Australian National University

Action Schema Networks and Monte Carlo Tree Search: The Best of Both Worlds

William Shen

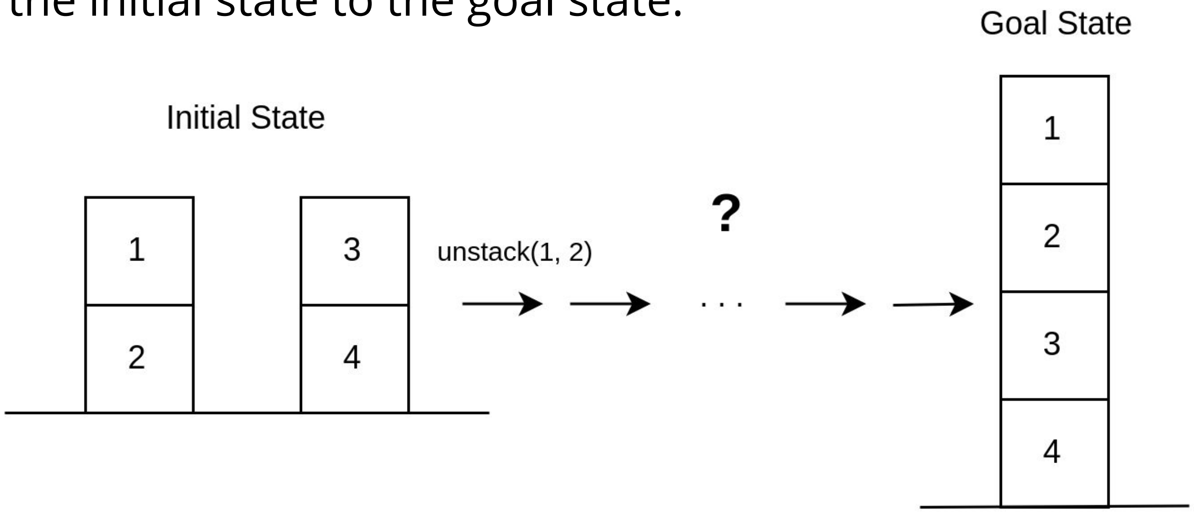
Supervisors: Felipe Trevizan, Sam Toyer, Sylvie Thiébaux, Lexing Xie

COMP3770

October 18th 2018

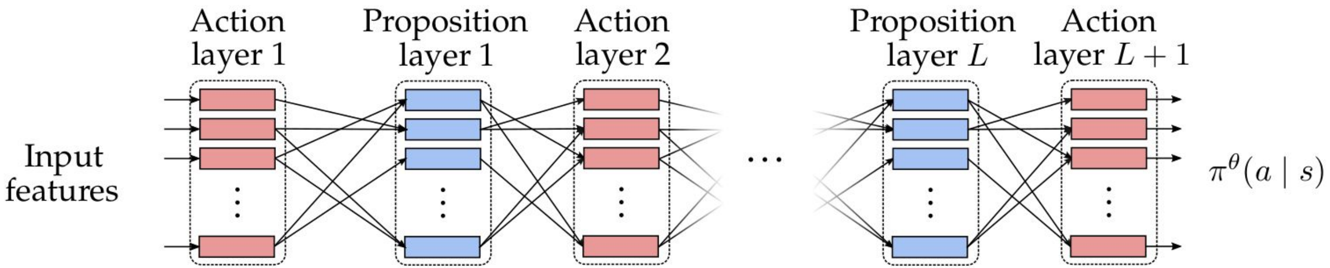
What is planning?

- World has initial state, goal state, and actions we can apply in each state.
- **Goal**: find a policy $\pi : A \times S \rightarrow [0, 1]$, which favours actions that take us from the initial state to the goal state.



Action Schema Networks (ASNets)

- Introduced by Toyer et al., 2018 [1].
- Neural network to learn a generalised policy, by learning local knowledge of the environment.
- **Generalised policy**: a policy that can be applied to similar problems to the ones that the network was trained on.
- Well-suited to problems where there is a ‘trick’ that can be used to avoid traps.
- But what if there is no trick, or the domain changes? Combine ASNets with search!

