

Back to basics: Benchmarking Canonical Evolution Strategies for Playing Atari

Patryk Chrabaszcz, Ilya Loshchilov, Frank Hutter

MU5IN259 - Intelligence Artificielle pour la Robotique

Réalisé par : Wissam AKRETCHÉ et Madina TRAORÉ

Encadré par : Monsieur Olivier SIGAUD



1. Introduction
2. Algorithmes étudiés dans l'article
3. Présentation des résultats obtenus par les auteurs
4. Présentation de nos premiers résultats et comparaison avec ceux obtenus par les auteurs
5. Extensions
6. Présentation de nos résultats finaux

Introduction

Méthodes d'apprentissage par renforcement et stratégies évolutives utilisées pour entraîner l'agent sur le jeu Pong :

- OpenAI ES
- Canonical ES
- CEM
- CMA-ES
- DQN

Algorithmes étudiés dans l'article

- Algorithme évolutionnaire basique
- La démarche est la suivante:
 - Tirer aléatoirement un vecteur de paramètres θ
 - Générer λ nouveaux candidats en ajoutant du bruit gaussien à θ
 - Évaluer chaque candidat
 - Mettre à jour θ en lui affectant la moyenne pondérée des μ meilleurs candidats

- Algorithme évolutionnaire appartenant à la classe NES "Natural Evolution Strategies"
- La démarche est similaire à celle de l'algorithme précédent
- À la dernière étape, les coefficients de θ sont mis à jour en suivant le gradient naturel vers une "fitness" plus élevée

Présentation des résultats obtenus par les auteurs

Présentation des résultats obtenus par les auteurs

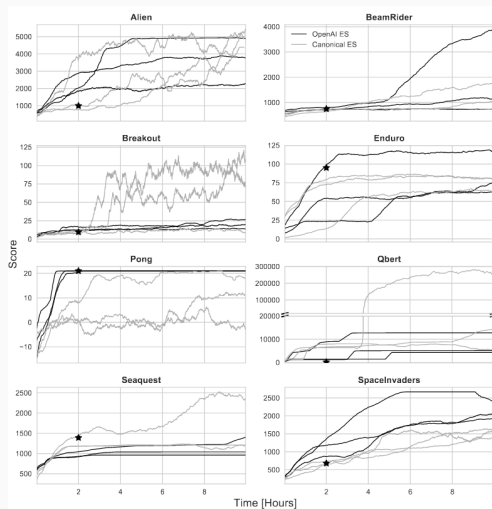


Figure 1: Courbes d'apprentissage obtenues pour les 8 jeux Atari testés par les auteurs : évolution du score en fonction du temps

Présentation des résultats obtenus par les auteurs

| | OpenAI ES 1 hour | OpenAI ES (our) 1 hour | Canonical ES 1 hour | OpenAI ES (our) 5 hours | Canonical ES 5 hours |
|---------------|---------------------|---------------------------|------------------------|----------------------------|-------------------------|
| Alien | 994 | 3040 ± 276.8 | 2679.3 ± 1477.3 | 4940 ± 0 | 5878.7 ± 1724.7 |
| Alien | | 1733.7 ± 493.2 | 965.3 ± 229.8 | 3843.3 ± 228.7 | 5331.3 ± 990.1 |
| Alien | | 1522.3 ± 790.3 | 885 ± 469.1 | 2253 ± 769.4 | 4581.3 ± 299.1 |
| BeamRider | 744 | 792.3 ± 146.6 | 774.5 ± 202.7 | 4617.1 ± 1173.3 | 1591.3 ± 575.5 |
| BeamRider | | 708.3 ± 194.7 | 746.9 ± 197.8 | 1305.9 ± 450.4 | 965.3 ± 441.4 |
| BeamRider | | 690.7 ± 87.7 | 719.6 ± 197.4 | 714.3 ± 189.9 | 703.5 ± 159.8 |
| Breakout | 9.5 | 14.3 ± 6.5 | 17.5 ± 19.4 | 26.1 ± 5.8 | 105.7 ± 158 |
| Breakout | | 11.8 ± 3.3 | 13 ± 17.1 | 19.4 ± 6.6 | 80 ± 143.4 |
| Breakout | | 11.4 ± 3.6 | 10.7 ± 15.1 | 14.2 ± 2.7 | 12.7 ± 17.7 |
| Enduro | 95 | 70.6 ± 17.2 | 84.9 ± 22.3 | 115.4 ± 16.6 | 86.6 ± 19.1 |
| Enduro | | 36.4 ± 12.4 | 50.5 ± 15.3 | 79.9 ± 18 | 76.5 ± 17.7 |
| Enduro | | 25.3 ± 9.6 | 7.6 ± 5.1 | 58.2 ± 10.5 | 69.4 ± 32.8 |
| Pong | 21 | 21.0 ± 0.0 | 12.2 ± 16.6 | 21.0 ± 0.0 | 21.0 ± 0.0 |
| Pong | | 21.0 ± 0.0 | 5.6 ± 20.2 | 21 ± 0 | 11.2 ± 17.8 |
| Pong | | 21.0 ± 0.0 | 0.3 ± 20.7 | 21 ± 0 | -9.8 ± 18.6 |
| Qbert | 147.5 | 8275 ± 0 | 8000 ± 0 | 12775 ± 0 | 263242 ± 433050 |
| Qbert | | 1400 ± 0 | 6625 ± 0 | 5075 ± 0 | 16673.3 ± 6.2 |
| Qbert | | 1250 ± 0 | 5850 ± 0 | 4300 ± 0 | 5136.7 ± 4093.9 |
| Seaquest | 1390 | 1006 ± 20.1 | 1306.7 ± 262.7 | 1424 ± 26.5 | 2849.7 ± 599.4 |
| Seaquest | | 898 ± 31.6 | 1188 ± 24 | 1040 ± 0 | 1202.7 ± 27.2 |
| Seaquest | | 887.3 ± 20.3 | 1170.7 ± 23.5 | 960 ± 0 | 946.7 ± 275.1 |
| SpaceInvaders | 678.5 | 1191.3 ± 84.6 | 896.7 ± 123 | 2326.5 ± 547.6 | 2186 ± 1278.8 |
| SpaceInvaders | | 983.7 ± 158.5 | 721.5 ± 115 | 1889.3 ± 294.3 | 1685 ± 648.6 |
| SpaceInvaders | | 845.3 ± 69.7 | 571.3 ± 98.8 | 1706.5 ± 118.3 | 1648.3 ± 294.5 |

