Histoire de la modélisation du Go Apprentissage et Recherche Monte Carlo

Journée de la CERNA

12 Juin 2016

Tristan Cazenave

LAMSADE

Université Paris-Dauphine

Tristan.Cazenave@dauphine.fr

Plan

- La préhistoire : les années 60-70
- Patterns et Alpha-Béta : les années 80-90
- Recherche Monte-Carlo : les années 2000
- Deep Learning depuis 2015
- Recherche Monte Carlo et Apprentissage pour l'optimisation

La préhistoire

- Le premier programme de Go a été écrit par D. Lefkovitz en 1960.
- Le premier article scientifique sur la programmation du Go a été publié en 1963 par Remus.
- Le sujet du papier était l'application du machine learning au Go.
- Le premier programme de Go à battre un être humain débutant complet est celui d'Albert Zobrist en 1969.
- Le principe était de calculer une fonction de potentiel qui approximait l'influence des pierres.

La préhistoire

- Un apport important de Zobrist fut sa fonction de hachage.
- Le hachage de Zobrist a été repris par pratiquement tous les programme de jeu de plateau.
- Le principe est d'associer un nombre aléatoire à chaque intersection et à chaque couleur.
- La hachage d'une position est le XOR des nombres aléatoires correspondant à la position.

La préhistoire

- La deuxième thèse sur la programmation du Go fut celle de Ryder en 1971.
- Les premiers programmes utilisaient tous une fonction d'influence.
- Une pierre irradiait de l'influence sur les intersections voisines.
- Les pierres noires irradiaient une influence de signe opposé à l'influence des pierres blanches.
- L'influence décroissait avec la distance aux pierres.
- Par exemple dans Many Faces of Go l'influence était de 1/distance.

La préhistoire

- Bruce Wilcox a écrit le premier programme de Go à jouer mieux qu'un débutant complet.
- Il utilisait des représentations abstraites du damier et raisonnait sur les groupes de pierres.
- Sa théorie des sector lines divisait le damier en zones qui étaient ensuite analysées.

Patterns et Alpha-Béta

Patterns et Alpha-Béta

- La génération suivante de programmes de Go utilisait les patterns pour conseiller des coups.
- Ils utilisaient aussi des Alpha-Béta sélectifs pour les calculs tactiques.
- Les premières compétitions de programmes de Go sont apparues.

Patterns et Alpha-Béta

- Plusieurs modules spécialisés sur les différents aspects du jeu :

  Fuzeki, Joseki, santé des groupes, vie et mort, capture, connexions, formes…
- Le nombre de connaissances à avoir pour s'améliorer croît avec le niveau du programme.

Recherche Monte-Carlo

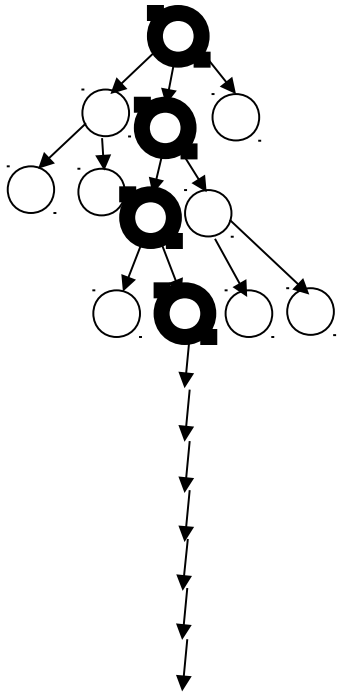
Monte-Carlo Go

- 1993 : premier programme de Monte Carlo Go
  - Gobble, Bernd Bruegmann.
  - Comment la nature jouerait elle au jeu de Go ?
  - Recuit simulé sur deux listes de coups
  - Statistiques sur les coups.
  - Une unique règle : ne pas se boucher un oeil.
  - Résultat : Programme de Go 9x9 moyen.
  - Avantage : très simple.

UCT

- UCT : Dilemme Exploration/Exploitation pour les arbres (2006).
- Jouer des parties aléatoires (playouts).
- Exploitation : choix du coup qui maximise la moyenne des parties aléatoires commençant par ce coup.
- Exploration : ajouter un terme de regret (UCB).

UCT



1) descente de l'arbre

2) partie aléatoire

3) mise à jour de l'arbre

RAVE

- Une amélioration importante pour le Go, Hex et d'autres jeux :

  Rapid Action Value Estimation (RAVE).

- RAVE combine la moyenne des parties aléatoires qui commence avec le coup avec la moyenne des parties aléatoires qui contiennent le coup.

GRAVE

- Generalized Rapid Action Value Estimation (GRAVE) est une modification simple de RAVE.
- Elle utilise le dernier nœud de l'arbre qui a plus de n parties aléatoires pour calculer les valeurs RAVE.
- C'est une amélioration importante de RAVE pour le Go, Atarigo, Knightthrough et Domineering.