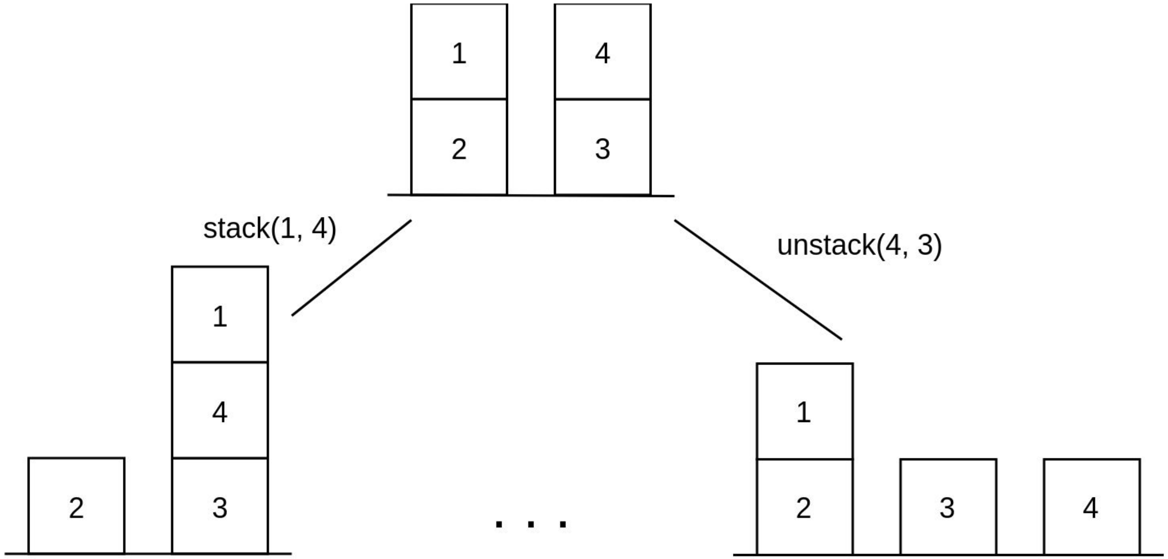


[3]

Australian National University

Monte Carlo Tree Search

- Domain independent heuristic tree search algorithm, based on random sampling. Used in Alpha-Go.



Upper Confidence Bound 1 Applied to Trees

- Need to balance exploration and exploitation when selecting which action to sample.
- Exploration term converges to 0 as $C^k(n_c) \to \infty$

Number of times state has been visited.

Bias term

$$\mathrm{UCB1}(n_d, n_c) = \underbrace{B}_{\text{Bias term}} \cdot \overbrace{\sqrt{\frac{\log C^k(n_d)}{C^k(n_c)}}}^{\text{Exploration}} - \overbrace{Q^k(n_c)}^{\text{Exploitation}}$$

Number of times action has been applied in state

Estimate of cost to reach goal

Combining ASNets and UCT

Combine power of search with local knowledge of the environment.

1. **Learn what we have not learned**
   - ASNet may fail to generalize to problems it has never seen before during training.
2. **Improve suboptimal learning**
   - It can be very difficult to train a neural network
3. **Robust to changes in the environment or domain**
   - The specific problem may change, or probabilities of non-deterministic actions can change.

Using ASNets as a Rollout Policy

- Guide the sampling estimates towards what ASNet believes are 'promising' parts of the search space.
  - Stochastically sample an action from ASNet's policy π
  - Select action with maximum probability in π
- **Good when**: ASNet has learned some knowledge of the environment.
- **Bad when**: ASNet has learned a policy that is completely misleading and thus will misguide the search.

Using ASNets in Action Selection

- Bias what ASNets believes are good actions into UCB1 for exploration.
- Influence of ASNets decays over time.

Policy evaluated on an ASNet

Influence constant

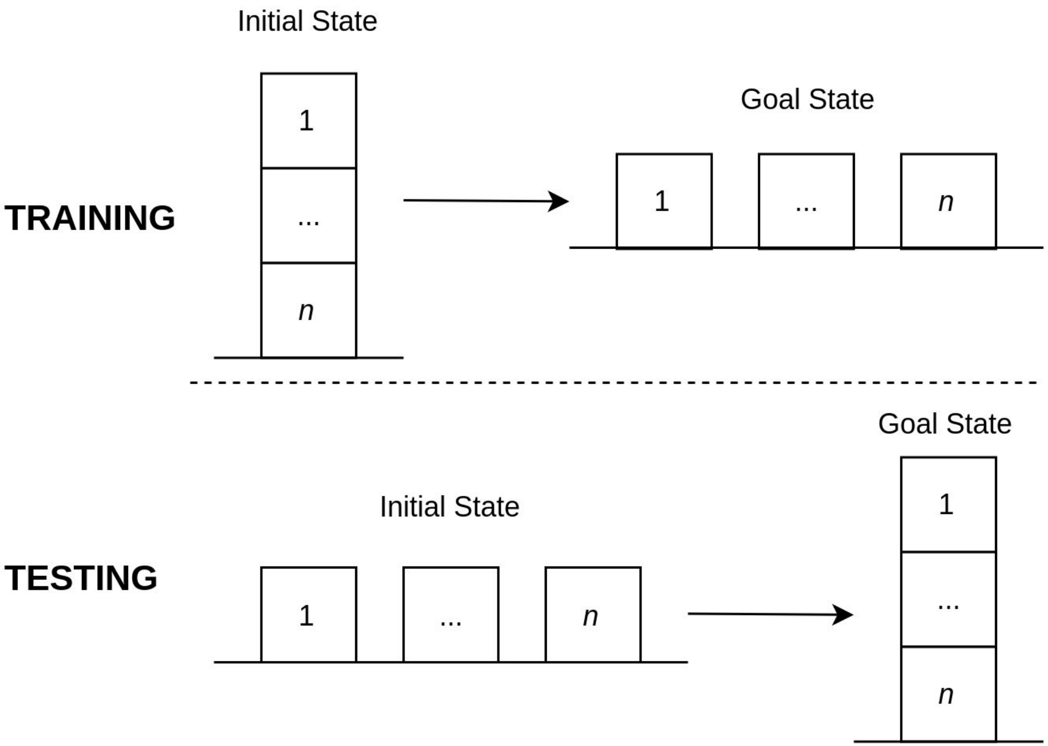
Number of times action has been applied in state

