ÉTENDRE LE CONCEPT DE GÉNÉRALISATION EN APPRENTISSAGE PAR RENFORCEMENT

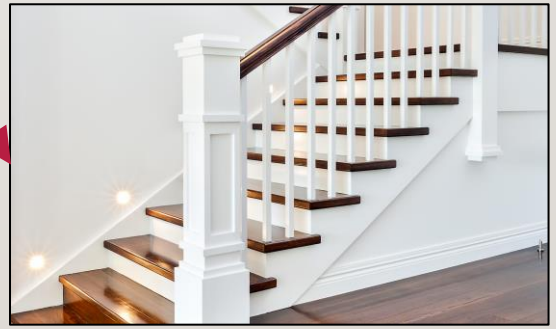
RONNIE LIU (20154429)

POUR LE COURS IFT 3150

MOTIVATION ET OBJECTIFS

2

Environnement entraîné



Différents types de surfaces lesquelles le robot n'est pas habitué

Surapprentissage

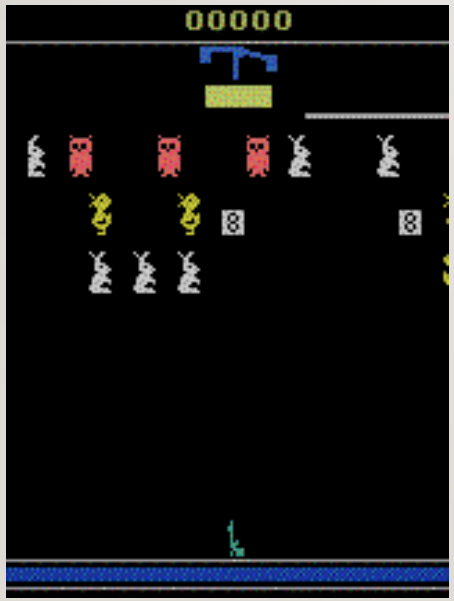
MOTIVATION ET OBJECTIFS

OBJECTIFS DU PROJET

1. Maîtriser les fondements de RL

2. Proposer une méthode d'entraînement afin de mieux **généraliser** pour des tâches similaires.

3. Appliquer et illustrer que cette méthode augmente la performance de l'agent à bien généraliser.



Carnival

Assault

Space Invaders

REVUE DE LITTÉRATURE

1. Article *Playing Atari with Deep Reinforcement Learning* https://arxiv.org/pdf/1312.5602.pdf

- 7 jeux Atari avec même réseau de CNN.
- But: Comparer la performance des agents en fonction de différents algorithmes classiques de RL.

SIMILARITÉS	DIFFÉRENCES
• Modèle CNN pour notre politique • Même processus prétraitement des images avec un seul canal au lieu de 4	• Entraîner sur de différents environnements lors de la phase d'entraînement et de test

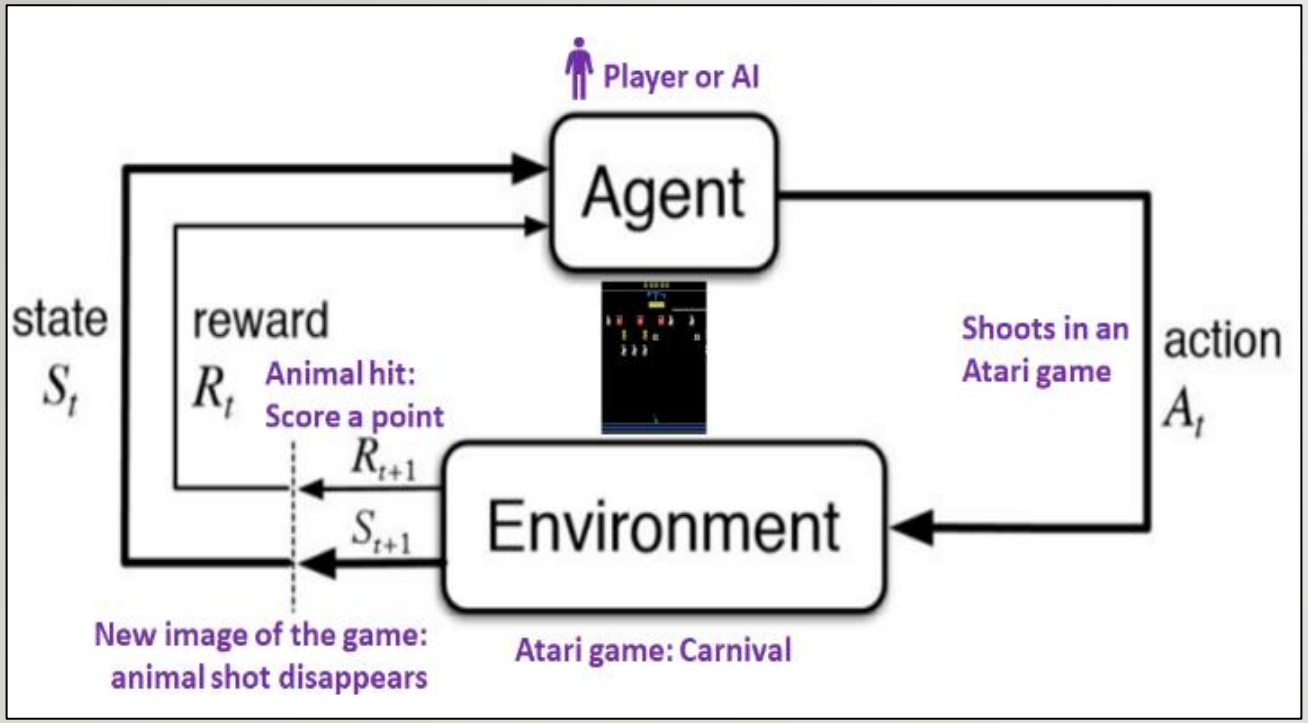
2. Article *Offline Reinforcement Learning with Implicit Q-Learning* https://arxiv.org/pdf/2110.06169.pdf

