How to Pick Strawberries Safely: Objective Reward Shaping with Visual Complexity

Rory Buckley
University College Dublin
Dublin, Ireland
rory.buckley@ucdconnect.ie

Gregory O'Hare
University College Dublin
Dublin, Ireland
gregory.ohare@ucd.ie

Rem Collier
University College Dublin
Dublin, Ireland
rem.collier@ucd.ie

ABSTRACT

Reward shaping is an important means of reducing the computational costs associated with deep reinforcement learning in environments with sparse rewards. However, serious concerns exist over the subjectivity and self-centred motivations introduced with artificial reward design. This paper presents an objective approach to reward shaping. Reward is generated from the maximisation of visual complexity in the environment. The current approach outlined in this paper relies on a measure of structural complexity. Reward generated from the maximisation of structural complexity demonstrates significant performance increases against the sparse reward of survival in a base-building game testbed. Despite these promising early results, the use of only structural complexity in the measure limits generalisation. Future research will seek to incorporate behavioural complexity for a more complete measure.

KEYWORDS

Reinforcement Learning, Reward Shaping, Image Reconstruction

1 INTRODUCTION

Deep Reinforcement Learning (Deep RL) has seen an explosion in interest over the last several years. Beginning with early work on Atari games[22], followed by Google DeepMind's mastery of Go[28], the state-of-the-art has now achieved super-human performance in complex games such as Starcraft 2[34] and Dota 2[2].

Following this string of achievements, Deep RL has been heralded as sufficient for Artificial General Intelligence (AGI). DeepMind, the lab behind much of these ground-breaking achievements, have recently claimed "Reward is Enough"[29]. They suggest the maximisation of reward with Deep RL is all that is required to achieve general human-like intelligence.

Criticism from the natural science community has already been leveled at DeepMind. Criticism is particularly focused on the suggestion that maximisation of an arbitrary reward function is sufficient for human-like intelligence. In public interviews, neuroscientist Churchland[7] criticises DeepMind for ignoring the social and moral aspects of human intelligence. From her work on the biological underpinnings of our moral intuitions, Churchland suggests a reward function for human-like AGI needs to have other social and moral considerations outside of individual self-interested maximisation of reward. We are not merely machiavellian maximisers, motivated by pure self-interest, it is suggested. We do not aim to maximise our own individual reward at all costs.

The fear of such a machiavellian maximiser precedes this criticism. The nightmare of an AI interested only in the selfish maximisation of its own reward function was epitomised in the thought-experiment of the Strawberry Picker[8]. Proposed by Elon Musk, the Strawberry Picker seeks only to maximise a single reward function chosen by its self-interested creators. It levels whole cities to create more farmland, all in the singular pursuit of increasing strawberry production. Such a machiavellian maximiser successfully maximises its chosen reward function, but does so with such single-minded self-centred purpose that it respects nothing of the existing environment around it. Reward functions chosen by self-interested creators may be "enough" to learn general intelligence, but not enough to learn to respect anything else. Humanity, our economy, and the wider ecology may suffer devastation as a result.

Arbitrary reward functions are not enough. Particularly those subjectively shaped by human desires. A long history of problems is associated with the introduction of artificial rewards designed by human hand. A move to rewards that can be considered objective is imperative. Particularly in the face of a machiavellian maximiser emerging from an AGI with misguided reward design.

Unfortunately, existing rewards that can be objectively sourced in an RL environment are typically sparse. Their sparsity can worsen the already poor learning efficiency of RL, and necessitate prohibitively expensive computing costs. Escalating computing costs puts an objective reward shaping out of reach. A means of generating a frequent objective form of reward is required to speed up learning so as to avoid an inevitable cost-cutting return to subjective design in the future.

This paper presents the results obtained with a promising objective reward shaping candidate, the maximisation of visual environmental complexity. Early results demonstrate it offers significant performance increases over sparse rewards in certain scenarios. However, the current measure of visual complexity relies solely on structural complexity, limiting it to domains such as base-building games where it correlates with success. Future research will seek a more complete measure incorporating behavioural complexity.

