

Generalising Discrete Action Spaces with Conditional Action Trees

Christopher Bamford
Game AI Group
Queen Mary University of London
London, UK
c.d.j.bamford@qmul.ac.uk

Alvaro Ovalle
Game AI Group
Queen Mary University of London
London, UK
a.ovalle@qmul.ac.uk

Abstract—There are relatively few conventions followed in reinforcement learning (RL) environments to structure the action spaces. As a consequence the application of RL algorithms to tasks with large action spaces with multiple components require additional effort to adjust to different formats. In this paper we introduce *Conditional Action Trees* with two main objectives: (1) as a method of structuring action spaces in RL to generalise across several action space specifications, and (2) to formalise a process to significantly reduce the action space by decomposing it into multiple sub-spaces, favoring a multi-staged decision making approach. We show several proof-of-concept experiments validating our scheme, ranging from environments with basic discrete action spaces to those with large combinatorial action spaces commonly found in Real Time Strategy (RTS) style games.

Index Terms—action spaces, reinforcement learning, factorised policies, multi-agent, real-time strategy, actor-critic methods.

I. INTRODUCTION

Training reinforcement learning agents to solve environments with large, complex action spaces is a notoriously difficult task [1]. Several methods have been proposed to try to either reduce the space of actions by re-using model outputs for different action types [2], [3], provide side information to facilitate the exploration of large numbers of possible actions [4]–[6], or simplify the manipulation of the action spaces through action embeddings via mechanisms such as attention and graphs networks [7]–[9]. In this paper we propose a *Conditional Action Tree* as a paradigm to generalise several of these methods. *Conditional Action Trees* can be used to describe action spaces in a way that naturally reduces the required policy model output size whilst also allowing action parameterisation and action reduction using invalid action masking. We show how many of the action spaces frequently found in single, multi-agent and Real Time Strategy (RTS) games can be described using *Conditional Action Trees*. We also show that agents that have access to *Conditional Action Trees* as part of their state observations can learn high performing policies. We present several experiments where we purposefully modify the action space of a game environment to include several increasingly more complex features, whilst keeping the observation space and game mechanics consistent. In these experiments we show that agent operating

with *Conditional Action Trees* maintains the performance of those operating with common action space constructions while significantly compressing the number of outputs, or *logits*, required to furnish the policy distribution.

In addition to these experiments we also perform several ablation studies to show various possible modifications to the *Conditional Action Tree* formulation and how they can affect training.

The results suggest that the *Conditional Action Trees* could offer an alternative to generically handle complex combinatorial action spaces with multiple components. As part of this work, *Conditional Action Trees* are made available for all environments in the Griddly Framework [10].

II. BACKGROUND

In a discrete action setting, reinforcement learning (RL) has typically been adapted to environments with simple and small action spaces. Accordingly the implications that the size could have for the agent have been relatively overlooked. Let us start by considering a single actor-critic agent with a small repertoire of actions that consists in motion operations (e.g. up, down, left and right). In this setting, a policy will provide a probability distribution weighting each of the four directions. The agent then can sample this policy to select which action to apply to the environment. However these small manageable action spaces tend to be confined to either simple games or toy environments. The situation changes as we move to tasks requiring combinatorial actions such as in robotics [11], finance [12] and games that involve action spaces with several moving parts and interdependent components. For the latter, RTS games provide instances of actions spaces that can be particularly complicated. To consider a few examples, StarCraft II [3], μ RTS [13] and BotBowl [14] allow control of multiple individual units either by selecting their locations and then issuing commands to those units. Some of the units can perform certain types of actions that are not accessible to other units. Furthermore, some of those actions in turn require additional parameters. For instance, selecting a combat unit that can target several potential locations in the game requires to specify them. Moreover, the particular type of combat actions might be tied or dependent on the unit selected. Several techniques have been proposed to handle this kind of

action spaces. [1] proposes several ways of shaping actions spaces and their relative advantages and disadvantages across several games. For the rest of this section we briefly review two strategies for action space shaping that have been recently proposed in the literature.

A. Parameterised Actions

Parameterised action spaces commonly take the form of an action a made from two components c_0, c_1 where the first component is a *type* of action and the second is a *parameter*. In [2], this action space shaping strategy was applied in the *RoboCup 2D Half-Field-Offense* environment to beat the state-of-the-art hard-coded bots. The first action component defines whether the agent will *dash*, *turn*, *tackle* or *kick*. The second component defines continuous parameters for each of these actions. Four sets of parameters are used, however only one of them is used at each time-step depending on the action type selection. In larger environments such as RTS games, requiring parameters for every action quickly becomes infeasible as the number of action types increases. To contextualise the effect this can have for the size of the policy representation consider the example of *BotBowl*. The game contains 17 action types that require an x and y position parameter. If we proceeded to parameterise the action space, 17 sets of x, y positions would need to be predicted at each time step. From the point of view of an RL agent, the problem is exacerbated if we consider that the policy would have to specify each combination of x and y position. In a traditional *BotBowl* map (25×5) this would lead to a policy that requires to output $17 + 17 \times 25 \times 15 = 6,382$ logits (i.e. unnormalised scores) to parameterise these actions. The number grows exponentially with the map size, a 30×30 map for example, would require $17 + 17 \times 30 \times 30 = 15,317$ logits.