- Expériences sur les humanoïdes
- But: Illustrer la haute performance de l'algorithme IQL en termes de généralisation

SIMILARITÉS	DIFFÉRENCES
• Application de l'algorithme IQL dans notre projet	• Utilisation des jeux Atari au lieu des humanoïdes

CONTEXTE THÉORIQUE

APPRENTISSAGE PAR RENFORCEMENT (RL)



* But: Gagner le plus de points possibles (récompenses)

↓ Optimiser sur la…

* Politique $\pi(a \mid s, \theta)$: distribution sur des actions *a* étant donné un état *s*

↓ En se servant…

* Fonction de Qualité: Espérance des futures récompenses obtenues par l'agent à un état initial *s* avec une action choisie *a*.

$$Q_\pi(s,a) = \mathop{\mathrm{E}}_{s\sim p(s_0)}\left\{\sum_{t=0}^{\infty}\gamma^t R(s_t,a_t,s_{t+1}) \mid s_0 = s, a_0 = a\right\} pour\ 0 < \gamma < 1$$

$$\pi^* = \underset{a \in A}{\operatorname{argmax}}\, Q^*(s,a)$$

CONTEXTE THÉORIQUE

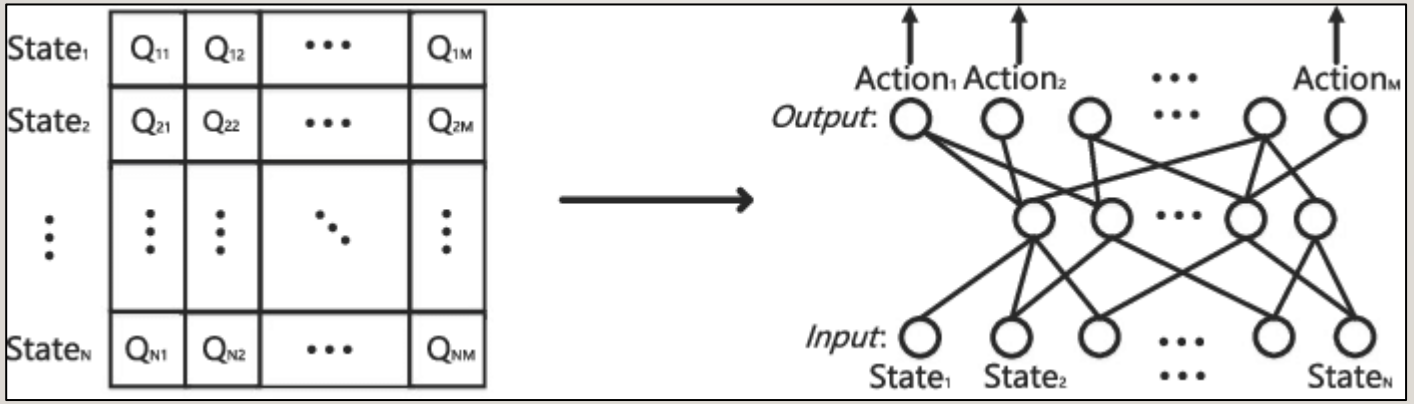
Q-LEARNING ET DEEP Q-LEARNING (DQN)

Principe derrière Q-Learning

- Epsilon-Greedy: exploration de l'environnement

$$Q(s_t, a_t)_{new} = Q(s_t, a_t)_{old} + \alpha \left[r + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)_{old} \right]$$