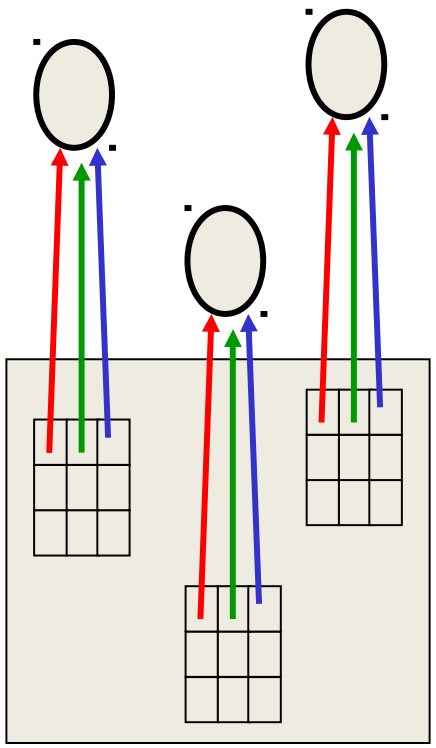
MCTS

- La recherche Monte-Carlo a permis une amélioration notable des programmes de Go.

Deep Learning

Les caractéristiques partagées
(l'approche actuellement dominante des réseaux de neurones)

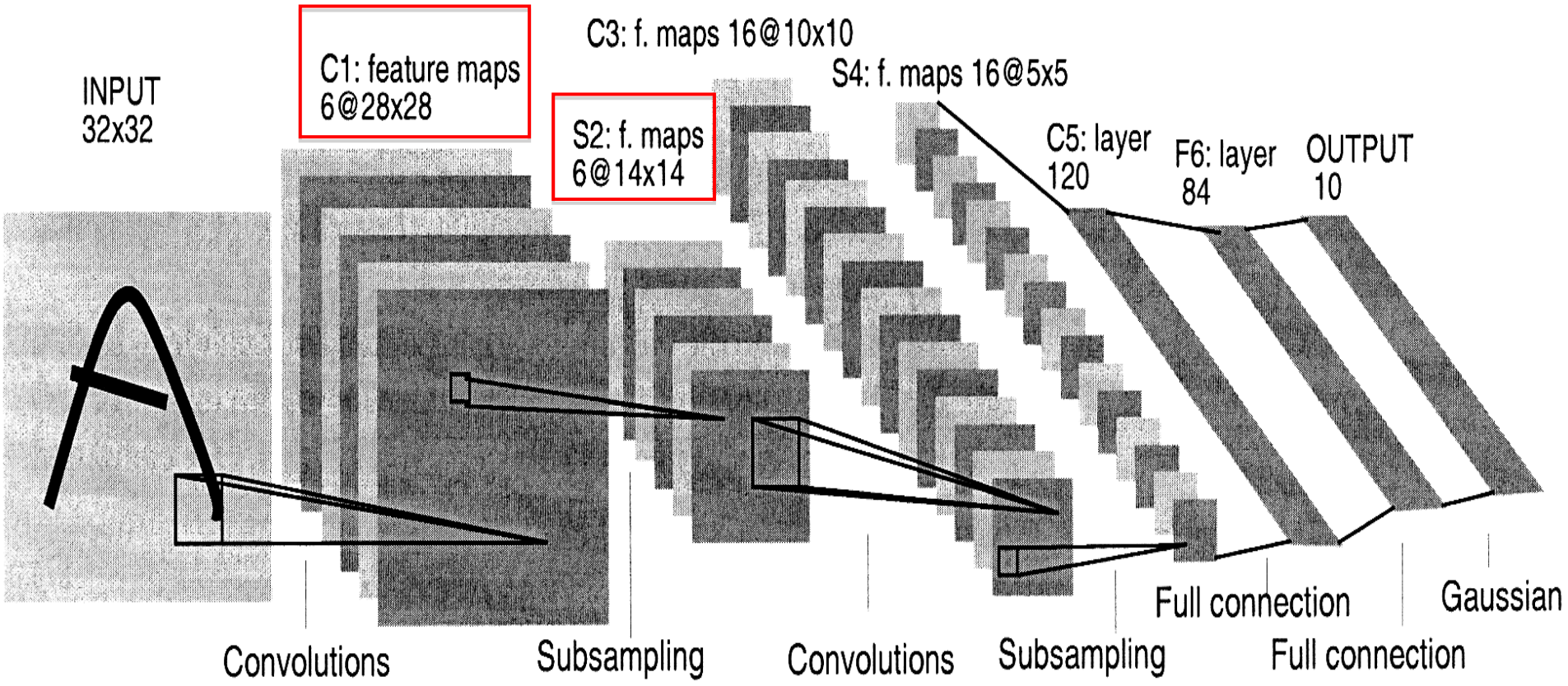
- On utilise des copies du même détecteur de caractéristique pour les différentes positions.
- Cela permet de réduire le nombre de paramètres à apprendre.
- On utilise plusieurs plans de caractéristiques qui ont chacun leur propres détecteurs.