B. Autoregressive Policies

The curse of dimensionality and the combinatorial explosion faced in certain action spaces renders the typical action selection approach highly impractical. An alternative comes from reflecting on the structural relations that exist in complex action spaces. For example, when comparing among all the potential actions that an agent could choose to enact, not all of them will belong to the same level of abstraction. Some actions will be more self-contained, whereas others may need a group of actions to be properly contextualised. Actions can also manifest to an agent as affordances, that is, arising from its coupling with the environment at a particular moment. Moreover, some actions also exert some degree of influence on each other, for instance mutual exclusivity or forming other types of associations.

It is possible to capture some of these notions more concretely by representing a policy in a more expressive manner. In [3] the authors suggest an autoregressive model of the form:

$$\pi(a|s) = \pi(c_0, \dots, c_k|s) = \prod_{k=0}^K \pi(c_k|c_{<k}, s) \quad (1)$$

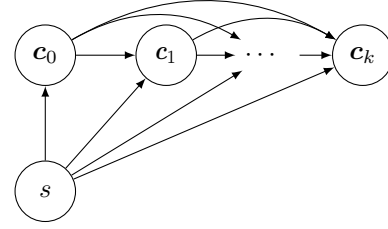


Fig. 1: An autoregressive policy can be graphically represented as a directed acyclic graph where we can illustrate the dependency of a component c_k on the previous components $c_{<k}$.

to decompose the action space into a sequence of sub-spaces. Instead of obtaining a in the full action space, an agent samples multiple sub-actions or components c_k that depend on the previous $c_{<k}$ choices (illustrated in Fig. 1). [3] explores the usage of conditional policies within the context of StarCraft II. However they relax the constraints imposed by the autoregressive model opting for a policy $\pi(a|s) = \prod_k^K \pi(c_k|s)$. In [15], the approach is extended substantially as the architecture considers a conditional policy that captures the context of previous actions through different embeddings. The action subspace decomposition is facilitated by an *invalid action masking* scheme to prevent the agent from selecting actions that are invalid or cannot be performed in the current state, where a function identifier c_0 determines the number of subsequent function arguments $c_1 \dots c_k$. From the point of view of the implementation and the capacity required by a policy, this form of decomposition implies a significant reduction in the number of actions that are effectively considered. Here we also take an approach that places an invalid action masking scheme as a crucial component to tackle the issue of policy decomposition and describe it in more detail in Section III-C.

III. CONDITIONAL ACTION TREES

Conditional action trees (CAT) offer a generalisation of discrete action spaces to provide an interpretation of action selection as the process of traversing along a chained sequence of action components with different levels of dependency. To complete the characterization of a *Conditional Action Tree* we first need to define three main elements: *Action Trees*, *Valid Action Trees*, and finally *Conditional Masking*.

A. Action Trees

We start by formulating a single action as a list of a fixed number of components $a = \{c_0, c_1 \dots c_n\}$, where $c_k \in C_k$. That is, each component takes a value from a set of possible elements. Actions in the same component level are mutually exclusive. For example, *move left* and *move right* must be options within a single component C_k . The possible values of C_k are determined first, by the specification of environment, and second by the values of previous selections $C_k = f(c_0, c_1, \dots, c_{k-1})$.

These restrictions naturally allow the components to form a tree structure, where a path from the root node to any leaf

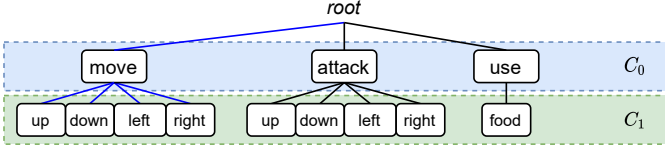


Fig. 2: An action tree consisting of nine possible actions and two components $C_0 = 3$ and $C_1 = 4$. The possible actions are move or attack in any of the four directions, attack in any of the four directions and finally to use a food item. *Move*, *Attack* and *Use* cannot be performed at the same time.

forms the action. An example of an *action tree* is shown in Fig. 2. Note that under this specification an environment requiring the agent to specify a single atomic action at each time step results in an *action tree* with a single component, $a = \{c_0\}$. Parameterised action spaces that contain an action type and a discrete action parameter can also be described by action trees with two components, $a = \{c_0, c_1\}$.

Previous work has touched upon the idea of using trees as a formalization of action spaces with multiple components such as in [16], where the tree structure is referred to as a *Hierarchical Action Space*. Other works have used action spaces that are similar to those used in this paper as examples of action trees. The *Global Action Space* in [17] for example can also be described as an *action tree*.

B. Valid Action Trees

We define a *valid action tree* as a sub-tree of an action tree at a particular environment state, where the nodes of the sub-tree correspond to *possible* actions in that state. For example, consider the tree in Fig. 2, an agent in a state where there are no enemies surrounding it and does not have food in its inventory has a *valid action tree* only consisting of the left-most *move* branch and its children.

In the context of reinforcement learning, a valid action tree is provided by the environment at each time step. Valid action trees are then used to construct the *Invalid Action Masks* which are described next in Section III-C. These masks index the child nodes that are available in the full action tree.

