AlphaGo

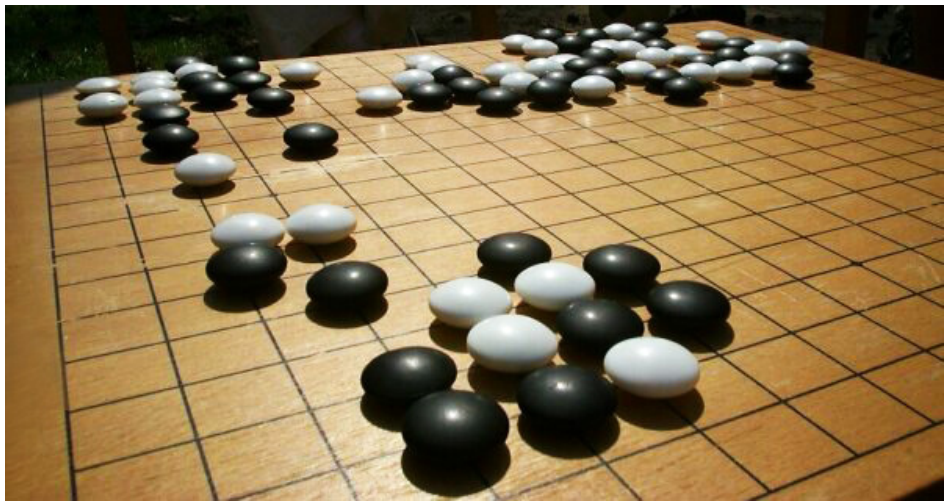
Линдеманн Никита

20 апреля 2019 г.

Игра Go

Go

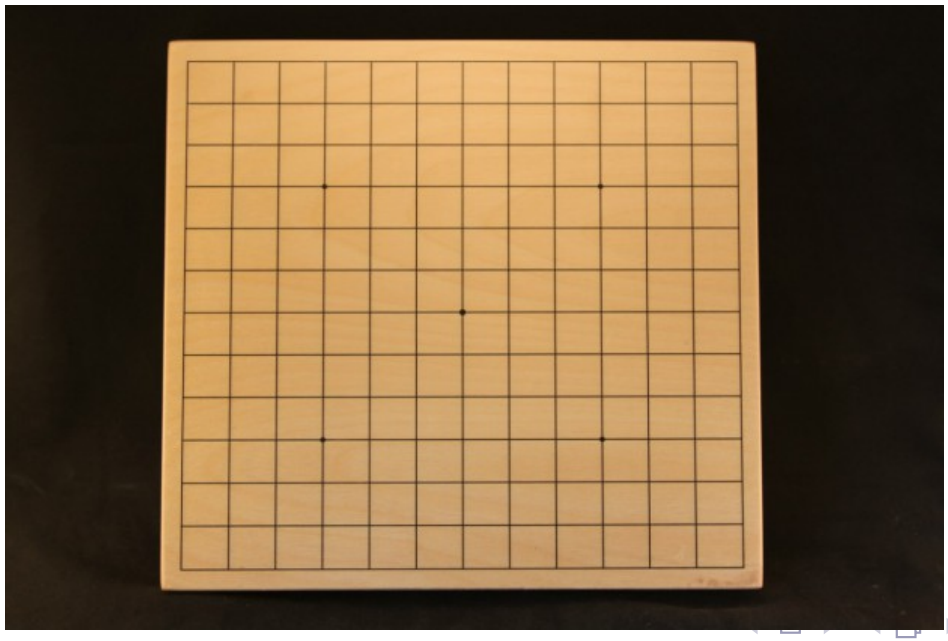
Го – логическая настольная игра с глубоким стратегическим содержанием, возникшая в Древнем Китае, по разным оценкам, от 2 до 5 тысяч лет назад.



Правила Go

Правила игры в Go

Играют два игрока, один из которых получает чёрные камни, другой — белые. Цель игры — отгородить на игровой доске камнями своего цвета большую территорию, чем противник.
Перед началом игры доска пуста. Первыми ходят чёрные. Затем белые. Далее ходы делаются по очереди. Размеры игрового поля бывают 9×9, 13×13 и 19×19.