Q-Learning vs DQN



CONTEXTE THÉORIQUE

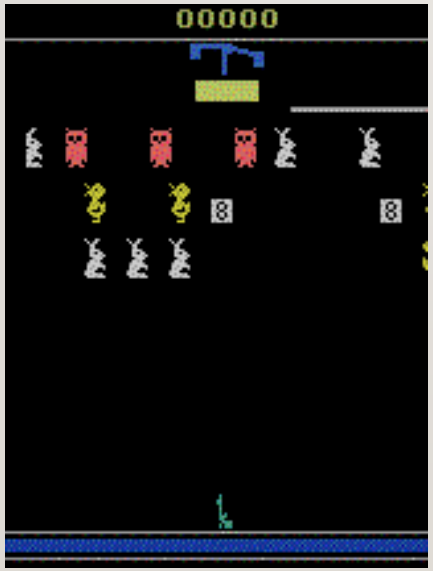
GÉNÉRALISATION EN APPRENTISSAGE PAR RENFORCEMENT

- Fonction de perte ou Récompenses obtenues? Quelle mesure doit-on utiliser?
- Application de cette analogie avec les jeux Atari

Environnement d'entraînement

Environnement de test

Haute performance

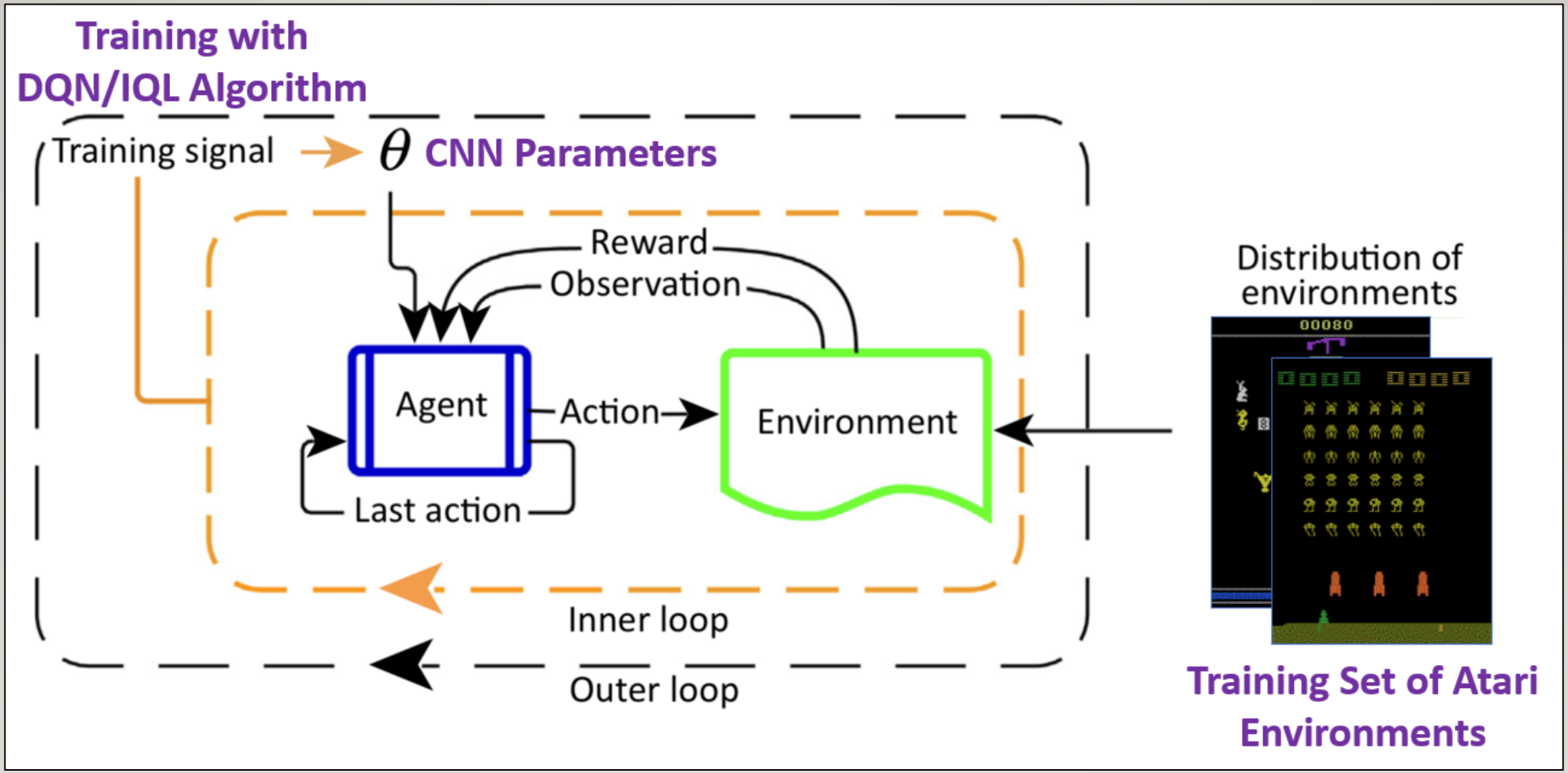


≠

Mauvaise performance

MÉTHODOLOGIE

ZERO-SHOT META LEARNING



MÉTHODOLOGIE

PRÉTRAITEMENT ET MODÈLE CNN

9

Image originale de taille 3 x 210 x 160

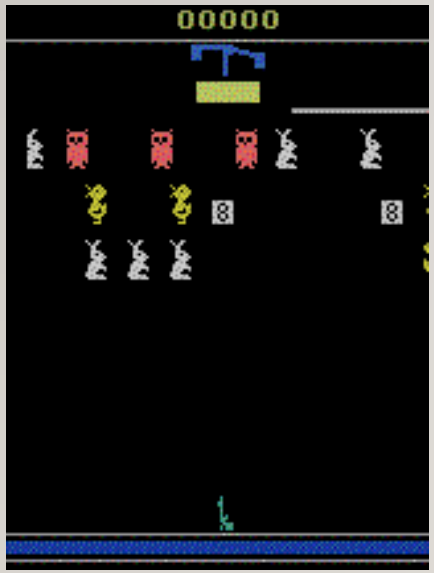
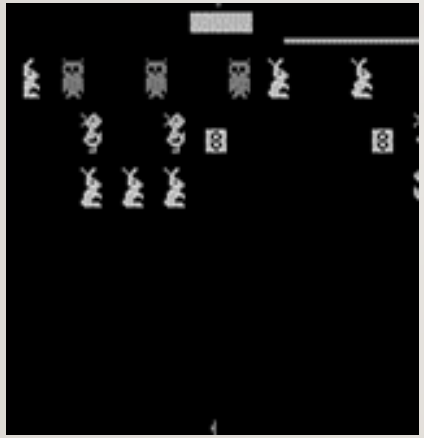


Image prétraitée de taille 1 x 84 x 84



CNN Model

Taken from the following paper https://arxiv.org/pdf/1312.5602.pdf

Input: 256 x 1 x 84 x 84
Convolution Layer 1: 16 filters of 8x8 with stride 4 → output: 256 x 16 x 20 x 20
Convolution Layer 2: 32 filters of 4x4 with stride 2 → output: 256 x **32 x 9 x 9**
[Flattened Tensor]
FC Layer 1: → output: 256 x 256
FC Layer 2 → output: 256 x **# types of actions**

Actions avec leurs valeurs de qualité correspon-dantes

MÉTHODOLOGIE

ALGORITHME PROPOSÉ: IQL

Q-LEARNING (OFF-POLICY)

$$Q(s_t, a_t)_{new} = Q(s_t, a_t)_{old} + \alpha \left[r + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)_{old} \right]$$