C. Invalid Action Masking

Invalid action masking (IAM) [13] is a technique used to stop agents from sampling actions that are invalid in a particular game state. IAM is useful in environments where the action space is large, and some of the actions are only available in certain states. For example in RTS games [3], [4], [15], the agent’s action may consist of selecting a unit or units from a large list, and then issuing commands to those units. The commands sent to those units can also be unique to particular unit types. This results in a large number of options in the action space that are invalid. In policy gradient and actor critic methods in deep RL, IAM is applied to the logits, $\mathbf{l} \in \mathbb{R}^n$, produced by a neural network by replacing the logits corresponding to invalid actions with large negative numbers.

This forces the probability of selecting those actions to tend towards 0.

For instance, let us assume a compound policy constituted by K independent components, such that $\pi(a|s) = \prod_k^K \pi(c_k|s)$. This type of action policy could be described by:

$$\pi(a|s) = [\pi(c_0|s), \pi(c_1|s), \dots, \pi(c_K|s)] \quad (2)$$

For each of the components in equation 2, a value is selected following a softmax sub-policy. We can create a mask to modify the logits to assign large negative numbers to actions deemed as non-viable or inaccessible. The modified logits result in $\hat{\mathbf{l}} = \mathbf{l} + \mathbf{m}$ where $-\infty < m_i \ll 0$. It then follows that the *masked* logits alter the probability of a value of c_i of being sampled:

$$\pi(c_i|s) = \begin{cases} 0 & \text{if } m_i \rightarrow -\infty \\ \frac{e^{l_i}}{\sum_j^K e^{l_j}} & \text{if } m_i = 0 \end{cases}$$

In PySC2 [3], μ RTS [5] and BotBowl [14] action masks can be constructed from lists of available actions that are provided by the environment implementations, however, these action masks do not take into account that the masking of some sub-actions can depend on the sampled values of others. As an example, in an environment with units that are selected by coordinates and the set of available actions for each unit is disjoint, the mask for the available actions is dependent on the selection of the unit. Masks that are naively constructed using these lists can still lead to select actions that are not available, as the list does not take into account the selection of the unit. [18] introduces a two-step method for generating masks where the unit location is selected using masked logits and then a second mask is generated based on that selection. This significantly improves training as the mask for unit actions is dependent on the selected unit.

D. Conditional Masking

The two-step method of masking in [18] can be generalised to an n-step masking method when the environment provides a valid action tree as described in Section III-B. We refer to this generalization of action selection and masking as a *Conditional Action Tree (CAT)*.

A CAT is constructed by adding a mask at each node of a valid action tree, defining which child nodes of the complete action tree are available. An action is constructed by starting at the root node of the valid action tree and selecting a child node from the masked distribution. This child node contains the mask to use for the next component. Thus, first the mask is obtained as $\mathbf{m}_{k+1} \sim p(\mathbf{m}_{k+1}|c_k)$, to produce a masked sub-policy to sample a component $c_{k+1} \sim p(c_{k+1}|\mathbf{m}_{k+1}, s)$. This process continues until all action components have been sampled. The full compound policy, as illustrated in Fig. 3, is factorised as:

$$\pi(a|s) = p(\mathbf{m}_0) \prod_k^K p(\mathbf{m}_{k+1}|c_k) p(c_{k+1}|\mathbf{m}_{k+1}, s) \quad (3)$$

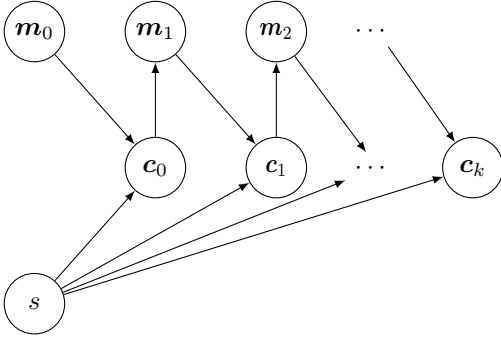


Fig. 3: A graphical model representing the policy as a joint distribution of masks m and components c . In a CAT, a component c_0 is sampled from the options allowed by the mask m_0 . The next mask m_1 depends on c_0 which in turn constrains the next possible component c_1 . The process is repeated until all c_k have been sampled.

IV. ACTOR-CRITIC WITH CONDITIONAL ACTION TREES

A. IMPALA

The description of the action spaces provided by CAT is naturally agnostic to the choice of the RL algorithm. We examine this perspective within the context of IMPALA, an actor-critic based framework introduced in [19]. Unlike A3C [20] or other similar distributed approaches where the agents share their gradients, IMPALA considers the acting and the data collection as independent from the learning step. That is, it separates the learners who are in charge of computing the gradients and sharing the most recent parameters, from the actors whose role is to execute a policy, only sharing back with the learners the observations gathered during an episode.