Предшественники

Deep Blue (1997 год)

1. Библиотека дебютов
2. 8000 настраеваемых признаков оценки позиции
3. Форсированный вариант, итеративное углубление, таблицы перестановок

Новизна Alpha Go

AlphaGo (2015 год)

1. Нет дебютной базы
2. Нейронная сеть, обученная на большом количестве партий
3. Неизвестно, как оценивается позиция (феномен эмерджентности)
4. Глубокое обучение

Основные понятия

Глубокое обучение (англ. Deep learning) — совокупность методов машинного обучения (с учителем, с частичным привлечением учителя, без учителя, с подкреплением), основанных на обучении представлениям (англ. feature/representation learning), а не специализированным алгоритмам под конкретные задачи.

Обучение с подкреплением (англ. reinforcement learning) — один из способов машинного обучения, в ходе которого испытуемая система (агент) обучается, взаимодействуя с некоторой средой.

Общие подходы к играм с полной информацией

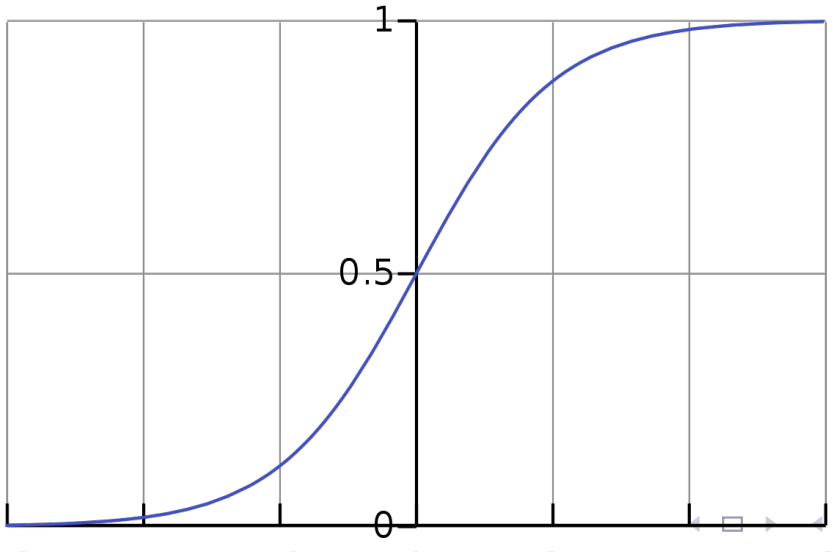
Го является игрой с полной информацией. Теоретически для любой игры с полной информацией существует оптимальная стратегия. Чтобы найти оптимальную стратегию, нужно обойти полное дерево игры. Для большинства игр этот метод непрактичен, так как размер дерева может быть очень большим. Его можно оценить как b^d, где b – степень ветвления дерева игры (то есть примерное число возможных ходов в каждой позиции), а d – глубина дерева игры (то есть примерная длина партии). Для го $b \approx 250$, $d \approx 150$, в то время как для шахмат $b \approx 35$, $d \approx 80$.