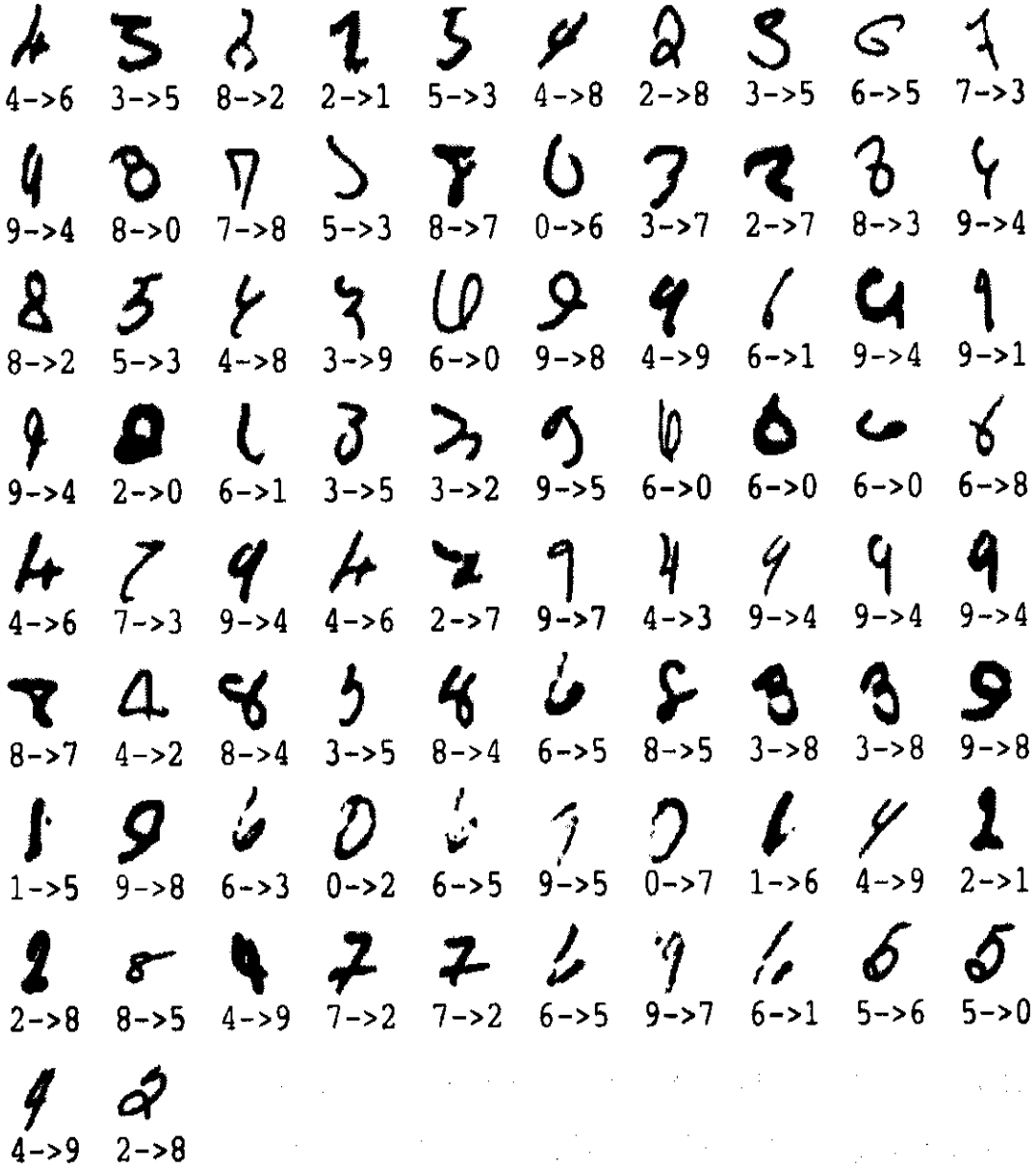


Le Net

- Yann LeCun et ses collaborateurs ont dévelopé des réseaux de reconnaissance de l'écriture manuscrite qui ont obtenus des performances remarquables.
- Ils comportent de nombreuses couches cachées.
- De petits réseaux sont répliqués sur toute l'image.
- Le Net a été utililsé pour lire 10% des chèques en amérique.
- http://yann.lecun.com