B. V-trace and masking

As an actor-critic, IMPALA learns $V_\theta(s)$ parameterised by θ to be used as part of the baseline, and a policy π_ϕ parameterised by ϕ . Each actor executes their own policy μ by retrieving the latest policy π from the learner. Meanwhile the learner updates continuously the parameters θ and ϕ . As the process occurs in parallel and in a decoupled manner, there will be a discrepancy between the policy μ from an actor and π . Namely, the trajectories $(s_t, a_t, r_t \dots)$ collected by an actor come from a policy μ that has become obsolete with respect to π . IMPALA proposes to address these off-policy corrections by introducing a v-trace target,

$$v_t = V(s_t) + \sum_{i=t}^{t+n-1} \gamma^{i-t} \left(\prod_{j=t}^{i-1} u_j \right) \delta_i V \quad (4)$$

where $\delta_i V$ corresponds to a temporal difference term,

$$\delta_i V = \rho_i(r_i + \gamma V(s_{i+1}) - V(s_i))$$

the v-trace adjusts the weight of the contributions provided by the actors through the presence of two truncated importance sampling weights $\rho_i = \min(\bar{\rho}, \frac{\pi}{\mu})$ and $u_j = \min(\bar{u}, \frac{\pi}{\mu})$. Thus

the second part of v-trace target acts as a correction term. For example assuming $\bar{\rho}$ and $\bar{u} \geq 1$, if $\mu > \pi$ the learner would downweight the observations and actions followed by the actor. Intuitively, if this ratio tends towards a low number it indicates that the policies have diverged significantly. The extent to which more recent $\delta_i V$ affect the update of a previous V is captured by the product of $u_{t:i-1}$ where \bar{u} serves as a hyperparameter controlling the convergence speed towards V . In turn, ρ determines to which V we converge. A $\bar{\rho}$ close to 0 leads convergence towards a V^μ as the correction term becomes negligible in the v-trace target.

It is important to note that for CAT we do not just consider a single set of importance sampling weights $\{\rho, u\}$ but instead we must account for multiple corrections dependent on the various sub-policies such that $\rho_{k,i} = \min(\bar{\rho}, \frac{\pi(c_k|m_k,s)}{\mu(c_k|m_k,s)})$ and $u_{k,j} = \min(\bar{u}, \frac{\pi(c_k|m_k,s)}{\mu(c_k|m_k,s)})$ for a sub-policy k . Moreover, we must synchronize the masks applied to c_k in both π and μ . Similarly, for updating the policy parameters ϕ we adapt,

$$\rho_{k,i} \nabla_\phi \log \pi_\phi(c_k|m_k,s)(r_t + \gamma v_{t+1} - V_\theta(x_t))$$

to consider the inclusion of the masks and to propagate the gradients to all sub-policies.

V. EXPERIMENT SETTING

A. The "Clusters" Game

We perform our experiments in the *Clusters* environment provided by Griddly [10]. *Clusters*¹ is a game in which coloured *boxes* must be clustered together in specific locations defined by the environment level. The environment contains five levels with a set of movable coloured *boxes* and a single fixed-position *block* of each colour. The agent receives a reward of +1 each time it pushes a coloured *box* against a fixed location *block* of the same colour. When a coloured *box* is pushed against its respective *block*, it becomes a *block* itself. If all *boxes* are converted to *blocks* the episode is completed successfully. Some levels also contain *spikes* which give the player a negative reward (-1) and terminate the episode if the agent or any *boxes* collide with them.

The observation space of the agent consists of a 5×5 grid where the agent itself is situated at the center-bottom of the grid as shown in Fig. 4. Each cell of the 5×5 grid contains 10 binary values describing whether an object is present in each cell. The 10 objects are as follows: three (red, green, blue) coloured *boxes* and three associated *blocks*, *walls*, *spikes*, the agent and finally a *broken box* which only appears in the final state of an episode if a coloured *box* is pushed against *spikes*.

B. Action Space Variations

By default, the agent's movement is restricted to moving forward one position, or rotating ± 90 degrees every step. *Boxes* are "pushed" by the agent when the agent attempts to move into the cell occupied by the *box*.

¹<https://griddly.readthedocs.io/en/latest/games/Clusters/index.html>

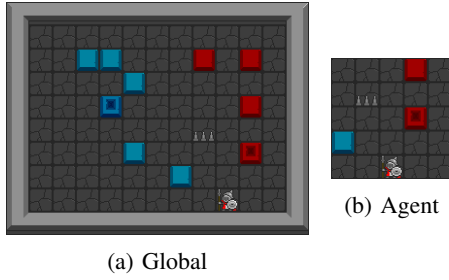


Fig. 4: An example of a level in the Clusters game, showing (a) the entire game and (b) the viewpoint of the agent.

In our experiments, we modify these action spaces to make it increasingly more complex whilst keeping the game mechanics, observation space and reward scheme consistent. This allows us to test the Conditional Action Tree formulation on different action spaces with minimal influencing factors. The only significant change we make to the environment across experiments is when we remove the avatar and allow the agent to move *boxes* independently by selecting their x and y coordinates. These action space variations are explained below:

1) **Move (M)**: The first action tree variation is the default action space provided by the *Clusters* environment. The action space consists of rotate left, right and move forward. As mentioned in Section III-A, this is equivalent to an *Action Tree* with a single component $a = \{c_0\}$, with $c_0 \in \{0, 1, 2, 3\}$.