In addition to the promise of faster RL, it is hoped visual complexity will also provide a path to solving the problems associated with subjective reward design and self-centred motivations. Incorporating a motivation that inherently respects the existing structural and behavioural complexity of the environment is a promising direction towards safe RL. It promises to motivate only further complexification of the agent's environment, avoiding losses to complexity in economic assets and the surrounding ecology. It opens up a novel approach for future research that seeks means of avoiding dangers to the economy, human society, and life itself, posed by a machiavellian maximiser with self-centred motivations.

Proc. of the Adaptive and Learning Agents Workshop (ALA 2022), Cruz, Hayes, Silva, Santos (eds.), May 9-10, 2022, Online, https://ala2022.github.io/. 2022.

2 BACKGROUND

Reward shaping is not a new idea. But introducing artificial rewards into an RL scenario typically involved a myriad of human design choices. Historically, the subjectivity of human hand-crafting introduced perverse incentives, producing suboptimal and absurd results.

But reward signals considered objective are difficult to find in many scenarios. Existing rewards in an RL environment can be infrequently encountered. Such sparse rewards are a serious problem for viability as Deep RL enters into ever more complex environments with longer episodes of learning. Indeed, in many environments the sole reward encountered by an agent exists only upon successful completion of the episode. Such a survival signal is a maximally sparse reward. Given longer episodes of learning, large computing costs are then required to learn. OpenAI has recently modeled this trend of exploding computing requirements, and has predicted it will double every 3.4 months[26]. This is in stark contrast to the doubling of computing power every 18 months with Moore's law, a law which is predicted to end soon as well[31].

2.1 Subjective Reward Shaping

Reward shaping is the inclusion of artificial rewards in RL distinct from those found in the Markov-Decision-Process (MDP) underlying the environment. Reward shaping was conceived of as a means of introducing expert knowledge into RL. RL is known to be data sample inefficient, requiring many agent-environment interactions to learn. Expertise used to craft rewards could speed up learning. Subjective choices were then required to modify the original reward structure of the MDP.

Such subjective reward shaping was known to lead to sub-optimal behaviour[27] and frequently produced absurd results. It can introduce perverse incentives that lead away from the goal state, and has no guarantees of policy invariance. It fails to guarantee policy learnt with the modified reward structure will be optimal policy in the original unmodified MDP.

Potential-Based Reward Shaping (PBRS)[5] is a proposed solution for the sub-optimal behaviour learned with subjective reward shaping. Subjective design choices are limited to the design of potential functions that describe an agent's potential to reach the goal state. Policy invariance is theoretically guaranteed when the state described as having maximum reward by the potential function coincides with the goal state of the original MDP[23].

Critically, PBRS requires that expert knowledge of goal states is sufficient. Similarly, choices are left up to subjective design, such as the choice of heuristic describing distance to the goal. Complex MDPs may lack sufficient expert knowledge for the design of heuristics that can describe distance to goal states. In addition, simple distance-to-goal heuristics are known to lead to sub-optimal behaviour[33].

Subjective reward shaping increasingly appears an ill-fit for the ever more complex MDPs underlying state-of-the-art testbeds and real-world applications.

2.2 Objective Reward Shaping

Reward signals sourced from objective information alone would avoid all the issues associated with subjectivity in reward shaping. Several main approaches exist that produce reward signals without subjective design choices.

Learning-to-Learn methods would offer the ability to learn frequent reward signals from the existing sparse rewards of an RL environment. Evolutionary[24] methods and Meta-learning[36] have both been proposed as means of learning frequent reward signals to aid in RL. Dynamic PBRS[6] is the most promising, and learns potential functions with policy invariance guarantees. However, due to the overhead of running a secondary learning method on top of existing RL methods, it is uncertain whether these approaches can consistently offer the necessary speed-ups. This is particularly of concern in complex MDPs possessing only a survival signal.