L'architecture de LeNet5





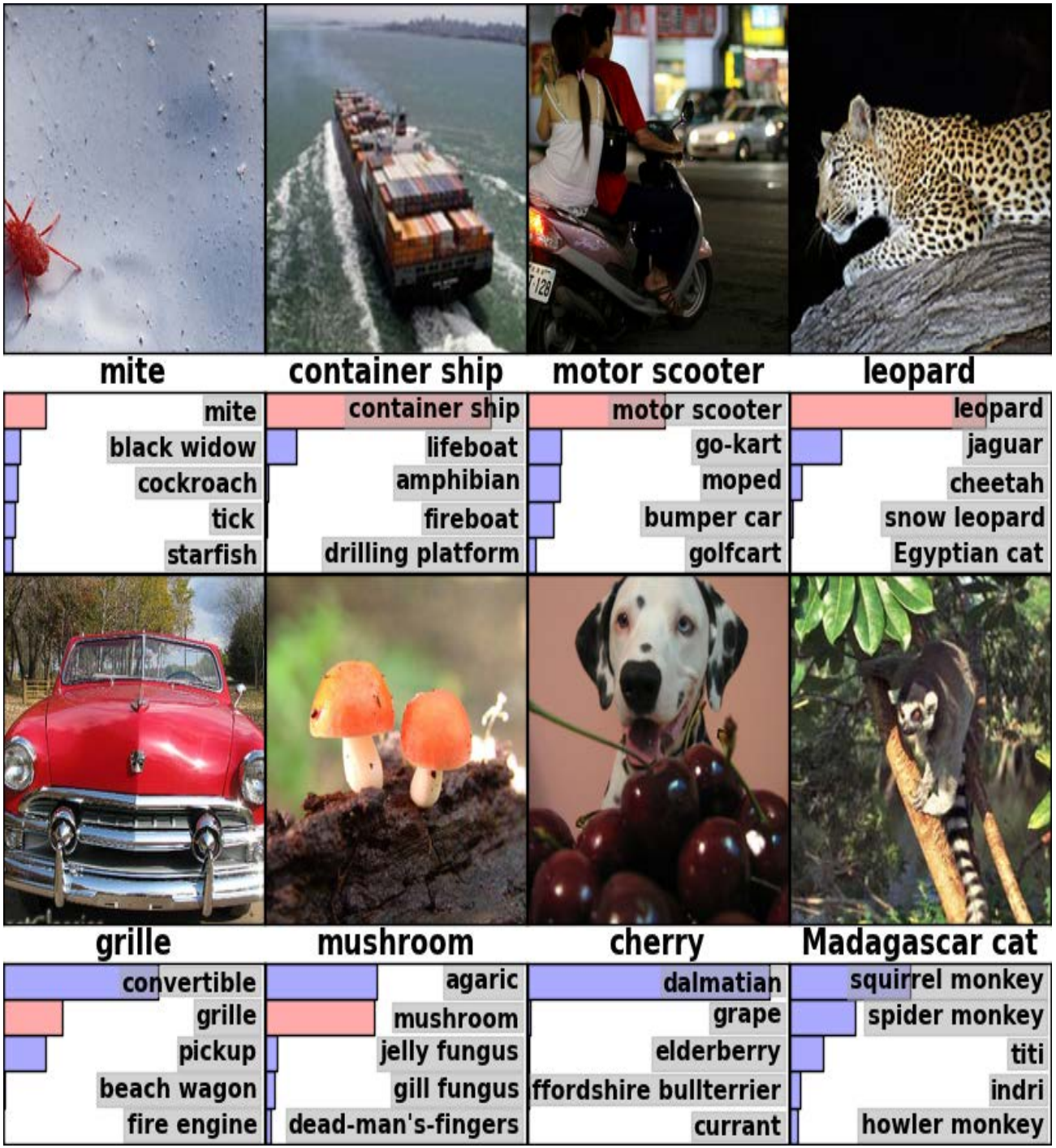
Les 82 erreurs de LeNet5

Les erreurs sont différentes de celles faites par les humains.

Les humains font de l'ordre de 30 erreurs.

La compétition ImageNet

- L'ensemble d'apprentissage contient 1.2 million d'images haute résolution.
- Les méthodes classiques de reconnaissance des formes ont été testées sur cet ensemble.
- Le but est de trouver la classe de l'image dans les 5 meilleures sorties.
- Alex Krizhevsky a dévelopé un réseau profond du même type que ceux utilisés par Yann Le Cun.
- Il utilise 5 couches convolutionnelles suivies de 2 couches complétement connectées.



Exemples d'images avec la sortie du réseau de neurones

La compétition ImageNet

- University of Tokyo 26.1 %
- Oxford University Computer Vision Group 26.9 %
- INRIA + Xerox 27.0 %
- University of Amsterdam 29.5 %

- University of Toronto (Alex net) 16.4 %

Le hardware

- Les réseaux convolutionnels sont gérés très efficacement par les cartes graphiques de Nvidia (avec Cuda).
- Elles contiennent des milliers de coeurs.
- Elles sont très efficaces pour les multiplications de matrices.
- Les échanges mémoires sont très rapides.
- Alex net a ainsi pu être entrainé en une semaine.
- Progrès incessant des cartes graphiques, moins chères, plus de coeurs, plus rapides => progrès du deep learning.

