AlphaZero: apprentissage par renforcement et réseaux de neurones à convolutions pour les jeux de plateau

Rémi Coulom

Juin 2018

Janvier 2016: Le choc AlphaGo

Des résultats spectaculaires

- Domine les meilleurs concurrents (99.8% de victoires)
- Première victoire contre un joueur humain professionnel (Fan Hui, champion d'Europe)

Mars 2016: Défaite d'un champion légendaire

Match contre Lee Sedol

- Victoire 4-1 pour la machine
- Choc immense dans le monde du go

Octobre 2017: AlphaGo Zero, Décembre 2017: AlphaZero

Apprendre à partir de zéro

- AlphaGo apprenait à imiter des parties d'experts
- AlphaGo Zero apprend des parties qu'il joue contre lui-même
- AlphaZero généralise l'approche aux échecs et au shogi

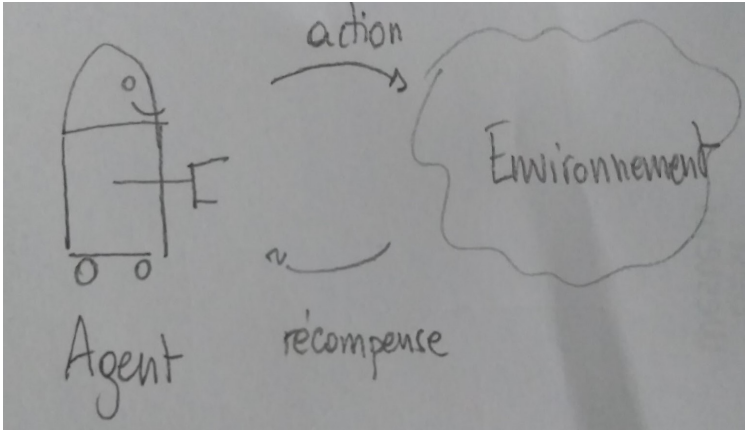
Règles $\rightarrow$ AlphaZero $\rightarrow$ IA super forte

- go
- échecs
- shogi
- gomoku

Plan de l'exposé

- L'algorithme AlphaZero
- Application au morpion

Apprentissage par renforcement



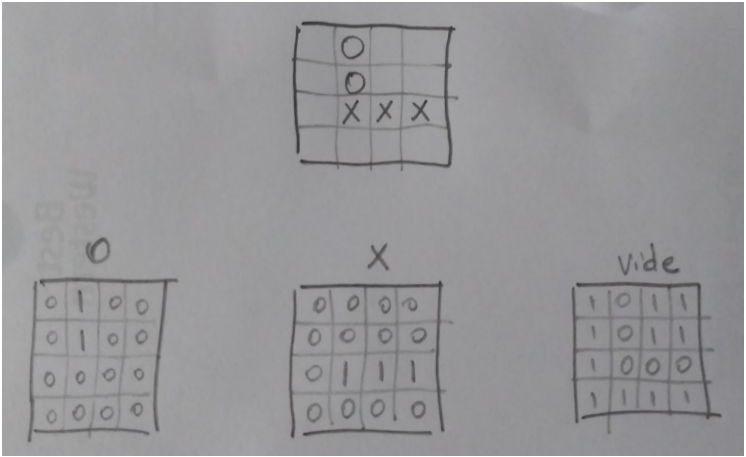
Acteur et critique

action a, état s:

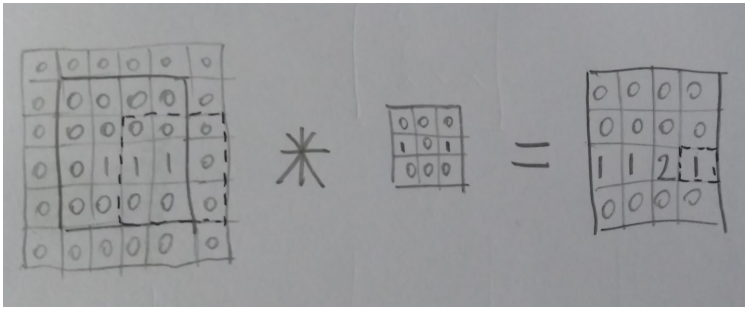
- acteur: $\pi(s, a)$, probabilité de a dans s
- critique: $V(s)$, espérance de récompense dans l'état s