Objective means of transferring expert knowledge have been proposed. RL from demonstration[3], RL from advice[11], and Preference-based RL[35] are examples of such attempts. Advice appears the most promising and offers policy invariance guarantees. Unfortunately, many complex real-world environments are open-ended and lack experts with sufficient objective knowledge to guide optimally without bias. Advice from subjective opinion may bias the agent in unknown ways in complex environments with sparse rewards. A sparsity or total absence of objective signals may leave other considerations without representation and lead unpredictably to a Strawberry Picker scenario.

Intrinsic motivation[1] provides rewards from objective modeling efforts. Curiosity[4] is one form, and provides a reward from prediction of future states. Reward results from surprise at state unpredictability and unfamiliarity. Curiosity is then only useful for exploring unknown states.

Empowerment is another form of intrinsic motivation. The empowerment method DIAYN[9] is the culmination of recent information-theoretic work on how the action-space of an RL agent can itself be used to generate rewards. DIAYN motivates the agent to maximise the diversity of empowering "skills" it can find in its action-space. Consequently, it is inevitably centred around the agent's own action space. This constraint precludes consideration of the diversity and complexity of structures and behaviour in the environment outside of the agent. The complexity of vital interests, such as economic capital and other agents, is unable to be captured by this approach. DIAYN appears unable to mitigate the risk of a machiavellian maximiser concerned only for its own interests. Indeed, an empowerment approach that fundamentally relies on the action-space alone constrains an agent to learning from information that is irrevocably self-centred.

The candidates above provide objective forms of reward useful for other purposes, but leave us without what is desired, reward that is frequent and fast, but also free from self-centred concerns and subjective bias. A novel inductive bias for RL appears to be required that would both speed up learning and eliminate subjectivity and self-centred motivations.

3 VISUAL ENVIRONMENTAL REWARDS

An inductive bias that is objectively sourced would best come from environmental information itself. Environmental information can provide both objective information on success in the environment, as well as provide frequent rewards, solving the issue of sparsity found with existing environmental rewards. Visuals appear to be the most reliable source of information on the environment. Rewards generated from visual information would also add no further cost to the existing fixed costs of image convolution in state-of-the-art end-to-end neural network approaches to Deep RL.

Characteristics of the environment need to be found that provide an inductive bias for learning optimal behaviour. The base-building RTS genre is explored, and taken as an exemplar of complex RL environments due to state-of-the-art use.

3.1 Thermodynamic Characteristics

Base-building RTS games focus on the need to acquire resources, invest them wisely into assets, and use strategy to protect against asset loss from adversity. Real-world scenarios also require living systems to learn similar characteristic behaviours. Living systems must also accumulate resources, invest them into assets, and protect against asset loss. These characteristic behaviours of living systems exist due to thermodynamic constraints. It appears thermodynamic constraints are implicitly shared between base-building games and reality itself. Natural scientific principles describing the means living systems thrive under thermodynamic constraints may provide insight into an inductive bias for RL environments implicitly constrained in this way.

In systems ecology the Maximum Power Principle[25] states that the most successful living system in an environment will maximize power, the dissipation of thermal energy over time. Thermodynamically this dissipation of thermal energy is associated with an increase in entropy. This increase in entropy through maximal power output satisfies the second law of thermodynamics and allows life to survive within thermodynamic constraints. The characteristic behaviours of a living system appear to be a sign it is maximizing power and increasing entropy. Working backwards, the increasing entropy of visual environmental information may provide a sign of optimal behaviour in an environment with thermodynamic constraints. One potential inductive bias for such environments may then involve measuring visual environmental entropy.

Evolutionary biology offers another alternative. McShea[20] describes a tendency for living systems to increase in complexity over time. As increases in complexity track with the successful persistence of living systems, it suggests increasing complexity will also track with the characteristic behaviours required to thrive under thermodynamic constraints. This then provides another potential inductive bias: measuring increases in visual environmental complexity.

A measure of visual entropy is trivial. The entropy of image pixel data can be used. A simple count of unique values suffices for RTS base-building games with simpler graphics. A measure of the diversity of pixel values would be needed for more advanced graphics. A measure of visual complexity, though, proves more ambiguous. Much ambiguity and debate already exists around how to create a measure of complexity for real-world living systems. A far more complicated measure than simply pixel value diversity would be needed to measure visual environmental complexity in an RTS base-building game.