Оптимизация

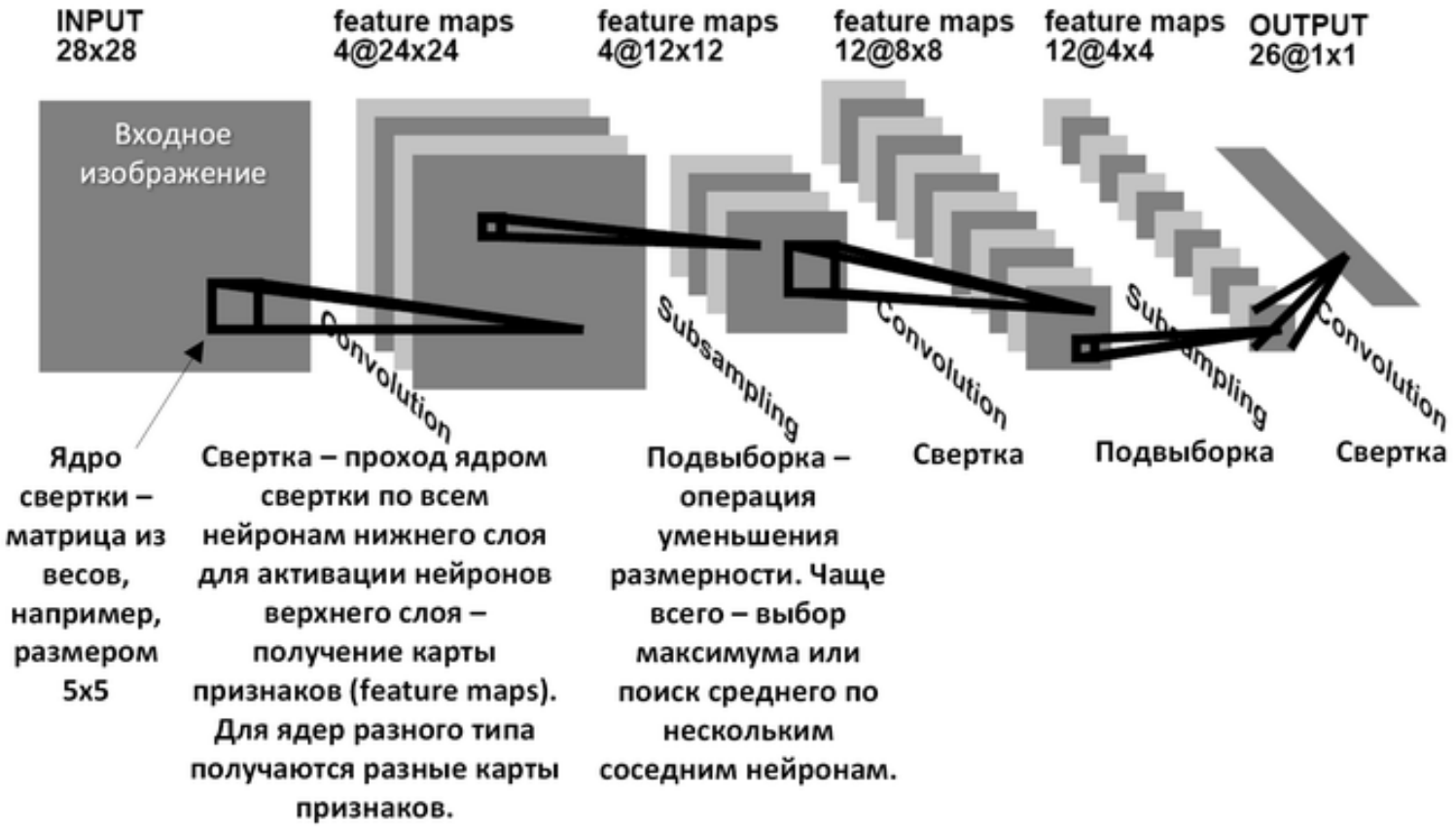
1. Оценочная функция: вместо того, чтобы рассматривать игру до конца, можно оценить промежуточную позицию (шахматы)
2. Сокращении степени ветвления просматриваемого дерева за счёт отбрасывания некоторых ходов (метод Монте-Карло)

AlphaGo и распознование образов

AlphaGo работает, используя сверточные нейронные сети. Они состоят из нескольких уровней нейронов. Каждый уровень получает на вход матрицу чисел, комбинирует их с некоторыми весами и, используя нелинейную функцию активации (например, логистическую функцию активации), выдаёт множество чисел на выходе, которые передаются на следующий уровень. Нейронные сети обучают на большом количестве изображений, постоянно корректируя веса, используемые для вычисления результата.



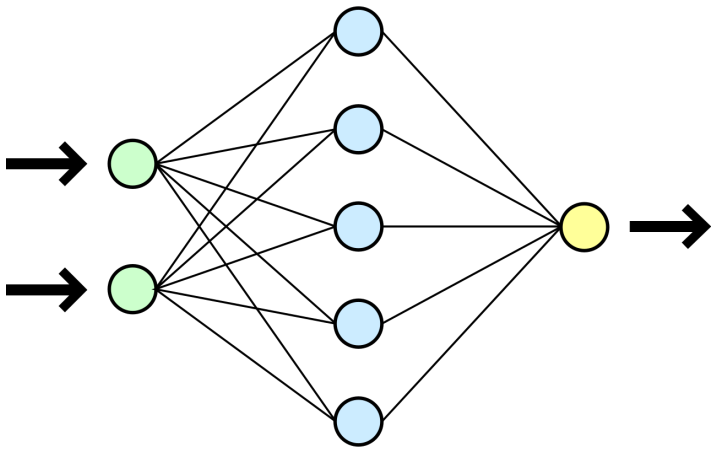
AlphaGo и распознование образов



Архитектура свёрточной нейронной сети

Концепт AlphaGo

AlphaGo использует свёрточные нейронные сети для того, чтобы оценить позицию или предсказать следующий ход. AlphaGo подаёт на вход нейронной сети позицию. Каждая позиция представлена как многослойная картинка 19×19, где каждый слой представляет описания простых свойств каждого пункта доски. Свойства – цвет камня, взятие камней и тд. Единственное нетривиальное свойство – это угрожает ли данной группе захват в лестницу. Всего используется 48 бинарных свойств. Таким образом каждая позиция представлена в виде таблицы $19 \times 19 \times 48$.



