultats, intérêts et discussions

Références : - Exploration: A Study of Count-Based Exploration for Deep Reinforcement Learning, Tang et al, 2016 - Large-Scale Study of Curiosity-Driven Learning, Burda et al, 2018

4. Model based Deep Reinforcement Learning : apprentissage du modèle.

- Mise en œuvre de l'apprentissage d'un modèle interne à l'agent devant représenter l'environnement.
- Étude des différentes stratégies de modélisation. Approche probabiliste ou déterministe.
- Entrainement d'un modèle dans son environnement "interne" et application à l'environnement cible.
- Étude du concept d' "imagination" (Deepmind), Imagination Augmented Agent. Exploitation d'un apprentissage libre avec modélisation des états futurs d'une manière interne. Études d'ablation.
- Résultats comparés

Références : - Imagination-Augmented Agents for Deep Reinforcement Learning, Weber et al, 2017 - Recurrent World Models Facilitate Policy Evolution, Ha and Schmidhuber, 2018.

5. Approches model-based : AlphaGo, AlphaGo Zero et dérivés

- Monte Carlo Tree Search (MCTS) : analyse de l'algorithme fondamental
- AlphaGo : analyse de l'apprentissage en quatre étapes, et utilisation de la MCTS pondérant les différents réseaux de neurones disponibles. Analyse de la performance et des résultats
- AlphaGo Zero : analyse des évolutions, utilisation de la MCTS au sein de l'apprentissage. Comparaison AlphaGo VS AlphaGO Zero
- AlphaZero : généralisation de l'approche AlphaGo Zero à d'autres approches
- Imitation Learning : définition et exemples
- Expert Iteration : utilisation de la MCTS pour modélisation interne d'un modèle expert permettant de mettre en oeuvre l'imitation learning.

Références : - Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm, Silver et al, 2017 - Thinking Fast and Slow with Deep Learning and Tree Search, Anthony et al, 2017

6. Scaling d'un apprentissage RL et algorithmes récents

- Analyse des possibilités de parallélisation GPU versus CPU. Stratégies d'approches et de mitigation. Vision "data-efficiency" des approches proposées.
- Approche distributive pour parallélisation plus importante des apprentissages
- Analyse de l'algorithme R2D2 : utilisation de modèles récurrents et parallélisation, analyse poussée des biais induits par la variation de l'état caché du réseau

Références : - Accelerated Methods for Deep Reinforcement Learning, Stooke and Abbeel, 2018 - Recurrent Experience Replay in Distributed Reinforcement Learning, Kapturowski et al, 2018

Sociétés concernées

Cette formation s’adresse à la fois aux particuliers ainsi qu’aux entreprises, petites ou grandes, souhaitant former ses équipes à une nouvelle technologie informatique avancée ou bien à acquérir des connaissances métiers spécifiques ou des méthodes modernes.

Positionnement à l'entrée en formation

Le positionnement à l'entrée en formation respecte les critères qualité Qualiopi. Dès son inscription définitive, l'apprenant reçoit un questionnaire d’auto-évaluation nous permettant d’apprécier son niveau estimé sur différents types de technologies, ses attentes et objectifs personnels quant à la formation à venir, dans les limites imposées par le format sélectionné. Ce questionnaire nous permet également d’anticiper certaines difficultés de connexion ou de sécurité interne en entreprise (intraentreprise ou classe virtuelle) qui pourraient être problématiques pour le suivi et le bon déroulement de la session de formation.

Méthodes pédagogiques

Stage Pratique : 60% Pratique, 40% Théorie. Support de la formation distribué au format numérique à tous les participants.

Organisation

Le cours alterne les apports théoriques du formateur soutenus par des exemples et des séances de réflexions, et de travail en groupe.

Validation

À la fin de la session, un questionnaire à choix multiples permet de vérifier l’acquisition correcte des compétences.

Sanction

Une attestation sera remise à chaque stagiaire qui aura suivi la totalité de la formation.