AlphaGo

Lee Sedol est le plus célèbre et le plus titré des joueurs 9p :

AlphaGo a gagné 4-1 contre Lee Sedol.

AlphaGo

- AlphaGo combine Monte-Carlo et Deep Learning.
- Il utilise des réseaux profonds de 13 couches.
- Ils sont convolutionnels.
- Ils ont de 128 à 256 plans de caractéristiques.
- Il utilise de nombreuses CPU et GPU en parallèle.

AlphaGo

- AlphaGo a été développé en quatre phases :
- Apprentissage sur des parties de joueurs forts => policy network.
- Il joue contre différentes versions de lui-même pour améliorer le policy network.
- Il apprend ensuite un value network qui évalue les positions à partir de millions de parties jouées contre lui-même.
- Il combine Monte-Carlo, policy et value network.

AlphaGo

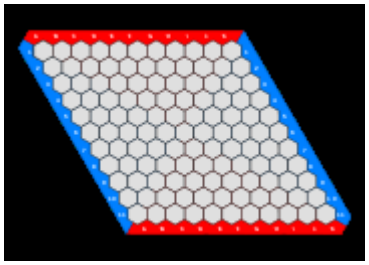
- Le policy network est utilisé pour essayer les meilleurs coups en premier au début de la recherche Monte-Carlo.
- Les parties aléatoires sont utilisées pour évaluer les coups.
- La probabilité de gain donnée par le value network est combinée avec le résultat des parties aléatoires.

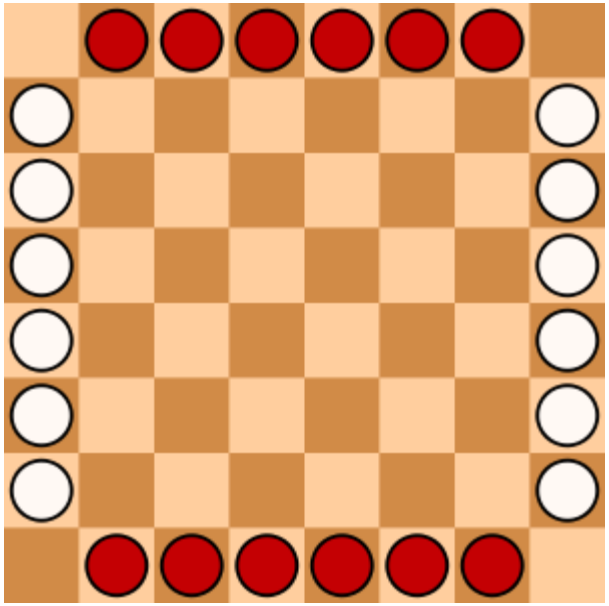
MCTS

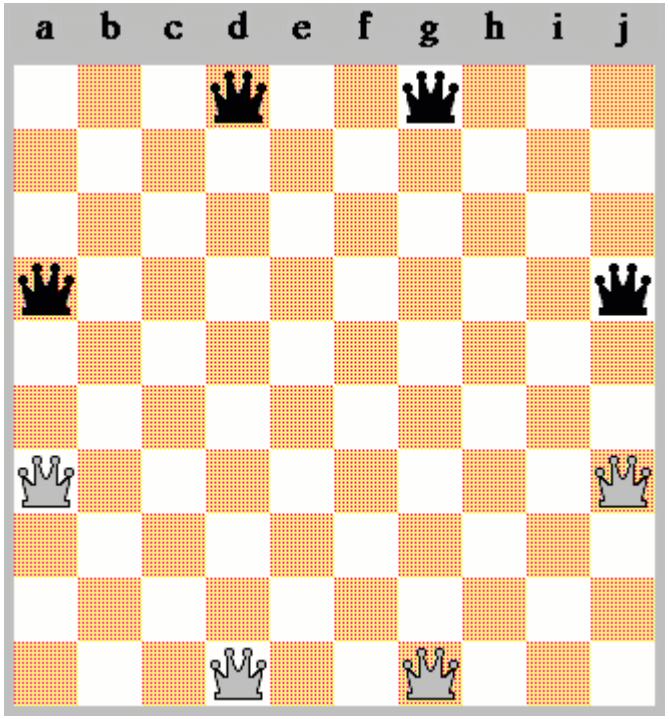
- General Game Playing = jouer à un jeu dont on ne connaît que les règles.
- Competition annuelle organisée par Stanford.
- Ary champion du monde en 2009 et 2010.
- UCT depuis 2007.

