TP06 & TP07

Améliorations du DQN et TradingEnv

Étude et implémentation des améliorations du Deep Q-Network (Double DQN, Dueling DQN, Prioritized Experience Replay) et application à un environnement de trading réaliste utilisant Gymnasium.

Double DQN (DDQN)

Problématique

Le DQN classique surestime les Q-values, menant à un biais optimiste qui déstabilise l'apprentissage.

Principe

Séparation des rôles entre deux réseaux :

- **Réseau Online :** Sélectionne l'action future
- **Réseau Target :** Évalue cette action

Fonctionnement

Cette séparation du rôle de sélection et d'évaluation réduit le biais optimiste en évitant que le même réseau ne sélectionne et n'évalue l'action.

Formule

> a* = argmax_a Q_online(s', a)
>
> target = r + γ·Q_target(s', a*)·(1-done)

Bénéfices

- Réduit le biais d'optimisme
- Améliore la stabilité de l'apprentissage
- Convergence plus fiable
- Fondamental pour l'apprentissage par renforcement stable

Dueling DQN

Architecture Innovante

Dueling DQN décompose la Q-value en deux composantes distinctes :

- **Valeur de l'état :** V(s) - importance intrinsèque de l'état
- **Avantage de l'action :** A(s,a) - avantage relatif de l'action

Motivation

Certains états sont intrinsèquement meilleurs que d'autres, indépendamment de l'action. Dueling permet à l'agent d'apprendre cette distinction, améliorant la généralisation.

Formule Combinée

> Q(s,a) = V(s) + (A(s,a) - mean(A(s,a')))

La soustraction de la moyenne des avantages assure l'identifiabilité.

Bénéfices

- Accélère l'apprentissage
- Améliore la généralisation
- Efficace dans les environnements où la valeur de l'état prime
- Amélioration architecturale majeure

Prioritized Experience Replay (PER)

Problématique

L'échantillonnage uniforme du buffer de rejeu est inefficace : toutes les transitions sont traitées de manière égale, même celles dont l'agent a peu à apprendre.

Principe Fondamental

PER remplace l'échantillonnage uniforme par un échantillonnage intelligent basé sur la **priorité** de chaque transition.

Mécanisme

- Les transitions "surprenantes" (erreur TD élevée) sont rejouées plus souvent

- Correction du biais d'échantillonnage via les IS weights

- Recuit du paramètre β pour progressivement réduire la correction

Formules Clés

> $p_i = |\delta_i|^\alpha$
>
> $w_i = (1/(N \cdot P(i)))^\beta$

où δ_i est l'erreur TD, α contrôle la priorité, β contrôle la correction.

Bénéfices

- Apprentissage plus efficace

- Convergence plus rapide

- Concentration sur les données informatives

- Amélioration majeure de la performance

Construction du TradingEnv

Observation

État du marché fourni à l'agent :

- Prix actuel et historique
- Volume d'échange
- Indicateurs techniques
- Position actuelle du portefeuille
- Cash disponible

Espace d'observation continu et normalisé pour l'apprentissage optimal.

Action

Décisions disponibles pour l'agent :

- **Achat (Buy)** : Augmenter la position
- **Vente (Sell)** : Réduire la position
- **Maintien (Hold)** : Conserver la position

Espace d'action discret avec 3 actions possibles à chaque pas de temps.

Récompense

Signal de performance fourni à l'agent :

- Basée sur le profit/perte du portefeuille
- Pénalité pour les coûts de transaction
- Ajustement pour le risque
- Récompense à chaque pas de temps

Fonction de récompense réaliste et alignée avec les objectifs d'investissement.

Indicateurs Techniques et Coûts

Indicateurs Techniques

Les indicateurs techniques fournissent à l'agent des signaux de marché pour prendre des décisions éclairées.

SMA (Simple Moving Average)

Moyenne arithmétique des prix sur une période fixe. Lisse les fluctuations court terme et identifie les tendances.

EMA (Exponential Moving Average)

Moyenne pondérée donnant plus de poids aux données récentes. Plus réactive aux changements de prix que la SMA.

Croisements de Moyennes

Les signaux d'achat/vente sont générés lorsque les moyennes mobiles se croisent, indiquant des changements de tendance.

Coûts de Transaction

Les coûts réels du trading (commissions, spreads) sont appliqués lors des changements de position.

> transaction_cost = price × cost_ratio

Réalisme de l'Environnement

- Coûts appliqués uniquement aux changements
- Réduit les décisions excessives
- Reflète les contraintes du marché réel
- Améliore la qualité des stratégies apprises

Intégration

Les indicateurs et coûts sont intégrés dans l'espace d'observation et la fonction de récompense pour guider l'apprentissage de l'agent.

Métriques d'Évaluation

Sharpe Ratio

Performance Ajustée au Risque

Mesure le rendement excédentaire par unité de risque (volatilité).

> Sharpe = (R - Rf) / σ

où R est le rendement, Rf le taux sans risque, σ la volatilité.

- Plus élevé = meilleur
- Normalise le risque
- Idéal pour comparer des stratégies

Max Drawdown

Perte Maximale Observée

Mesure la plus grande perte entre un pic et un creux du portefeuille.

> MDD = (Trough - Peak) / Peak

Exprimé en pourcentage négatif.

- Plus proche de 0 = meilleur
- Mesure le risque baissier
- Critique pour la gestion du risque

Buy & Hold

Baseline de Référence

Stratégie passive : achat initial, conservation jusqu'à la fin.