IMPLICIT Q-LEARNING (OFFLINE RL)

$$Q(s_t, a_t)_{new} = Q(s_t, a_t)_{old} + \alpha \left| r + \gamma \max_{a_{t+1} \in A,\ s.t.\ \pi(a_{t+1}|s_{t+1}) > 0} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)_{old} \right|$$

Algorithm 1 Implicit Q-learning

Initialize parameters $\psi, \theta, \hat{\theta}, \phi$.
TD learning (IQL):
for each gradient step **do**
$\psi \leftarrow \psi - \lambda_V \nabla_\psi L_V(\psi)$
$\theta \leftarrow \theta - \lambda_Q \nabla_\theta L_Q(\theta)$
$\hat{\theta} \leftarrow (1 - \alpha)\hat{\theta} + \alpha\theta$
end for
Policy extraction (AWR):
for each gradient step **do**
$\phi \leftarrow \phi - \lambda_\pi \nabla_\phi L_\pi(\phi)$
end for

1. Variable Violette : Fonction de perte de la valeur
2. Variable Verte : Fonction de perte des valeurs de qualité
3. Variable Bleue : Régression pondérée des avantages pour extraire la politique

EXPÉRIENCES

Numéro	Algorithme utilisé	Environnements d'entraînement	Environnement de test
1	DQN	Assault	Assault
2	DQN	Carnival, Space-Invaders	Assault
3	IQL	Assault	Assault
4	IQL	Carnival, Space-Invaders	Assault
Noter que l'expérience 4 est la méthode proposée du projet.			