Autres jeux

- Hex : 2009
- Amazons : 2009
- Lines of Action : 2009



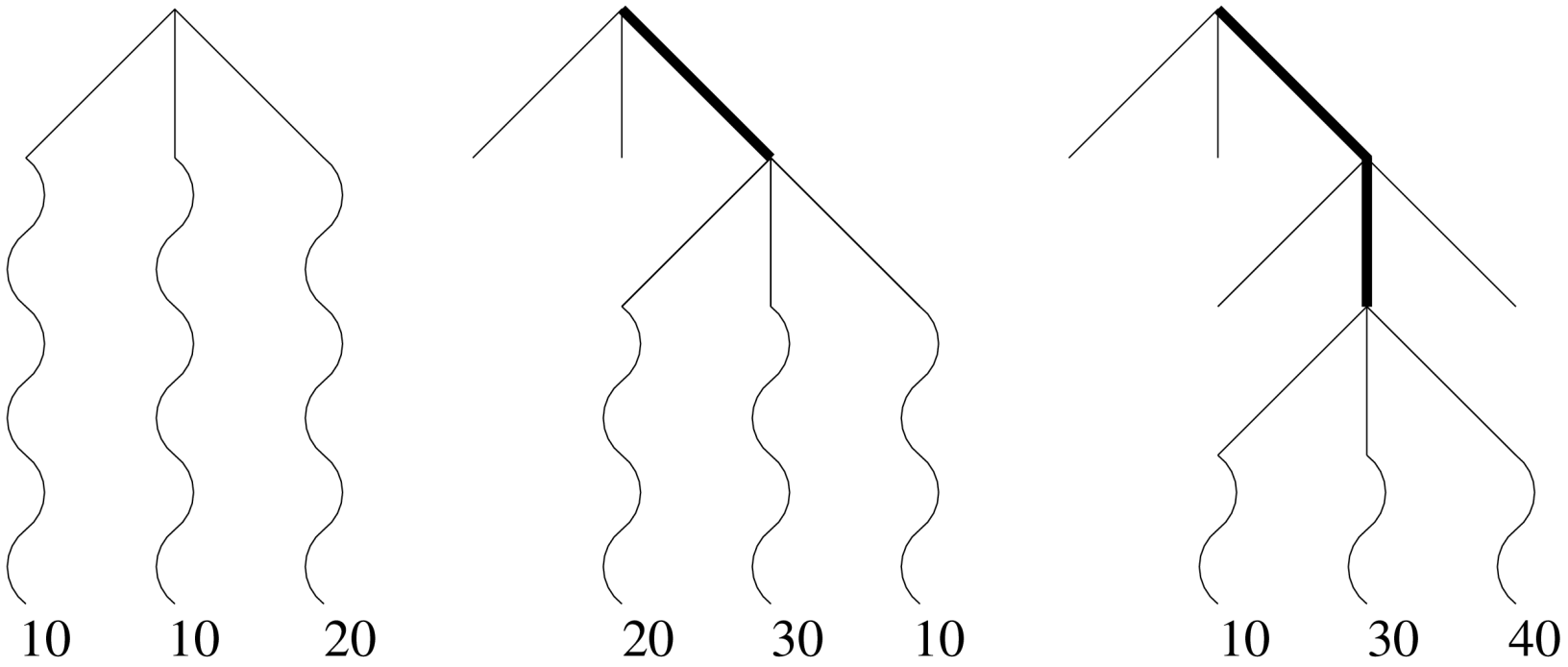




Single Agent Monte Carlo

- UCT can be used for single-agent problems.
- Nested Monte Carlo Search often gives better results.
- Nested Rollout Policy Adaptation is an online learning variation that has beaten world records.