Pour AlphaZero: un seul réseau de neurones, à la fois acteur et critique

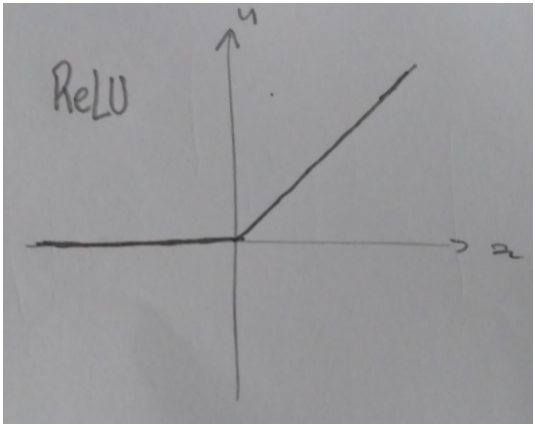
Architecture du réseau: entrées



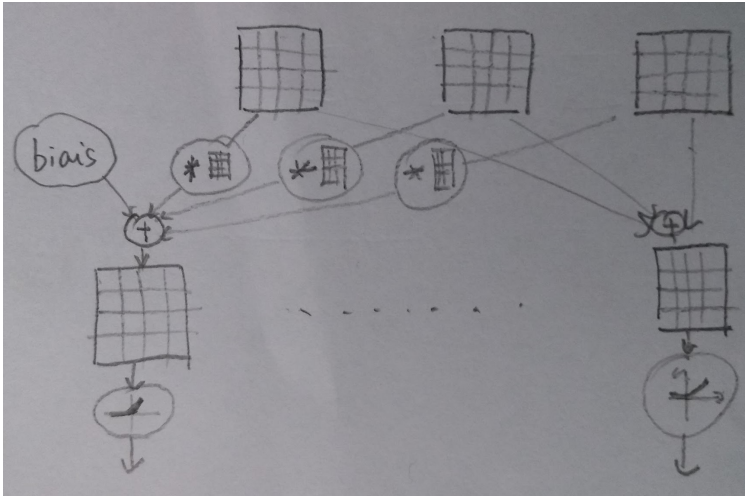
Architecture du réseau: convolution



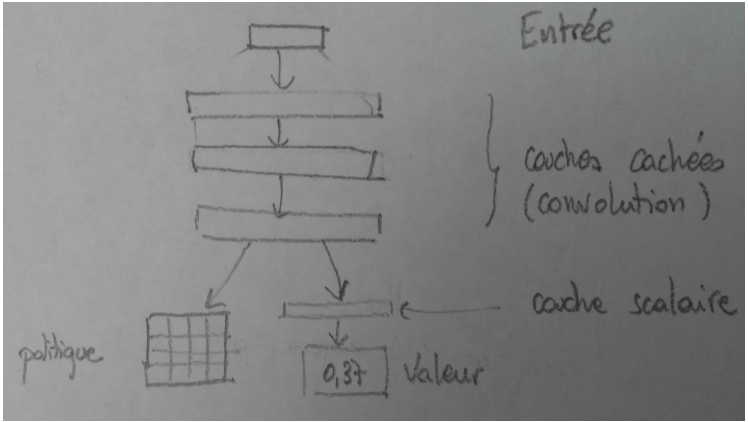
Architecture du réseau: ReLU



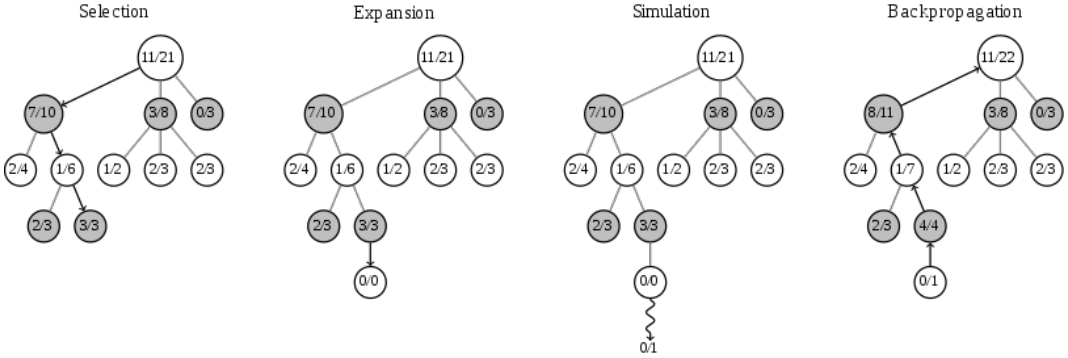
Architecture du réseau: une couche



Architecture du réseau: architecture complète



MCTS (Monte Carlo Tree Search)