3.2 Visual Complexity

Standard measures of complexity in information science prove inappropriate for use in measuring the complexity of real-world living systems. Measures of algorithmic complexity like kolmogorov complexity[17] produce absurd results when applied to real-world scenarios. Kolmogorov complexity is derived from the size of the smallest program required to reconstruct a system's output information. It relies on the incompressibility of this information. When applied to real-world systems it produces counter-intuitive results. Such a measure would rank the visual environmental information of a living system lower in complexity than random noise, due to the compressibility of regularities found in living structures.

Complex systems science offers a solution with the concept of effective complexity[10]. Gell-Mann's work sought to solve the counter-intuitive results found with kolmogorov complexity. The focus was put on respecting the regularities found in real-world systems that reduce measurements of their kolmogorov complexity when compared to random noise. Such regularities can be used to compress the program of reconstruction further than is possible for the incompressible irregularities of noise. To solve this, effective complexity sought to only measure the information content of regularities in a system, avoiding the information of irregularities.

Unfortunately, effective complexity proves ambiguous in its implementation. Criticism suggests the choice of implementation is debatable and open-ended[18]. A proxy measure of effective complexity is needed in lieu of a definitive measure. The means of generating the smallest program of reconstruction for a system's visual information, respecting only its regularities, is left open to interpretation.

Inspiration for the means of creating the program of reconstruction can be taken from scientific insights into the structure of real-world systems, as well as from practical work on deep learning with images. Evolutionary biologist McShea[19] notes that the increase in biological complexity over time involves an increase in *nestedness*, or hierarchical structure. Economist Simon[30] also notes that the complexity in real-world systems tends to be hierarchical. Their perspectives provide insight into the best means of capturing the feature regularities required for effective complexity. A means that captures hierarchies of feature regularities appears to be most suited for a measure of effective complexity in environments containing real-world systems.

Similarly, deep learning on images typically makes use of a hierarchical architecture. A convolutional neural network learns to produce a hierarchy of feature maps to best represent the construction of complex structures in the visual data.

Both the scientific insights of evolutionary biology and the best-practices of image reconstruction techniques point to the same general pattern of practically measuring complexity in real-world visual environmental information. Complex systems in reality tend to be hierarchically constructed, and a convolutional neural network would best capture the hierarchical structure of complex systems in realistic environments.

3.3 Visual Complexity Measure

A convolutional auto-encoder[13] was chosen to best measure the effective complexity of visual environmental information. This architecture learns to compress visual data through a bottleneck layer, and then reconstruct the image from only the compressed information out of the bottleneck. It trains on minimising image reconstruction inaccuracy. The bottleneck layer makes use of the information bottleneck principle[32] to learn a maximally compressed program of matrix multiplication that describes the construction of the system found in the visual data. Finding a maximally compressed program describing the construction of the visual data is what is required for both kolmogorov complexity and effective complexity.

This approach then is the best means of approximating the lower bound of the kolmogorov complexity of visual environmental information. It can also then be thought of as the best approach in measuring the effective complexity of the visual data. An auto-encoder has a finite bottleneck of only a certain number of latent dimensions. Random noise cannot be effectively reconstructed without maximal reconstruction inaccuracy. The information required to reconstruct random noise cannot pass through a bottleneck layer smaller in size than the visual data itself. By taking into account the reconstruction accuracy of the data, random incompressible noise can be excluded. A convolutional auto-encoder approach can then be assured to be measuring only regularities in the data, as effective complexity demands, and to not generate a high score for incompressible random noise, like kolmogorov complexity would provide. An auto-encoder approach then provides a practical means of measuring effective complexity in visual environmental information.

A measure of visual complexity that captures the hierarchical nature of effective complexity in natural systems was designed with a convolutional auto-encoder. The state-of-the-art variational auto-encoder (VAE)[14] was the specific implementation chosen. A Residual convolutional neural network (ResNet)[12] approach was chosen for the convolutional portion of the VAE. A ResNet can scale to 150+ layers. Otherwise convolutional networks are typically limited to only a few layers. A ResNet-VAE with surplus layers was desired to best capture the full hierarchy of regularities likely to exist in the visual information of structures in the RL environments. It was hoped a ResNet in the VAE would capture as many levels of feature regularities in the hierarchy as possible.