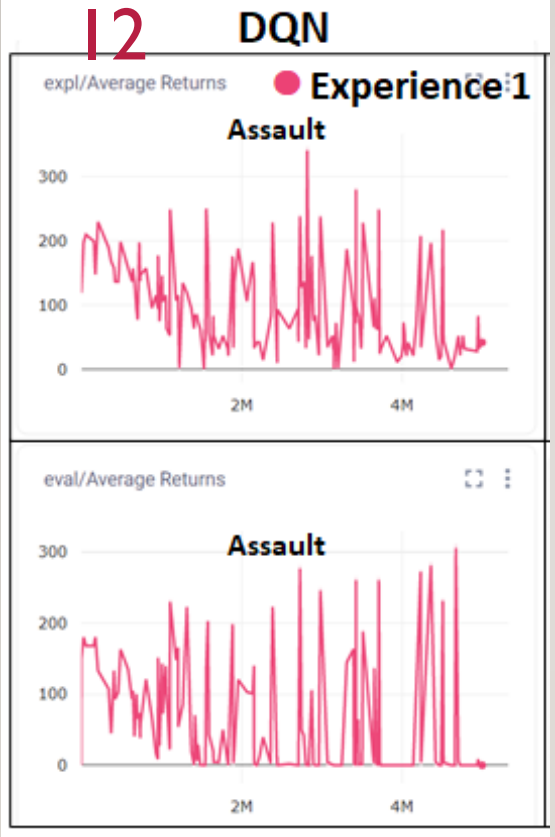
Utiliser les expériences 1 et 2 comme expériences de base.

Comparer les expériences 1 et 3 – DQN vs IQL avec un seul environnement d'entraînement

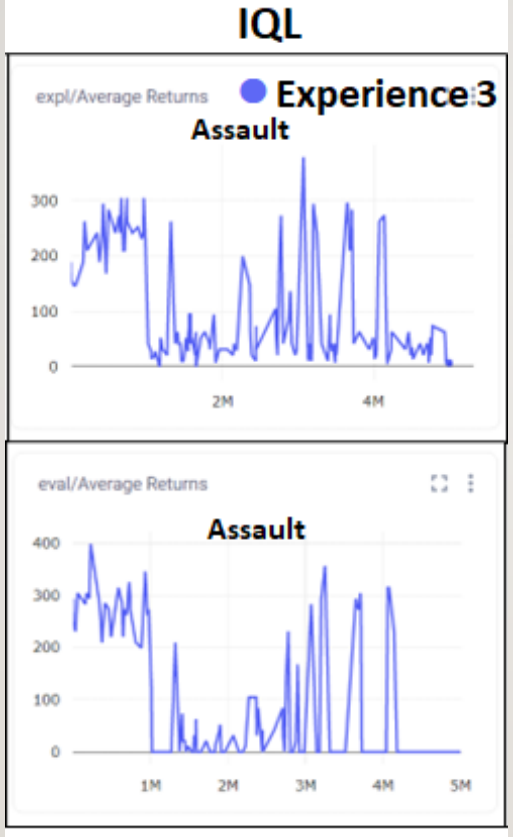
Comparer les expériences 2 et 4 – DQN vs IQL avec un ensemble d'environnements d'entraînement

Comparer les expériences 3 et 4 – Différents nombres d'environnement d'entraînement avec l'algo IQL

DISCUSSION

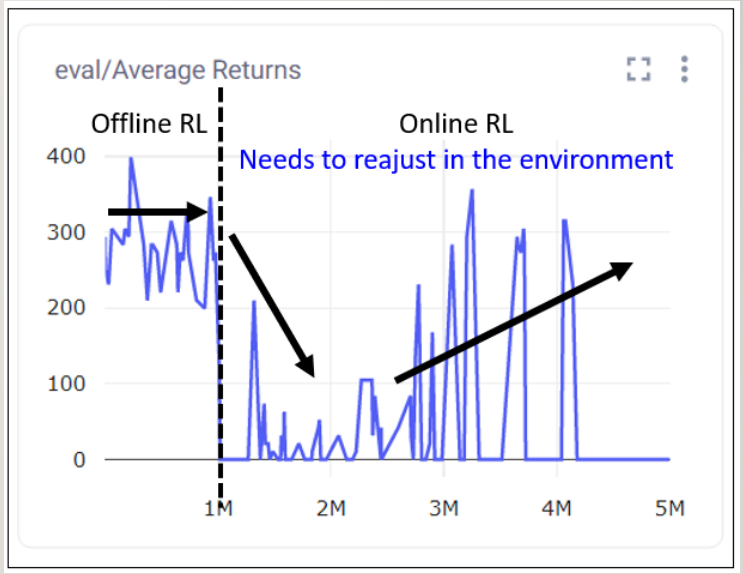


Score maximum
dans l'évaluation: 311 points
(avec 4.7 M de pas)