2) **Move + Push (MP)**: Next we modify the action space to consider that the agent can no longer *push* boxes by simply moving into the location occupied by them. We define a separate *push* action that must be performed in order to move any of the boxes. The *push* action has no effect unless there is a box directly in front of the agent. The *move* action is left unmodified, other than the fact that it can no longer be used to push boxes. As the *move* and *push* actions are mutually exclusive they are confined to the first level in the tree $C_0 = \{0, 1\}$, whilst the second component C_1 contains either the three move parameters or the single *push* parameter.

3) **Move + Push + Separate colours (MPS)**: This action space configuration contains the same modifications as the MP variant, however it splits the *push* component into three to account for the separate colours. The agent must select the correct *push* action, depending on which colour *box* it is pushing (i.e. push green, push blue, push red). Similarly to MP, the action space consists of two components, but the first one now contains the three different push actions instead of one, that is $C_0 = \{0, 1, 2, 3\}$. The second component C_1 remains the same.

4) **Move - Agent (Ma)**: To make the action space significantly larger we remove the agent and the associated ego-centric partial observability. Thus the input consists of the entire 13×10 grid with the same 10 binary digits per cell. The *boxes* are now moved first by selecting their x and y coordinates and then by issuing the direction where to move it. This action space has three components: $C_0 =$

	$ C_0 $	$ C_1 $	$ C_2 $	$ C_3 $	Total Logits
M	3				3
MP	2	3			5
MPS	4	3			7
Ma	13	10	4		27
MSa	13	10	4	3	30
Depth-2					
M	3				3
MP	4				4
MPS	6				6
Ma	130	4			134
MSa	130	12			142

TABLE I: This table shows the number of action components, their sizes in term of number of logits and the total logits needed in the policy output for the action space variations described in Section V-B. We also show the number of logits that are required in the *Depth-2* model.

$\{\text{valid } x \text{ coordinates}\}$, $C_1 = \{\text{valid } y \text{ coordinates}\}$ and $C_2 = \{0, 1, 2, 3\}$ referring to the movement directions *up*, *down* *left* and *right*.

5) **Move + Seperate colours - Agent (MSa)**: The final and largest action space we consider starts with the same formulation as *Ma*, but separates the colour components in the same way as done in *MPS*. This results in an action space with four components: x , y , action type and action parameters. An example of a conditional action tree for this space is shown in Fig. 5

C. Baselines

For each of the variations of the action space described in the previous section, we compare against two baselines which are designed to show the benefits and limitations of the CAT paradigm. The baselines modify only the way that the model interacts with the action space in terms of number of logits required. The number of actions and mechanics of the game are consistent.

1) **No Masking**: For the first comparison we use the same action components as a CAT but remove the Invalid Action Masking entirely. This means that the component selections are made independently of each other and invalid actions can be selected.

2) **Depth-2**: The second comparison also uses a conditional action tree structure, but flattens the action tree to only a depth of two. The separate x and y components (only available in *MSa* and *Ma*) are flattened into a single xy component. Additionally the action type and action parameter components are flattened into a single selection. This flattening process was also considered in [1] where multi-discrete actions are flattened into single discrete spaces. Table I shows the number of logits per-component for all experiments and the equivalent number of logits required in the depth-2 representation.

D. Masking Ablation

To show that structure of the tree and the resulting conditional masking has an effect on the learning of the policy, we perform an experiment where we relax the conditional

