**Réalisé par :**

Philippe GLASS

Abdoulaye HAMAT

**TP Apprentissage profond par renforcement**

****

1. **Description générale de notre implémentation**

**Choix effectués :**

Nous avons opté pour la librairie de deep learning Tensorflow.

Pour l’environnement CartPole-v1, la dimension d'entrée est de 4 pour les 4 valeurs que les états de l'environnement et la sortie est de deux nœuds correspondant aux valeurs Q pour chacune des deux actions possibles.

Ensuite, le modèle lui-même se compose de deux couches cachées de 64 nœuds chacune avec une activation ReLU. Notre fonction de perte est l'erreur quadratique moyenne car notre sortie est continue et non liée à une plage définie. Enfin, nous utiliserons un optimiseur Adam pour minimiser la fonction de perte. L'optimiseur Adam fonctionne bien sur les données bruyantes stochastiques, qui correspondent à nos données.

Pour les environnements Vizdoom, nous avons construit un réseau de neurones avec trois (3) couches de convolution successives avec une activation ReLU pour réduire la matrice de poids à apprendre. Ensuite, le modèle se compose d’une couche cachée de 800 nœuds avec une activation Relu et une sortie linéaire.

Enfin nous avons opté pour la stratégie epsilon-greedy, au début l’epsilon est à 1 pour avoir suffisamment d’expérience et atteint sa valeur finale qui est de 0.05 à la fin de l'entraînement du modèle.

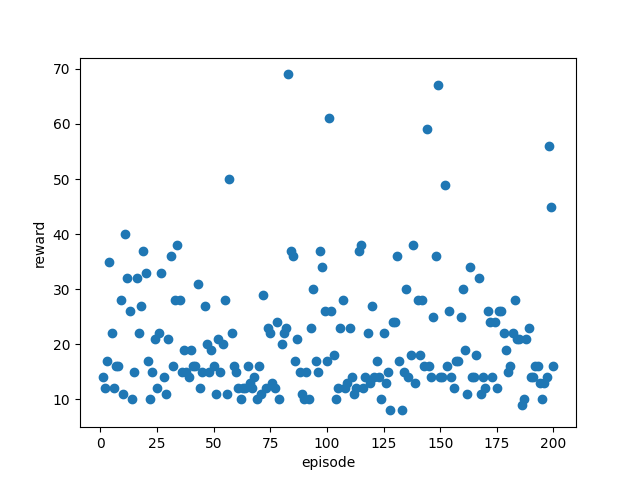
Concernant les paramètres, on s'était appuyé sur un article situé à l’adresse suivant: <https://arxiv.org/pdf/1605.02097.pdf>

Pour le duplicata du modèle, on a choisi la méthode qui consiste à recopier entièrement le modèle vers la cible à chaque fin d’épisode.

1. **Résultat**

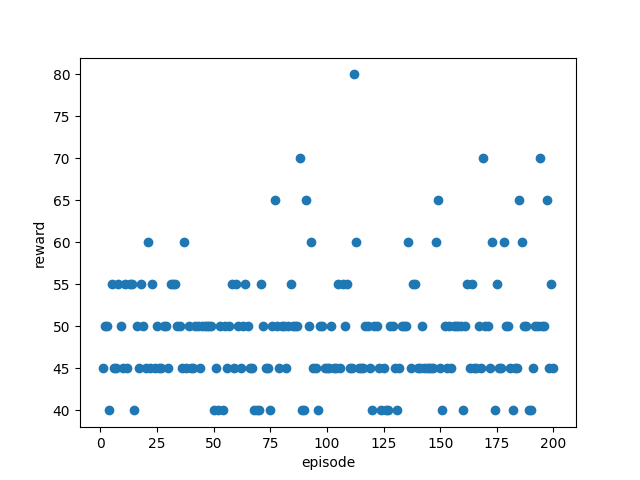
Environnement Cart-Pole:

1. Agent aléatoire



On voit très nettement sur ce graphique qu’avec l’agent aléatoire, on a les récompenses par épisode qui ne sont pas stables dû au fait que les actions sont choisies aléatoirement .

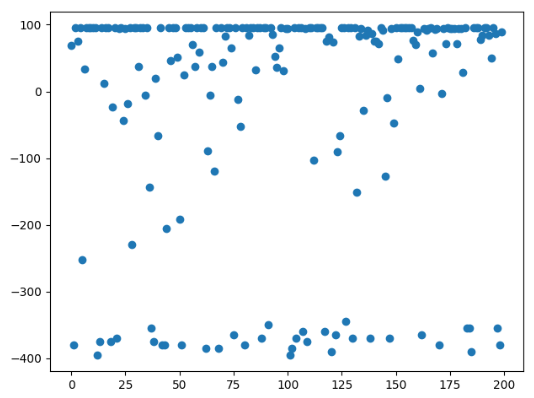
1. Agent Deep RL



Le graphique ci-dessus montre comment évoluent les récompenses sur 200 épisodes en utilisant l’agent Deep RL.

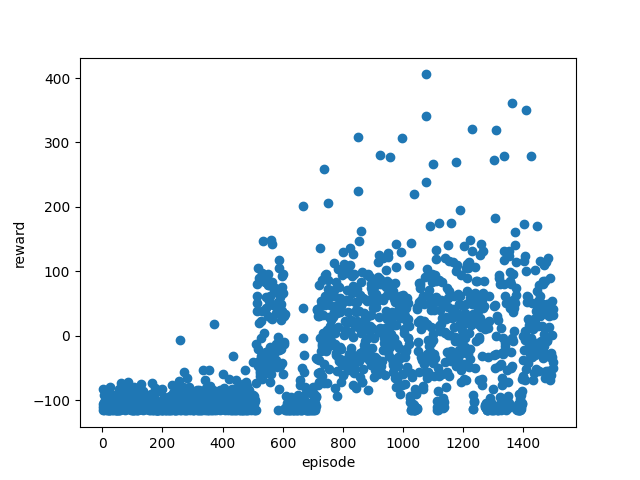
On voit très clairement que les récompenses sont plus élevées en moyenne que l’agent random et on remarque aussi une régularité sur les choix des actions.

Environnement Vizdoom-basic:



L’image ci-dessus illustre la somme des récompenses sur 200 épisodes dans l’environnement vizdoom-basic.

Environnement Vizdoom-Corridor:



Pour l’environnement Vizdoom-Corridor, on a pas réussi à avoir une politique gagnante à 100% mais le graphique nous montre qu’avec plus d'entraînement on pourrait espérer converger vers l’optimum.