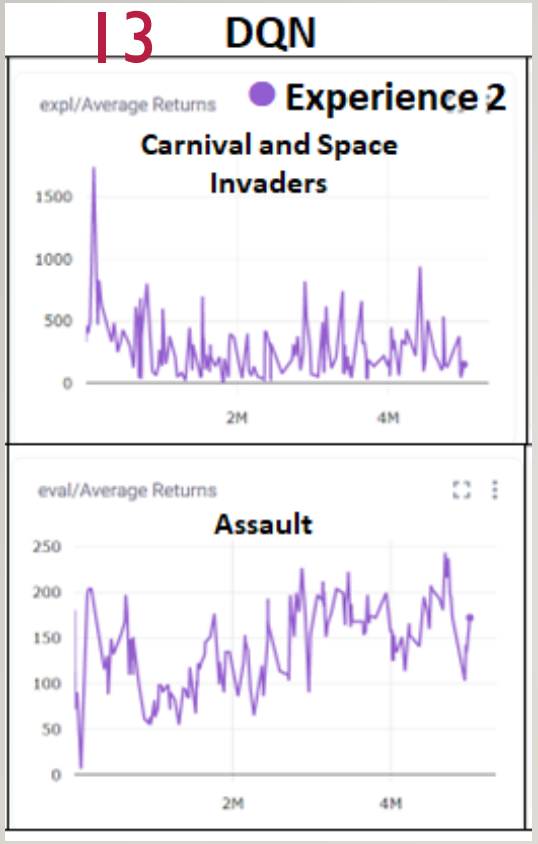


Score maximum
dans l'évaluation: 357 points
(avec 3.2 M de pas)

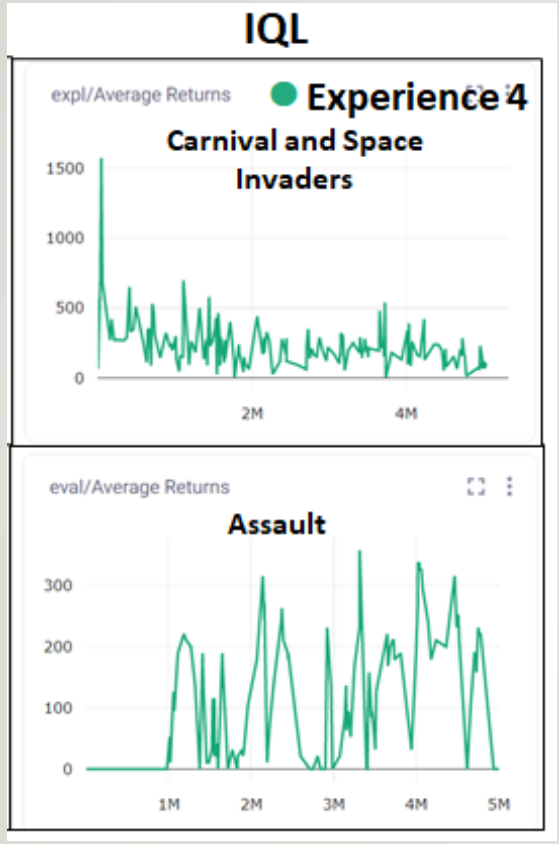
- Expérience 3 avec IQL a un score plus haut que l'expérience 1 avec DQN
- Phénomène étrange dans la courbe d'évaluation de l'expérience 3



DISCUSSION

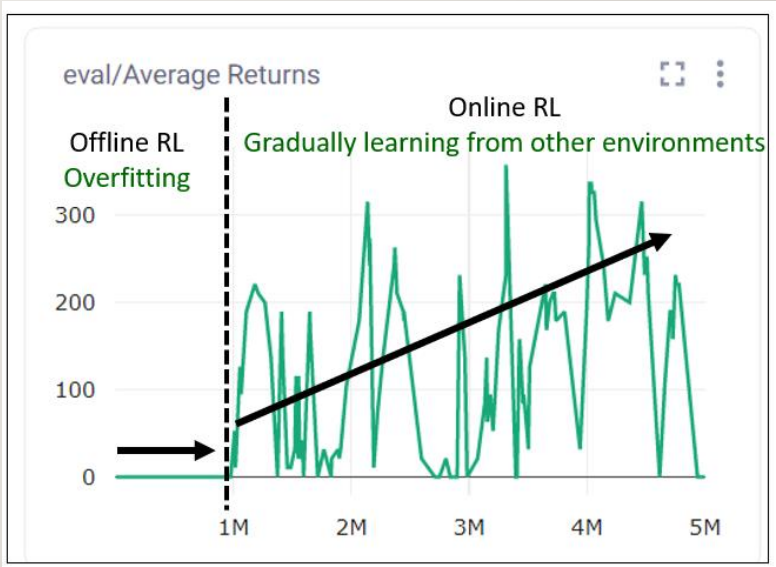


Score maximum
dans l'évaluation: 244 points
(avec 4.7 M de pas)