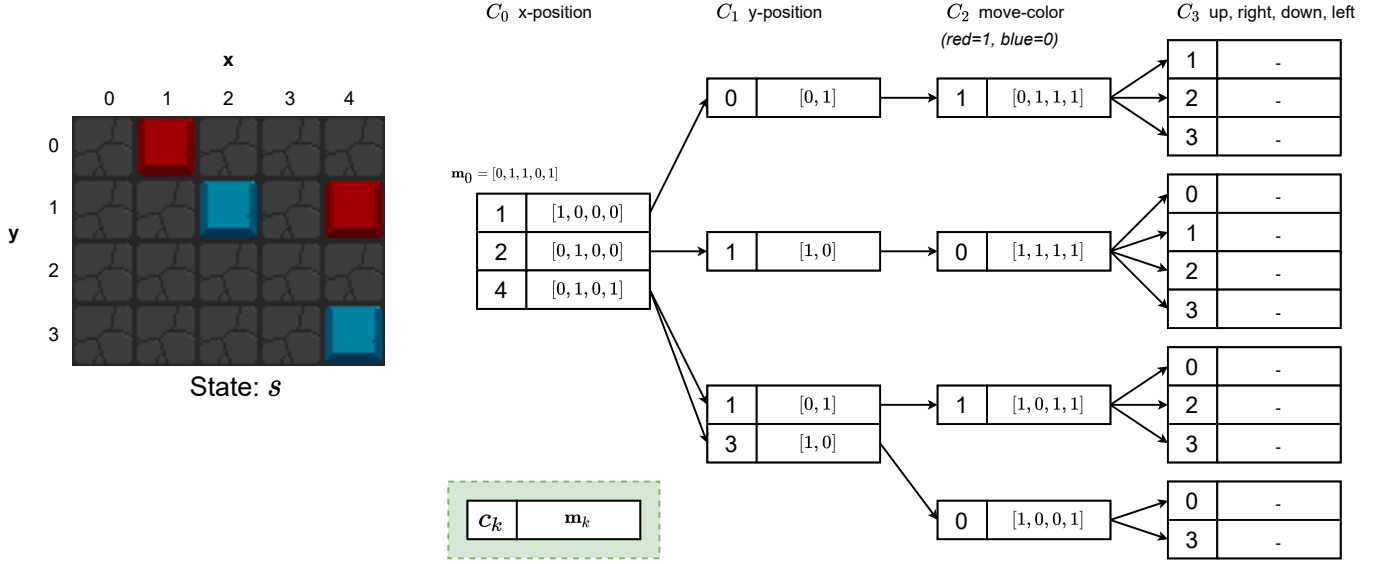


Fig. 5: Image of a conditional action tree from a 5x4 *Clusters* level configured in with the MSa action space as described in Section V-B. The agent is configured with an action space with 4 components, the agent selects which object to move by its position and the colour. It then proceeds to choose which direction to move the *box*. The CAT shown contains the selected action component c_k and the mask m_k for each possible valid combination of components.

masking restrictions and compare it against the fully conditional masking. We relax the conditional masking of the tree by collapsing the masking across the tree breadth-wise, so all masking is effectively a union of all the possible masks at each depth. This method is equivalent to applying a single mask to the entire action space with no consideration for dependencies between action selections. We refer to the relaxed *Collapsed* and full *Conditional* masking options in further sections as CAT_CL and CAT_CD respectively.

E. Model Architecture

We keep the model architecture consistent throughout all experiments as much as possible. The size of the model input observations differs between partially observable agent-based environments (**M**, **MP**, **MPS**) and unit-selecting environments (**Ma**, **MSa**). The partially observable environments have a $5 \times 5 \times 10$ observation space, while for the unit-selecting environments it is $13 \times 10 \times 10$. Additionally, the final layer in each experiment outputs the number of logits shown in Table I. The model itself contains two convolutional layers with padding 1 and kernel size 3 that up-scales the number of values in each channel to 32 and then 64 respectively, whilst keeping the width and height the same. After these layers, the output tensor is flattened and then passed through two linear layers with 1024 and 512 neurons. We then use a separate actor and critic head. The actor head contains a further two linear layers, first to compress to 256 nodes and then a final layer to output predicted logits. The critic head contains a single layer which outputs the single predicted value.

VI. RESULTS

In total we run 4 experiments on each variation of the action space of the *Clusters* game. The four experiments contain the two baselines as previously described, and two versions of masking (CAT_CL and CAT_CD).

The first variation **M** provides evidence that the formulation of conditional action trees generalise to simple action spaces. In this environment, all variations of the action space are almost identical and therefore have similar performance. Masks in this environment have little effect because only a few actions are ever invalid. **MP** and **MPS** variations begin to show that the fully conditional tree CAT_CD and the depth-2 action tree policies learn faster and plateau at high-scoring policies. Depth-2 action policies in these variations are in fact slightly better performing than the more hierarchical formulation of the Conditional Action Tree, in addition of using one less logit in their policies. The reason for this is that in the **MP** and **MPS** the structure of the associated action tree has a degree of 1 in all of the *push* nodes, making the tree structure redundant for the push actions. In cases like these, where parent nodes have only single children, it is more efficient to flatten these nodes into a single set of children.

Conditional Action Trees excel in the variations with the highest branching factors. **Ma** and **MSa** both require the policy to select an individual unit to perform an action at each time step. As expected, the depth-2 policy and CAT_CD have similar performance as they are both CATs, but CAT_CD splits the x , y selection into separate components, which results in a greater than 4x reduction in the number of logits required by the policy, with no loss in performance.