A hyper-parameter search was conducted and 256 latent dimensions in the VAE bottleneck layer were found to be the best balance between reconstruction accuracy and compression. The ResNet-VAE was trained for 2000 epochs, training on 320 images per epoch, and validating on 32 per epoch.

The feature maps produced by the convolutional filters of each layer in the encoder are considered to be describing the minimal information required for reconstruction of the visual data under maximal compression. The total entropy of these feature maps was then considered to be giving an approximate measure of the kolmogorov complexity of the visual data. It was further considered to be an approximate measure of the effective complexity of non-random regularities found in the visual data.

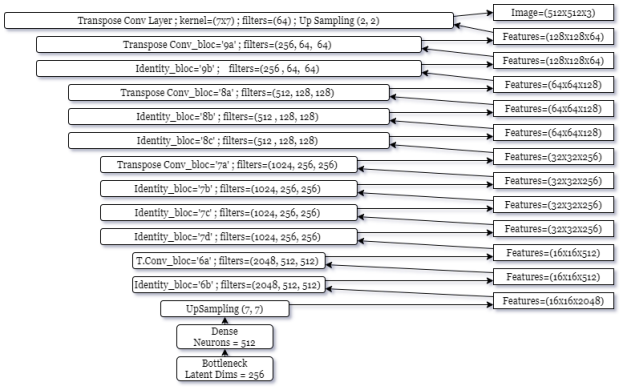

Figure 1: ResNet-VAE Decoder

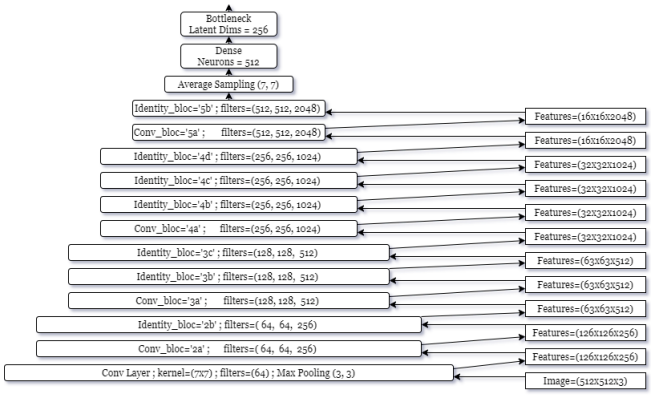

Figure 2: ResNet-VAE Decoder

Visual complexity was then taken as the effective complexity of the image normalised between the minimum value encountered so far and the max possible value.

3.4 Visual Complexity Equation

$$\sum_{i=1}^{n}(\sum_{j=1}^{m}(\sum_{k=1}^{l}H(i,j,k))) = r_{effective\_complexity}$$

$$\frac{(r_{effective\_complexity} - r_{min\_encountered})}{(r_{max\_possible} - r_{min\_encountered})} = r_{visual\_complexity}$$

where n is the number of $512 * 512$ sub-images in an image, m is the number of convolutional layers in the ResNet-VAE encoder, l is the number of filters in each layer, $H()$ is the entropy of a feature map, $r_{min\_encountered}$ is the minimum effective complexity encountered, and $r_{max\_possible}$ is the theoretical maximum effective complexity possible.

3.5 Visual Entropy

Measuring visual entropy by comparison was a far simpler calculation. Visual entropy was simply calculated as the number of unique values in an image.

4 TESTBED

The Tiberian Sun Grid World testbed has been created from Command and Conquer Tiberian Sun graphical assets and game values. Tiberian Sun is a base-building Real-Time-Strategy game similar to the state-of-the-art testbed StarCraft 2, but is less visually complicated. It shares with it similar real-world thermodynamic constraints. The testbed simulates resource acquisition, investment of resources into assets that present as visual information, and a visualised loss of these assets through destruction by an enemy adversary.