Стратегическая сеть

Для того, чтобы не рассматривать совсем плохие ходы, и тем самым сократить степень ветвления при поиске, AlphaGo использует стратегические сети (англ. policy networks) — нейронные сети, которые помогают выбирать хороший ход.
Одна из таких сетей (SL policy networks) может предсказывать ход, который в данной позиции сделал бы профессионал. Это 13-уровневая нейронная сеть получена обучением «с учителем»на 30 миллионах позиций. В качестве обучающего алгоритма использовался стохастический градиентный спуск для поиска максимального правдоподобия. Получившаяся нейронная сеть вычисляла распределение вероятностей среди всех возможных ходов в данной позиции. В результате нейронная сеть смогла правильно предсказывать ход, который выбрал человек, в 57% тестовых ситуациях (не использованных при обучении). Для сравнения лучший результат до AlphaGo был 44%.

Улучшенная стратегическая сеть

Стратегическая сеть была улучшена при помощи обучения с подкреплением, а именно сеть постоянно улучшалась, играя с одной из сетей, полученных ранее. В результате получилась стратегическая сеть (RL policy network), которая выигрывала у первоначальной сети 80% игр. Оказалось, что полученная стратегическая сеть, смогла выиграть 85% игр у самой сильной на то время открытой программы Pachi. Для сравнения, до этого лучшая программа, которая играла, не используя перебор вариантов, а только свёрточную нейронную сеть, выигрывала у Pachi 11% игр.