$$\overbrace{\underbrace{\frac{M * \pi(a \mid s)}{C^k(n_c)} + B * \sqrt{\frac{\log C^k(n_d)}{C^k(n_c)}}}_{\text{Exploit ASNets for exploration}}}^{\text{Exploration}} - \overbrace{Q^k(n_c)}^{\text{Exploitation}}$$

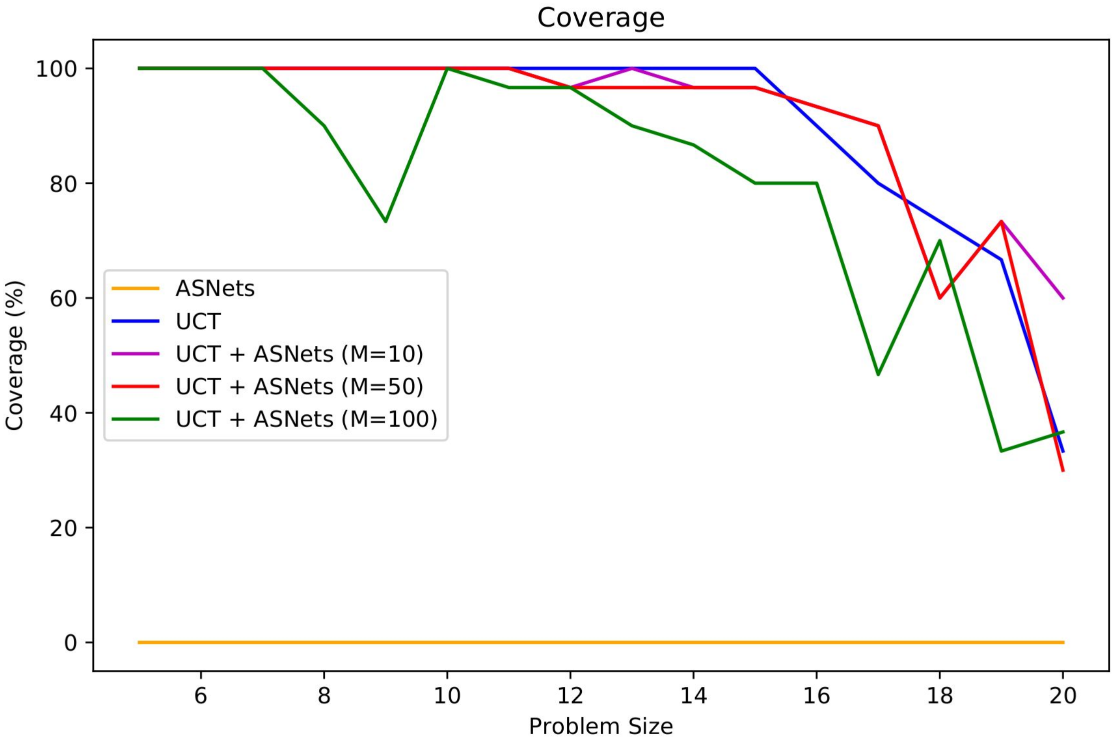
UCB1

Results - Stack Blocksworld

- ASNets trained on unstacking blocks from a single tower.
- Evaluated on stacking blocks into a single tower.
- **Worst-case behaviour**