> B&H = (Final - Initial) / Initial

Représente la performance d'un investisseur passif.

- Benchmark absolu
- Difficile à surpasser
- Référence pour l'agent

Résultats TP06 : Rewards Cumulés

Observations

DQN Classique : Progression irrégulière avec fluctuations importantes

Double DQN : Progression plus régulière et stable

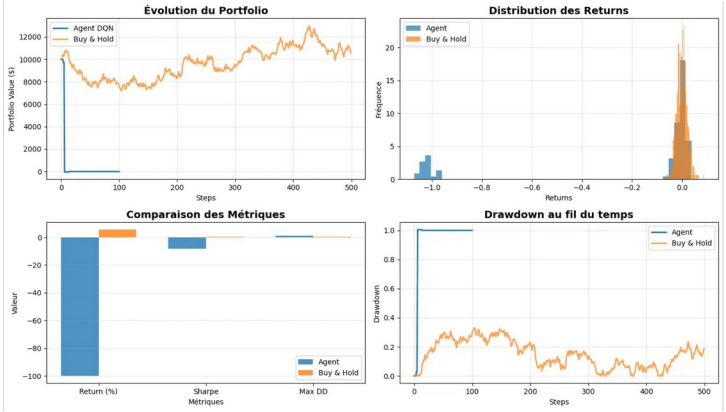
Dueling DQN : Rewards supérieurs et convergence plus rapide

Analyse

Les variantes améliorées (DDQN et Dueling) démontrent une stabilité théorique supérieure. La réduction du biais d'optimisme et la décomposition architecturale permettent un apprentissage plus fiable et une meilleure généralisation.

> **Conclusion :** Les améliorations du DQN sont théoriquement bénéfiques et se traduisent par une meilleure performance en termes de stabilité et de convergence.

Résultats TP07 : Agent vs Baseline



Observations

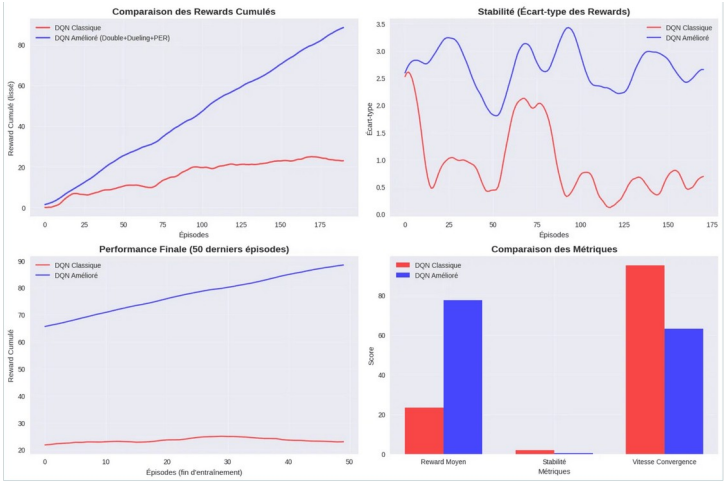
- **Buy & Hold :** Progression régulière, typique d'un marché haussier
- **Agent DQN :** Fluctuations importantes, souvent en dessous de la baseline
  - Décisions "catastrophiques" très tôt dans l'entraînement
  - Gestion de positions et coûts de transaction mal optimisés

Interprétation

L'agent n'a pas réussi à surpasser la stratégie passive. Ses décisions actives (achats/ventes) n'apportent pas de valeur ajoutée, voire dégradent la performance.

> **Conclusion :** Les améliorations du DQN apportent une meilleure stabilité théorique, mais l'application au trading nécessite un affinement de l'entraînement, de la gestion des coûts et de la stratégie d'exploration.

Comparaison DQN Classique vs Amélioré



Efficacité : Dueling améliore l'apprentissage en séparant la valeur de l'état et l'avantage de l'action, tandis que PER concentre les ressources sur les données les plus informatives.

Tableau Comparatif

Aspect	DQN Classique	DQN Amélioré
Stabilité	Biais optimiste	Réduit par DDQN
Architecture	Q(s,a) simple	Dueling (V+A)
Échantillonnage	Uniforme	PER intelligent
Convergence	Lente, instable	Rapide, stable
Performance	Variable	Supérieure

Avantages du DQN Amélioré

Le DQN amélioré (DDQN + Dueling + PER) combine trois améliorations majeures qui se renforcent mutuellement :

Robustesse : DDQN élimine le biais optimiste inhérent au DQN classique, garantissant une estimation plus fiable des Q-values et une convergence plus stable.

Conclusion et Perspectives

Synthèse

Les améliorations du DQN sont cruciales :

- **DDQN :** Réduit le biais d'optimisme
- **Dueling :** Améliore la généralisation
- **PER :** Accélère l'apprentissage

Ces techniques constituent l'état de l'art en apprentissage par renforcement moderne.

Défis en Trading

L'application au trading révèle des défis majeurs :

- Complexité de l'environnement
- Non-stationnarité des données
- Coûts de transaction réels
- Difficulté à surpasser Buy & Hold

La théorie ne suffit pas ; l'implémentation pratique requiert un affinement constant.

Perspectives

Directions futures pour améliorer les performances :

- Hyperparamètres optimisés
- Architectures plus complexes
- Stratégies d'exploration avancées
- Environnements multi-actifs
- Combinaison avec d'autres méthodes

La recherche continue est essentielle pour obtenir des agents réellement performants.

Merci

Fin de la Présentation