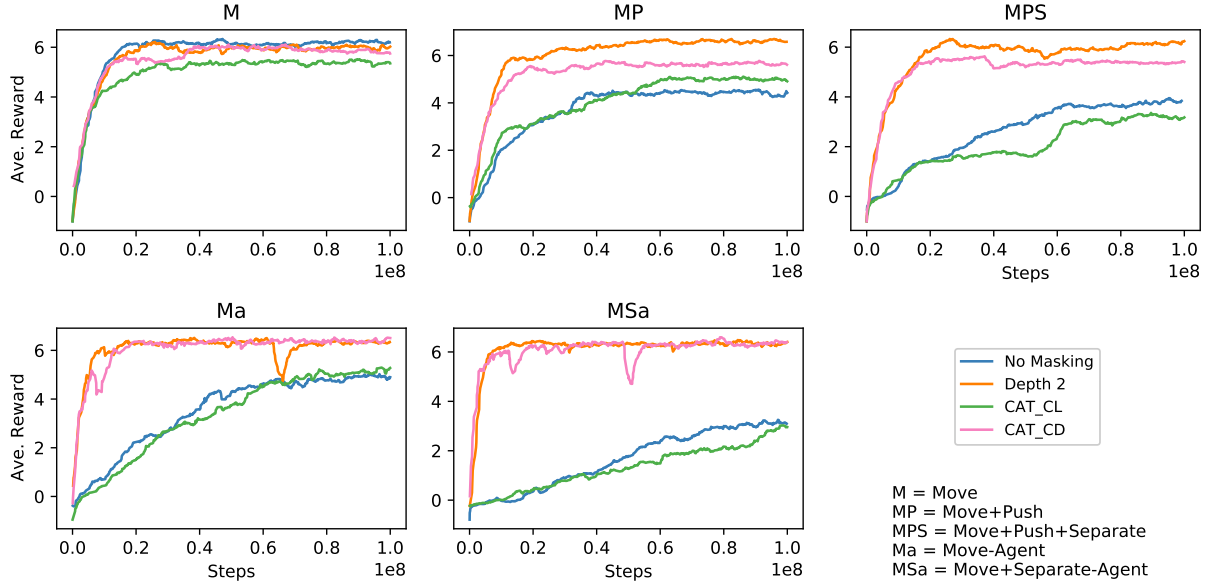


Fig. 6: The average episode reward during training of the 5 different action space variations as described in Section V-B. For each of the 5 action space variations, we compare three policies with the same action tree structure, but different masking methods: No Masking, CAT_CD (conditional) and CAT_CL (collapsed). We also provide a comparison with a model policy that uses an action tree limited to depth 2 as described in Section V-C2

The results for experiments on all five test environments with *Collapsed* (CAT_CL) masks are also shown in Fig. 6. We can see that with *Conditional Action Trees* the full *Conditional* (CAT_CD) masking strategy is important for efficient training, as the *Collapsed* masking strategy performs similarly to the *No Masking* Baseline.

VII. DISCUSSION

Trees are a useful data structure across many fields of computer science, and can provide a natural representation for action spaces. Although the formulation and the experimental setting focused on discrete action spaces, we hypothesize that in principle the formulation can be extended to continuous spaces by implementing a similar parameterisation to structure the action components associated to specific densities. To the best of our knowledge this direction has not been explored.

It is important to note that the degree of subtrees in a CAT should be taken into consideration when deciding on parts of the tree that could be flattened, as this can lead to unnecessary increase in policy size.

The current work presented the CAT formulation in five toy scenarios intended to recreate, with different levels of complexity, the conditions frequently exhibited in various single, multi-agent and RTS games. Further work will be required to analyse the behavior of the CAT in more complex domains such as μ RTS or BotBowl. Part of the current limitation resides in adapting these and other environments to provide *Valid Action Trees*. With a *Conditional Action Tree* the parameterised part of BotBowl’s action space could be reduced from 6392 logits to $25 + 15 + 17 = 57$

Other relevant research on how to handle large action spaces has applied techniques such as evolutionary algorithms [6], [21]. These proposals have also been tested in scenarios that require multiple actions per time-step. A naive approach to work with CATs within this context would be to recursively append the tree to its own leaf nodes, resampling until a condition specifying the required number of actions is fulfilled.

In its current form a CAT makes specific assumptions about the conditional dependencies between actions (Section III-D). Following [15], a potential future research avenue is to explore the possibility of modelling more complex dependencies. Namely, by contextualizing further the selection of a c_k with with an encoding learned from previous components $c_{<k}$.

A. Entropy reduction

The process by which an agent acquires a policy in reinforcement learning is essentially an exercise in reducing its behavioral uncertainty. Although it remains important that the agent maintains a level of flexibility that can support an adequate generalization [22], [23]. It must effectively have to be able to discriminate the actions that are beneficial from those that are not in a particular state. We can interpret this procedure of action discrimination as a process of entropy reduction, as an initial high entropy policy is transformed by redistributing the probability mass or density to weight those actions that have been identified as more favorable. For large action spaces it is evident that this process becomes more complex as the behavioral possibilities explode. The view posited by CAT is that it is possible to exploit the structure of the action space to facilitate the acquisition of

behaviorally relevant policies as we can state that for two arbitrary segments of the action space, C_i and C_j , the mutual information between them is $I(C_i; C_j) \geq 0$. This implies that there can be information to be gained by using existing relationships between actions. To express it differently, conditioning guarantees us that $H(C_i|C_j) \leq H(C_i)$ with equality iff $p(c_i, c_j) = p(c_i)p(c_j)$. Thus constructing an action tree becomes a tool that contributes to the process of entropy reduction at the level of the policy as it decomposes what is potentially a large flat action space into multiple smaller sub-spaces.

VIII. CONCLUSION

In this paper we have proposed a formalisation of a tree structure for representing discrete action spaces with any number of components. We have provided the required steps to adapt already existing action spaces to conform to a *Conditional Action Tree*. From a technical perspective, a side effect of imposing a structure to the action space is the reduction of the elements considered by a policy. The experiments showed that this modification does not reduce the sample efficiency during training and achieves comparable performance while resulting in significantly smaller models with less parameters.

As part of this work, Griddly [10] implements built-in functionality for generating *Valid Action Trees* and we provide all reproducible examples in a github repository². We also provide all training parameters, statistic and videos using Weights and Biases³.

We encourage the developers of reinforcement learning environments, especially those with large discrete action spaces to provide *Valid Action Tree* observations in their environments.