The testbed features a Battle Simulator and Image Generator. The Battle Simulator uses game values from an open-source version of Tiberian Sun running on the OpenRA engine. The game state after each battle is visually represented by the Image Generator. After each successful battle the Image Generator visualises advancement across the terrain, acquisition of a terrain grid, and the construction of assets on this grid. Assets previously constructed on a grid are visually destroyed after a failed battle, simulating destruction by the adversary. Together the Battle Simulator and Image Generator form the Tiberian Sun Grid World testbed.

4.1 Image Generator

The Tiberian Sun Grid World Image Generator uses graphical assets sourced from an online fan site for the game. These graphical assets exist as simple sprites and represent all units and structures in the game from all possible angles. The Image Generator overlays them on background terrain to generate images.

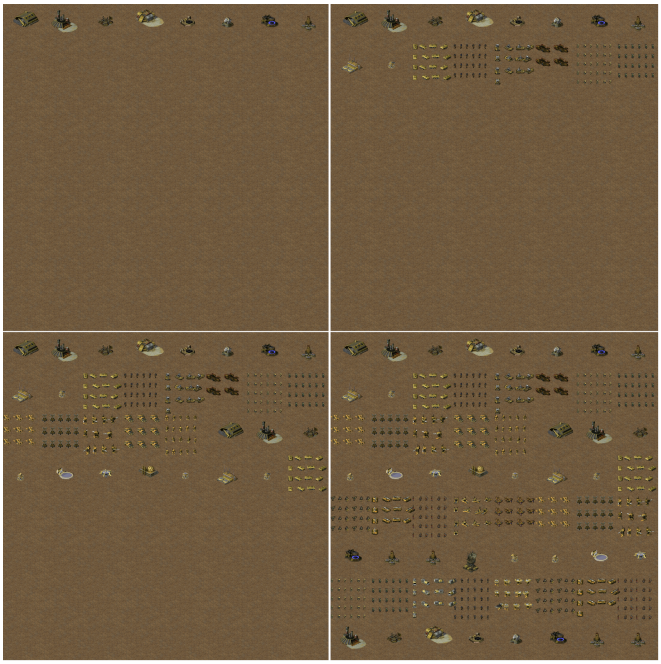

Figure 3: Advancement across terrain in Tiberian Sun Grid World

The Image Generator places the units and structures in a predefined expansion across the terrain of the grid world. Starting at the top left corner there is an expansion of structures and units left to right, and then top to bottom, across the terrain grid. The Tiberian Sun Grid World testbed uses this in conjunction with the Battle Simulator to visualise the advancement of an RL agent's structures and units across the grid world's terrain.

The Image Generator visually represents the acquisition of increasing numbers of contiguous terrain grids from successive victories in the Battle Simulator, and the conversion of terrain resources into either units or structures. This visualisation of an increasingly large and diverse area of units and structures across a previously undeveloped terrain creates a visual complexification of the economy, potentially discernible by measures of visual entropy or visual complexity applied to each successive image.

Conversely, the Image Generator visualises the loss of terrain corresponding to defeats in the Battle Simulator through the destruction of the units and structures previously built upon the foremost terrain grid. These contiguously lost terrain grids return to empty undeveloped terrain after defeat in the Battle Simulator, with the destruction of their contents. The loss of size and diversity of units and structures in the previously growing area of contiguous terrain grids is hoped to present as a loss of complexity of the RL agent's assets and economy in the environment. This loss is similarly hoped to be discernible to a measure of visual entropy or visual complexity applied to successive images showing the worsening game state.

4.2 Battle Simulator

The Tiberian Sun Grid World has at its core a Battle Simulator that simulates battles between a chosen friendly unit and a random enemy unit. The values for cost, hit-points, armour type, weapon damage and weapon reload delay were sourced from the open-source version of Tiberian Sun running on the OpenRA engine.