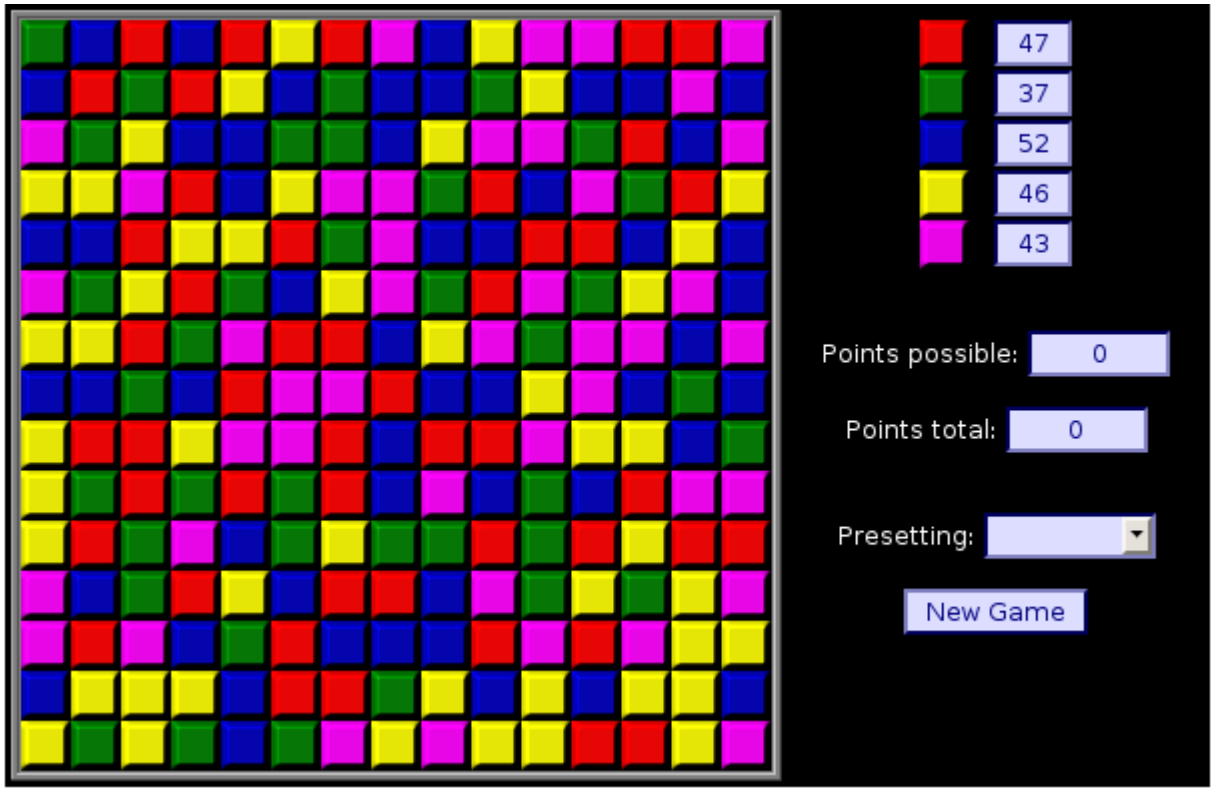
Nested Monte-Carlo Search



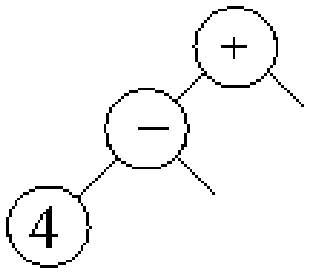
Nested Monte-Carlo Search

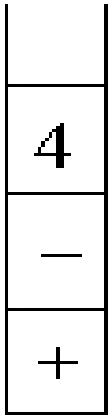
- Play random games at level 0
- For each move at level n+1, play the move then play a game at level n
- Choose to play the move with the greatest associated score
- Important : memorize and follow the best sequence found at each level

Same Game


47
37
52
46
43
Points possible:
0
Points total:
0
Presetting:
New Game

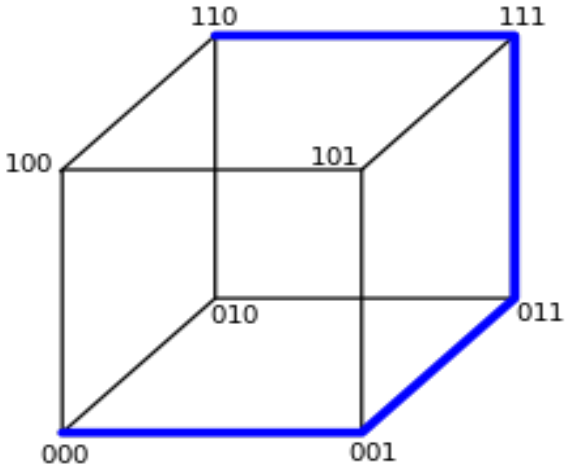
Monte-Carlo Discovery of Expressions





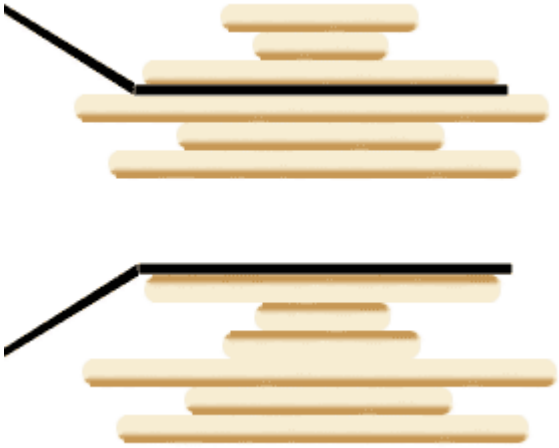
- Possible moves are pushing atoms.
- Evaluation of a complete expression.
- Better than Genetic Programming for some problems [Cazenave 2010, 2013].

Snake in the box



- A path such that for every node only two neighbors are in the path.
- Applications: Electrical engineering, coding theory, computer network topologies.
- World records with NMCS [Kinny 2012].

The Pancake Problem



- Nested Monte Carlo Search has beaten world records using specialized playout policies [Bouzy 2015].