MCTS avec un réseau de neurones

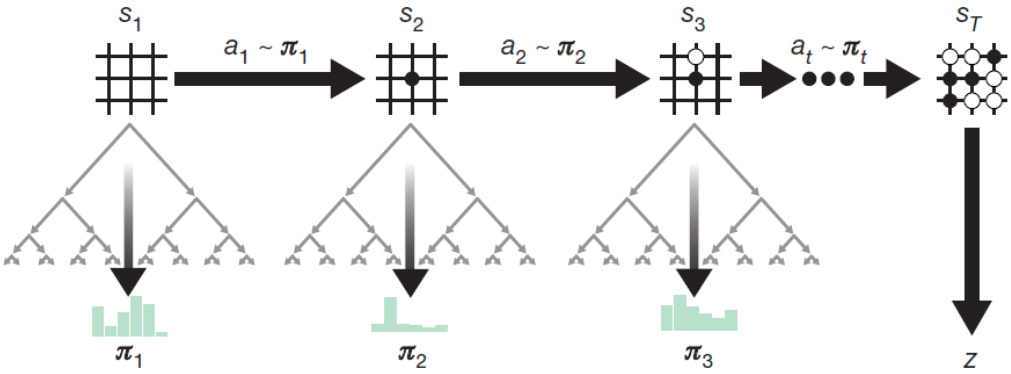
- On remplace la partie aléatoire par l'évaluation du réseau.
- La politique permet de guider la croissance de l'arbre.
- Faire la moyenne des évaluations est meilleur que min-max.

Selection de l'action a

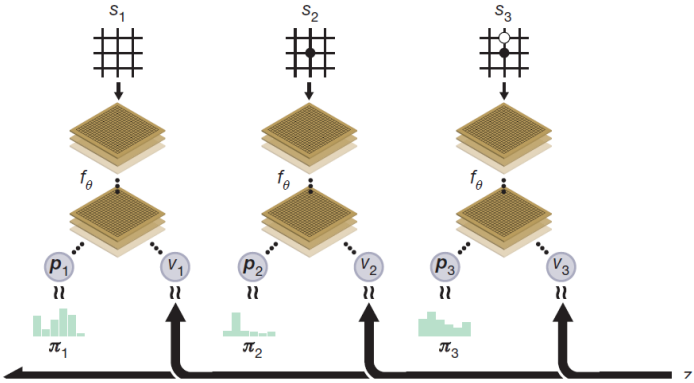
$$\max_a Q(s,a) + c\pi(s,a)\frac{\sqrt{\sum_b N(s,b)}}{1+N(s,a)}$$

- $Q(s,a)$: moyenne des évaluations
- c: coefficient d'exploration
- $\pi(s,a)$: probabilité de a
- $N(s,a)$: nombre de visites de a

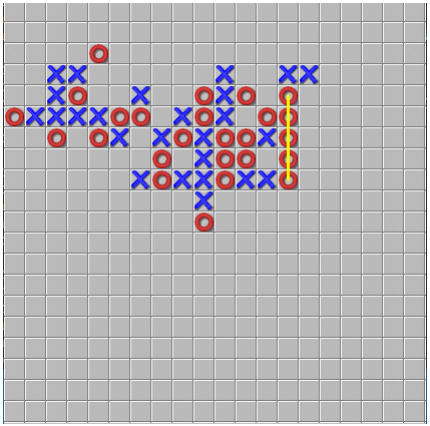
AlphaZero: 1. Jouer contre soi-même



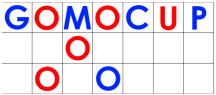
AlphaZero: 2. Apprendre



Application au morpion (gomoku narabe)



- Très populaire en Asie et Europe de l'Est.
- Les meilleures programmes ne battent les meilleurs humains que depuis très récemment.
- Une compétition annuelle entre programmes: la Gomocup.



Victoire contre Yixin

Le match

- Yixin: Vainqueur de la Gomocup, 7 fois consécutivement
- 5 secondes par coup, GPU: GTX 960M, CPU: i7-6700HQ à 2.60 GHz
- 41 ouvertures de piskvork, 1 partie de chaque couleur
- Résultat: victoire 48-34 (8 fois 2-0, 1 fois 0-2, 32 fois 1-1).

Conclusion

- Une méthode générique pour les jeux de plateau
- Pas nécessaire de construire des heuristiques à la main
- Fonctionne hyper bien