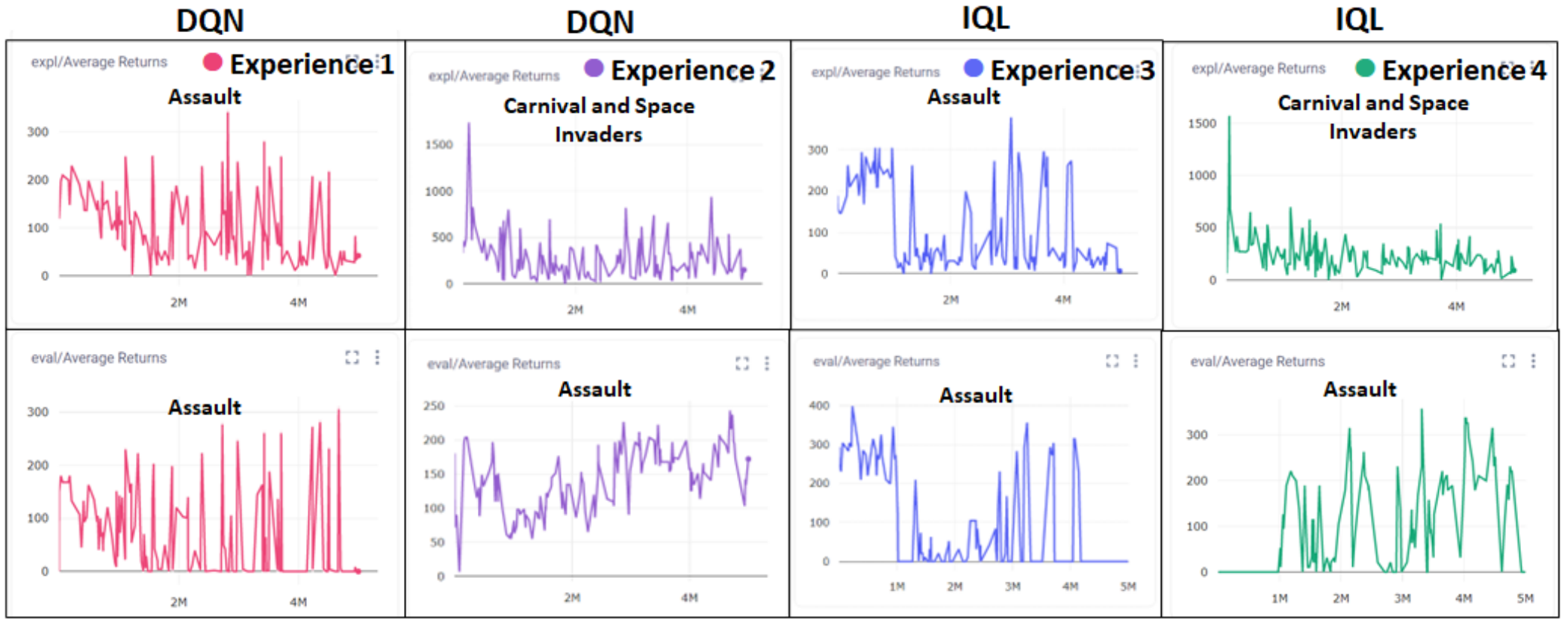
Score maximum
dans l'évaluation: 357 points
(avec 3.3 M de pas)

- Expérience 4 avec IQL a un score plus haut que l'expérience 2 avec DQN
- Phénomène étrange aussi dans la courbe d'évaluation de l'expérience 4