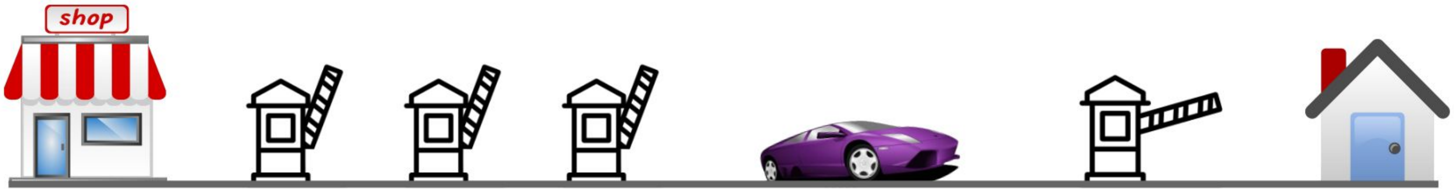
Results - Stack Blocksworld



- ASNets achieves 0 coverage.
- Plain UCT's coverage decays as the problem size increases
- UCT will correct for the bad information an ASNet gives it

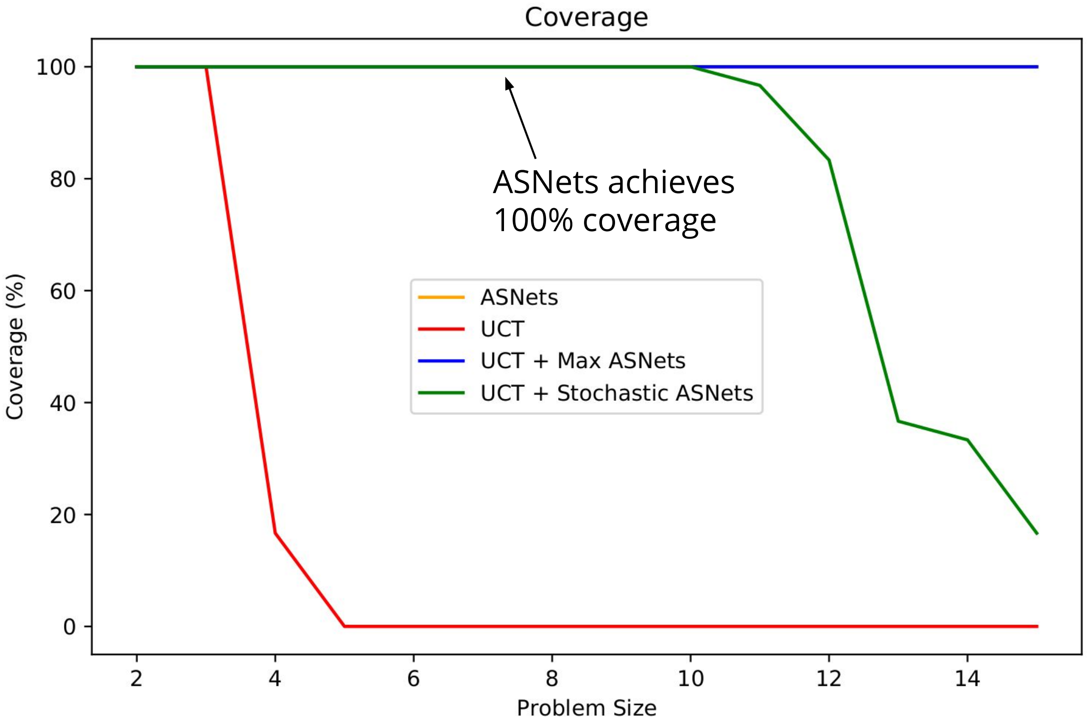
Results - Cosanostra Pizza

- Goal is to deliver pizza to the home, and return to the shop.
- On the road, there are toll-operators. If you do not pay the toll, they might drop the boom gate on your way back.
- Requires long reasoning chains.
- **Very difficult for plain UCT, but very easy for an ASNet.**



[3]

Results - Cosanostra Pizza



- ASNets has learned a ‘trick’ - i.e. pay the toll.
- UCT does not have a very long reasoning chain.
- Max ASNets: gives us a direct path to the goal.
- Stochastic ASNets: probability of path to goal decays as problem size gets larger.

Conclusion and Future Work

- Combining UCT with local knowledge learned by an ASNet can help improve suboptimal learning and be robust to changes in the problems.
- UCT can combat the misinformation given by an ASNet by reducing the network's influence over time.
- Using ASNets as a rollout policy in UCT can vastly improve performance over plain UCT.
- Interleaving planning with execution - train an ASNet whilst using UCT.
- Automatically adjust M in ASNet action selection based on past scenarios.

References

- [1] *Toyer, S.; Trevizan, F.; Thiébaux, S.; and Xie, L.*, 2018. Action Schema Networks: Generalised Policies with Deep Learning. In AAAI Conference on Artificial Intelligence (AAAI).
- [2] *Keller, T. and Helmert, M.*, 2013. Trial-Based Heuristic Tree Search for Finite Horizon MDPs. In ICAPS.
- [3] Some figures taken from Sam Toyer's thesis and presentation for ASNets. https://github.com/qxcv/asnets/blob/master/slides.pdf