Nested Rollout Policy Adaptation

- NRPA is NMCS with policy learning.
- It uses Gibbs Sampling as a playout policy.
- It adapts the weights of the moves according to the best sequence of moves found so far.
- During adaptation each weight of a move of the best sequence is incremented and the other moves in the same state are decreased proportionally to their weights.

Nested Rollout Policy Adaptation

- Each move is associated to a weight w_i.
- During a playout each move is played with a probability:

$$\exp(w_i) / \Sigma \exp(w_k)$$

Nested Rollout Policy Adaptation

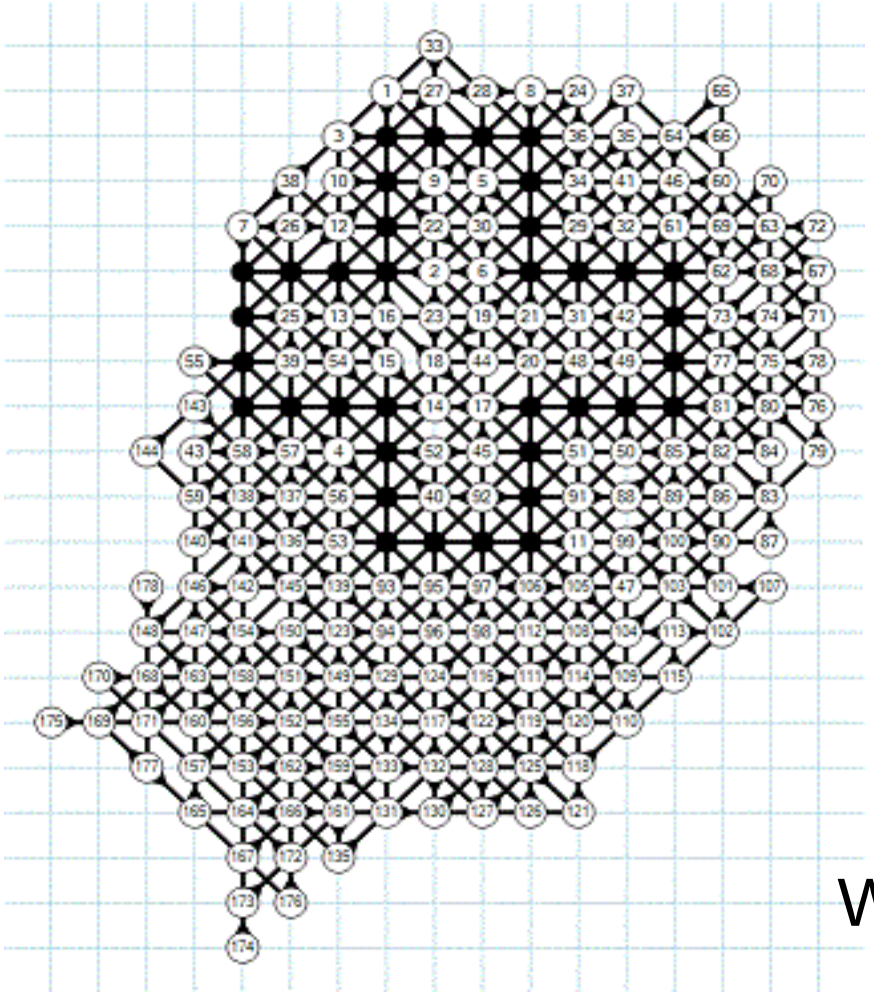
- For each move of the best sequence:

  $w_i = w_i + 1$

- For each possible move of each state of the best sequence:

  $w_i = w_i - \exp(w_i) / \Sigma \exp(w_k)$

Morpion Solitaire



World record [Rosin 2011]

Applications

- State of the art results for Logistics [Edelkamp & al. 2016].
- Improvement of some alignments for Multiple Sequence Alignment [Edelkamp & al 2015].

Playout Policy learning

Playout policy learning for games :

- Start with a uniform policy.
- Use the policy for the playouts.
- Adapt the policy for the winner of each playout.

Experimental results

	Size	Playouts	
		1,000	10,000
Atarigo	8 x 8	72.2	94.4
Breakthrough	8 x 8	55.2	54.4
Misere Breakthrough	8 x 8	99.2	97.8
Domineering	8 x 8	48.4	58.0
Misere Domineering	8 x 8	76.4	83.4
Knightthrough	8 x 8	64.2	64.6
Misere Knightthrough	8 x 8	99.8	100.0
Nogo	8 x 8	64.8	46.4
Misere Nogo	8 x 8	80.6	89.4

Conclusion

La recherche Monte Carlo combinée à l'apprentissage donne des résultats meilleurs que les autres algorithmes pour de nombreux problèmes :

- Jeux
- Puzzles
- Découverte de formules
- Snake in the box
- Pancake
- Logistique
- Alignement de séquences