ACKNOWLEDGMENT

We would like to thank Paulo Rauber, Simon Lucas and Shengyi Huang for their valuable feedback and to the ITS Research team at Queen Mary University of London for providing technical support necessary to run the simulations. This research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT. <http://doi.org/10.5281/zenodo.438045>.

REFERENCES

- [1] A. Kanervisto, C. Scheller, and V. Hautamäki, "Action space shaping in deep reinforcement learning," *arXiv*, apr 2020.
- [2] W. Masson, P. Ranchod, and G. Konidaris, "Reinforcement learning with parameterized actions," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI'16. AAAI Press, 2016, p. 1934–1940.
- [3] O. Vinyals, T. Ewalds, S. Bartunov, P. Georgiev, A. S. Vezhnevets, M. Yeo, A. Makhzani, H. Küttler, J. Agapiou, J. Schrittwieser, J. Quan, S. Gaffney, S. Petersen, K. Simonyan, T. Schaul, H. van Hasselt, D. Silver, T. Lillicrap, K. Calderone, P. Keet, A. Brunasso, D. Lawrence, A. Ekermo, J. Repp, and R. Tsing, "StarCraft II: A new challenge for reinforcement learning," *arXiv*, aug 2017.
- [4] M. Samvelyan, T. Rashid, C. S. de Witt, G. Farquhar, N. Nardelli, T. G. J. Rudner, C.-M. Hung, P. H. S. Torr, J. Foerster, and S. Whiteson, "The StarCraft multi-agent challenge," *arXiv*, feb 2019.
- [5] S. Ontañón, "The combinatorial multi-armed bandit problem and its application to real-time strategy games," in *Proceedings of the Ninth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, ser. AIIDE'13. AAAI Press, 2013, p. 58–64.
- [6] N. Justesen, T. Mahlmann, S. Risi, and J. Togelius, "Playing multi-action adversarial games: online evolutionary planning versus tree search," *IEEE Transactions on Games*, vol. 10, no. 3, pp. 281–291, sep 2018.
- [7] C. Berner, G. Brockman, B. Chan, V. Cheung, P. Debiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, R. Józefowicz, S. Gray, C. Olsson, J. Pachocki, M. Petrov, H. P. d. O. Pinto, J. Raiman, T. Salimans, J. Schlatter, J. Schneider, S. Sidor, I. Sutskever, J. Tang, F. Wolski, and S. Zhang, "Dota 2 with large scale deep reinforcement learning," 2019.
- [8] Q. Ma, S. Ge, D. He, D. Thaker, and I. Drori, "Combinatorial optimization by graph pointer networks and hierarchical reinforcement learning," *arXiv*, nov 2019.
- [9] P. Ammanabrolu and M. Hausknecht, "Graph constrained reinforcement learning for natural language action spaces," jan 2020.
- [10] C. Bamford, S. Huang, and S. Lucas, "Griddly: A platform for ai research in games," 2020.
- [11] D. Korenkevych, A. R. Mahmood, G. Vasan, and J. Bergstra, "Autoregressive policies for continuous control deep reinforcement learning," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, S. Kraus, Ed. California: International Joint Conferences on Artificial Intelligence Organization, aug 2019, pp. 2754–2762.
- [12] H. Yang, X.-Y. Liu, S. Zhong, and A. Walid, "Deep reinforcement learning for automated stock trading: an ensemble strategy," 2020.
- [13] S. Huang and S. Ontañón, "A closer look at invalid action masking in policy gradient algorithms," *arXiv*, jun 2020.
- [14] N. Justesen, L. M. Uth, C. Jakobsen, P. D. Moore, J. Togelius, and S. Risi, "Blood bowl: A new board game challenge and competition for AI," in *2019 IEEE Conference on Games (CoG)*. IEEE, aug 2019, pp. 1–8.
- [15] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. P. Agapiou, M. Jaderberg, A. S. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. L. Paine, C. Gulcehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, and D. Silver, "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, oct 2019.
- [16] Z. Fan, R. Su, W. Zhang, and Y. Yu, "Hybrid actor-critic reinforcement learning in parameterized action space," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, S. Kraus, Ed. California: International Joint Conferences on Artificial Intelligence Organization, aug 2019, pp. 2279–2285.
- [17] S. Huang and S. Ontañón, "Comparing observation and action representations for deep reinforcement learning in μ rts," *arXiv*, oct 2019.
- [18] S. Huang and S. Ontanon, "Measuring generalization of deep reinforcement learning applied to real-time strategy games,"
- [19] L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning, S. Legg, and K. Kavukcuoglu, "IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures," *arXiv*, feb 2018.
- [20] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *International conference on machine learning*. PMLR, 2016, pp. 1928–1937.
- [21] H. Baier and P. I. Cowling, "Evolutionary MCTS for multi-action adversarial games," in *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, aug 2018, pp. 1–8.
- [22] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning," in *Aaai*, vol. 8. Chicago, IL, USA, 2008, pp. 1433–1438.
- [23] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1352–1361.

²<https://github.com/Bam4d/conditional-action-trees>

³https://wandb.ai/chrisbam4d/conditional_action_trees