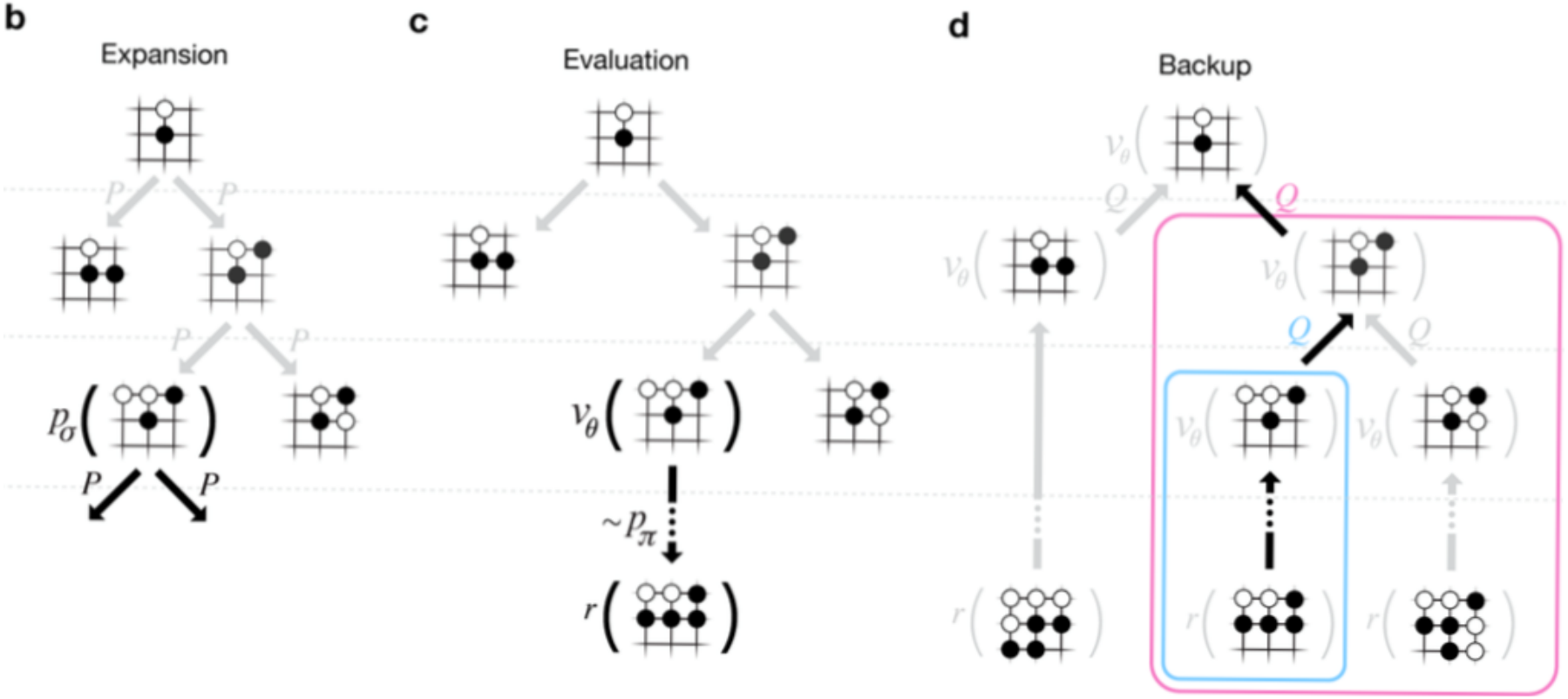
Оценочная сеть

Для сокращения глубины поиска AlphaGo использовала оценочную сеть (англ. value network). Эта нейронная сеть оценивает вероятность выигрыша в данной позиции. Эта сеть является результатом обучения на 30 миллионах позиций, полученных при игре улучшенной стратегической сети с собой. Для каждой из этих позиций оценивалась вероятность выигрыша методом Монте-Карло: устраивался турнир из многих партий, в которых улучшенная стратегическая сеть, построенная на прошлом этапе, играла сама с собой, начиная с этой позиции. После этого оценочная сеть была обучена на этих данных. Обучение заняло одну неделю на 50 GPU. В результате получилась сеть, которая могла предсказывать для каждой позиции вероятность выигрыша, при этом используя в 15000 раз меньше вычислений, чем метод Монте-Карло.

AlphaGo в игре

AlphaGo осуществляет перебор вариантов при помощи метода Монте-Карло для поиска в дереве следующим образом. AlphaGo строит частичное дерево игры, начиная с текущей позиции, производя многочисленные симуляции игры. Для каждого хода в дереве записывается оценка, которая специальным образом зависит от оценок хода, полученных при помощи стратегической и оценочной сетей, от результата случайных партий в предыдущих симуляциях и от количества предыдущих симуляций, выбравших этот ход (чем чаще выбирался раньше этот ход, тем ниже оценка, чтобы программа рассматривала больше разнообразных ходов).

AlphaGo в игре



AlphaGo в игре

Вначале каждой симуляции AlphaGo выбирает ход в уже построенном дереве, с максимальной оценкой. Когда симуляция доходит до позиции, которой нет в дереве, эта позиция добавляется в дерево, вместе со всеми ходами, разрешёнными в этой позиции, которые оцениваются при помощи стратегической сети. Далее, как в методе Монте-Карло, игра симулируется до конца без ветвления. В этой симуляции каждый ход выбирается случайно с вероятностью, полученной при помощи быстрой стратегической сети.

AlphaGo в игре

В конце симуляции, в зависимости от результата, обновляются оценки ходов в построенном дереве. Таким образом каждая симуляция начинается с текущей игровой позиции, доходит до конца, и в результате одной симуляции в текущем дереве раскрывается одна позиция.
Авторы программы обнаружили, что на этом этапе выгоднее использовать не улучшенную стратегическую сеть, а первоначальную (SL policy network). Как считают авторы, это связано с тем, что профессиональные игроки выбирают более разнообразные ходы, чем улучшенная сеть, что позволяет программе рассматривать больше вариантов. Таким образом улучшенная стратегическая сеть не используется во время игры, но её использование существенно для построения оценочной сети, когда программа обучается, играя сама с собой.

Вдохновленные успехами AlphaGo

Facebook также разрабатывает программу для игры в го, Darkforest, которая тоже основана на машинном обучении и поиске в дереве. На начало 2016 года Darkforest показал сильную игру против других компьютеров, но не смог выиграть у профессионала. По силе Darkforest оценивается на уровне программ Crazy Stone и Zen.

1 марта 2016 года разработчики программы Zen, компания DWANGO и исследовательская группа глубинного обучения Токийского университета (которая создала программу Ponanza для игры в сёги, победившую человека) объявили о совместном проекте «Deep Zen Go Project», с целью победить AlphaGo в течение 6—12 месяцев. Японская ассоциация го обещала поддержать проект. В ноябре 2016 года Deep Zen Go проиграла со счётом 2-1 самому титулованному игроку Японии Тё Тикуну.