RÉSULTATS SOMMAIRES

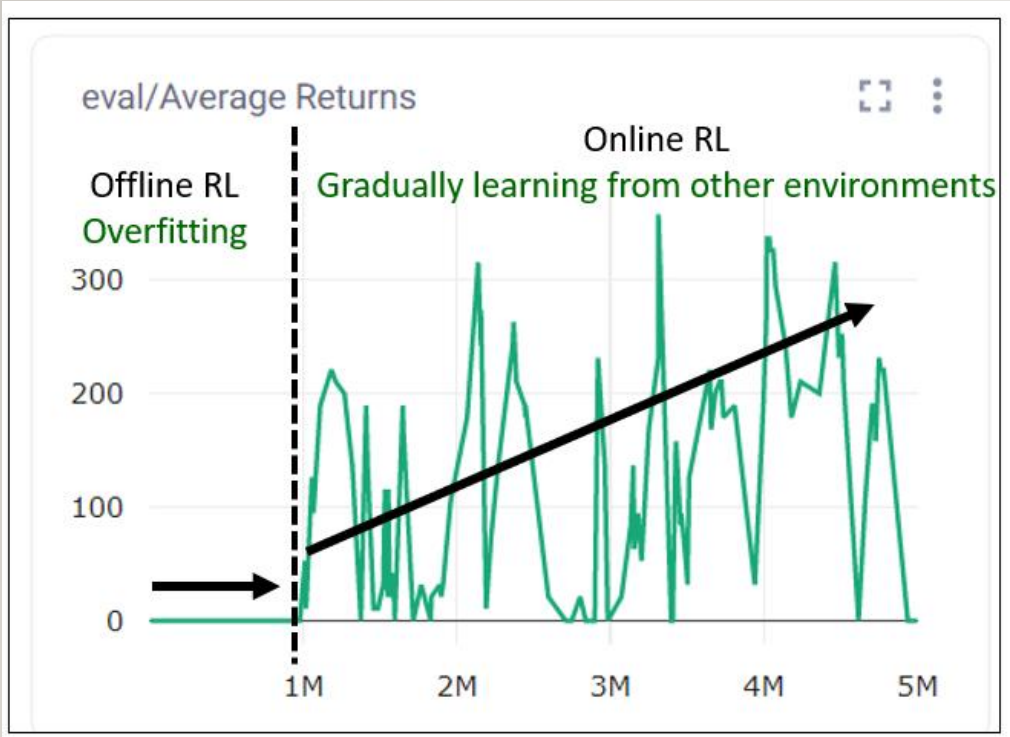
Graphique I. Récompenses totales moyennes durant la phase d'entraînement (haut) et de test (bas) en fonction de nombre d'étapes (pas)



Numéro d'expérience	Récompense maximale	Nombre d'étapes pour atteindre la valeur maximale
1	311 points	Environ 4.7 millions
2	244 points	Environ 4.7 millions
3	357 points	Environ 3.2 millions
4	357 points	Environ 3.3 millions

SOURCE D'ERREURS

15



- Implémentation erronée dans l'algorithme IQL
  - IQL, un algorithme qui améliore la généralisation selon l'article, mais presence de surajustement ici.
- Taille du Replay Buffer pour stocker les données

CONCLUSION

16

- Méthode proposée: Zero-Shot Meta Learning avec l'algorithme IQL

- Environnement d'entraînement = de test
  - IQL: score de 357 points > DQN: score de 311 points.

- Deux environnements d'entraînement et un nouvel environnement de test
  - IQL: score de 357 points > DQN: score de 244 points.

- IQL avec plusieurs environnements est la meilleure méthode pour que l'agent puisse s'adapter dans différents jeux

Contributions futures

- Mettre l'accent sur divers choix d'hyperparamètres

- Application d'un algorithme de Meta-Learning pour approfondir davantage le concept de généralisation en RL

LIENS DES IMAGES (RÉFÉRENCES)

Diapositive 2
Robot: https://172118l113.rsc.cdn77.org/data/images/full/41137/robot-dog.jpg?w=600?w=430
Tapis: https://lh5.googleusercontent.com/p/AF1QipOkbQ_tZ8LPZ8hHs5h6eIKAuqJreW3jSvudXO8w
Terre: https://thumbs.dreamstime.com/b/earth-ground-texture-as-background-nature-environment-environmental-backdrop-175502313.jpg
Escalier: https://boiseriesmetropolitaines.com/wp-content/uploads/2020/03/escalier-residentiel.jpg

Diapositive 3
Space Invaders: https://www.gymlibrary.ml/environments/atari/space_invaders/
Carnival: https://www.gymlibrary.ml/environments/atari/carnival/
Assault: https://www.gymlibrary.ml/environments/atari/assault/

Diapositive 5
Schéma de RL: https://i.stack.imgur.com/eoeSq.png

Diapositive 6
Q-Learning vs DQN:
https://www.researchgate.net/publication/352158682/figure/fig4/AS:1031386303582213@1622913065565/Difference-between-Q-Learning-and-DQN.png

Diapositive 8
Schéma de Meta-Learning: https://lilianweng.github.io/posts/2019-06-23-meta-rl/

Diapositive 9
Algorithme de IQL: https://arxiv.org/pdf/2110.06169.pdf