Figure 2: Comparaison des algorithmes OpenAI ES et Canonical ES : moyenne sur 30 sessions d'entraînement du score final obtenu pour chaque jeu. Les résultats significativement meilleurs (test U de Mann-Whitney) sont représentés en bleu.

Présentation de nos premiers résultats et comparaison avec ceux obtenus par les auteurs

Comparaison avec de nos résultats avec ceux des auteurs

| | OpenAI ES (1h) | Canonical ES (1h) | OpenAI ES (5h) | Canonical ES (5h) |
|---|-------------------|----------------------|-------------------|----------------------|
| Résultats obtenus par les auteurs | 21 | 8.2 | 21 | 21 |
| Nos résultats | 15 | 5 | 17 | 16 |

Tableau 1: Comparaison des scores finaux moyens obtenus pour le jeu Pong (21 étant le score maximal)

Extensions

CEM (Cross-Entropy Method)

- Se base sur une optimisation probabiliste appartenant au domaine de l'optimisation stochastique
- Le principe est le suivant :
 - On génère des échantillons de données aléatoires en utilisant un ensemble de paramètres dynamiques
 - On met à jour les paramètres permettant de générer les nouvelles données aléatoires en se basant sur les meilleurs échantillons de la génération courante

CMA-ES (Covariance Matrix Adaptation Evolution Strategy)

- Algorithme évolutionnaire
- Adapte sa moyenne et sa matrice de covariance en s'appuyant sur les meilleures solutions de la génération précédente
- L'algorithme peut ainsi décider d'élargir son espace de recherche lorsque les meilleures solutions sont éloignées les unes des autres ou de le réduire lorsqu'elles sont proches
- Au fil des itérations, la population se concentre autour de l'optimum global

DQN

- Algorithme d'apprentissage par renforcement profond
- Il prend en entrée l'observation courante et retourne les Q-valeurs des couples (état, action)
- Il estime la qualité d'effectuer une action à partir d'un état donné

Présentation de nos résultats fin- aux

Présentation de nos résultats finaux

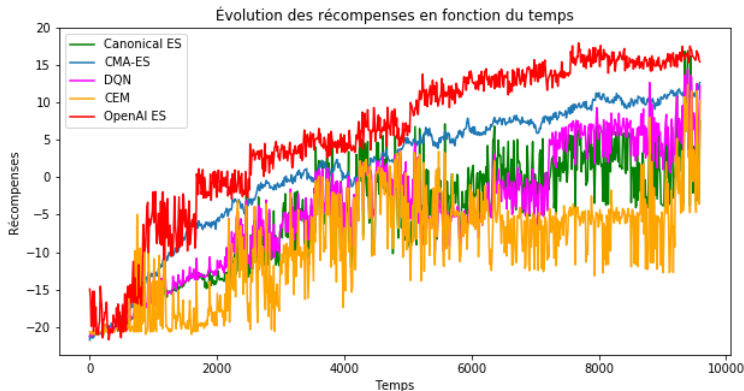






Figure 3: Évolution des récompenses en fonction du temps sur le jeu Pong lors de la phase d'apprentissage

Conclusion

- Les algorithmes évolutionnaires deviennent de réels concurrents aux algorithmes d'apprentissage par renforcement profond
- Combiner les forces des deux méthodes pourrait conduire à des performances inégalées

-  Patryk Chrabaszcz, Ilya Loshchilov, Frank Hutter. *Back to Basics: Benchmarking Canonical Evolution Strategies for Playing Atari*. arXiv:1802.08842.
-  Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, Martin Riedmiller. *Playing Atari with Deep Reinforcement Learning*. arXiv:1312.5602.
-  Brandon Amos, Denis Yarats. *The Differentiable Cross-Entropy Method*. arXiv:1909.12830
-  Saliman et al. *Evolution Strategies as a Scalable Alternative to Reinforcement Learning*. arXiv:1703.03864.
-  Nikolaus Hansen. *The CMA Evolution Strategy: A Tutorial* . arXiv:1604.00772.

Merci pour votre attention