The Battle Simulator determines if the friendly unit chosen by the RL agent to face off against the visible enemy unit will beat it in terms of cost-effectiveness, cost per unit time taken to destroy it. The RL agent will advance across the terrain and acquire the next grid of terrain resource if it is victorious. It will then automatically use this terrain resource to construct units and structures on it, complexifying its in-game economy. Conversely, failure to deploy a friendly unit with superior cost-effectiveness to the enemy will result in loss of a terrain grid, destruction of its constructed contents, and a loss of complexity for the RL agent's in-game economy.

4.3 Battle Simulator Equation

$$a_{cost\_eff} = b_c/((b_h/(a_{wd} * a_{vba})) * a_{wrd})$$
$$b_{cost\_eff} = a_c/((a_h/(b_{wd} * b_{vba})) * b_{wrd})$$
$$if : a_{cost\_eff} > b_{cost\_eff} = a_{victory}$$
$$elif : b_{cost\_eff} > a_{cost\_eff} = b_{victory}$$

where a_{cost_eff} is friendly cost-effectiveness, b_c is enemy unit cost, b_h is enemy unit health, a_{wd} is the damage the friendly unit's weapon inflicts, a_{vba} is the modifier for the friendly unit's weapon against the enemy unit's armour, and a_{wrd} is the friendly unit's weapon reload delay.

5 EVALUATION

Reward signals generated from measures of visual complexity and visual entropy were evaluated in the Tiberian Sun Grid World testbed. A second evaluation in a modified testbed was also performed. The Battle Simulator underlying the Tiberian Sun Grid World was replaced by the classic benchmark Cartpole-v0. A second MDP was sought due to concerns that Tiberian Sun unit values, and their resulting battle successes, may be overly correlated with the visual complexity of the sprites representing them.

5.1 Tiberian Sun Grid World

Reward signals Visual Complexity and Visual Entropy were trialed in the Tiberian Sun Grid World battle simulator. They were compared against Survival Signal and a Perfect Potential Function.

Survival Signal solely relies on a reward signal that only exists at the end of an episode of learning. It exists to demonstrate the issues of maximally sparse rewards, like the signals of victory or survival, that are only encountered at the ends of episodes of learning.

Perfect Potential Function is a reward signal that exists consistently after each battle and gives a perfect reward of +1 when a battle is won, and -1 when a battle is lost. It received the Survival Signal reward at the end also.

Visual Complexity and Visual Entropy both had to rely on purely visual rewards from an image generated after each battle, until they arrived at the end and received the numeric Survival Signal reward also. They were trialed in learning episodes of different lengths. Differing numbers of battles were required to be won in an episode before the Survival Signal reward was received. Episodes containing 10 battles, 20 battles and 40 battles were trialed.

The RL Agent used with each reward signal was the Advantage Actor Critic (A2C), a single threaded version of the Asynchronous Advantage Actor Critic (A3C)[21]. An epsilon-greedy approach to exploration was used during training, with random moves decaying with each training episode. The epsilon-decay curve governing this exploration behaviour is found in the following graphs.

Each reward signal was trialed in a training session with a max number of environmental actions. Failure to reach the end after a certain number of steps caused the environment to reset. Testing took place in a separate environment after a certain amount of training time. The proportion of battles won in each test episode was recorded, and a moving average of the last 15 such test episodes was generated per training session. 250 training sessions were repeated for each reward signal. The following graphs represent the mean-average of the test data from 250 training sessions.

5.1.1 Visual Complexity in Tiberian Sun Grid World. We see Visual Complexity tracks consistently with the Perfect Potential Function. Similar to the Perfect Potential Function, it displays the desired characteristic of faster learning with increasing episode length compared with the use of Survival Signal alone. This demonstrates that visual information of environmental state can be used to generate a reward signal that increasingly outperforms Survival Signal under increasing episode length.

5.1.2 Visual Entropy in Tiberian Sun Grid World. Visual Entropy as a reward signal underperforms against Visual Complexity in the Tiberian Sun Grid World. This is particularly apparent at 40 battles per episode, where it is noticeably slower at learning than Visual Complexity. In the longest episode of learning it is clearly intermediate in learning speed between Survival Signal and Visual Complexity, suggesting Visual Complexity is the superior one in longer episodes of learning.

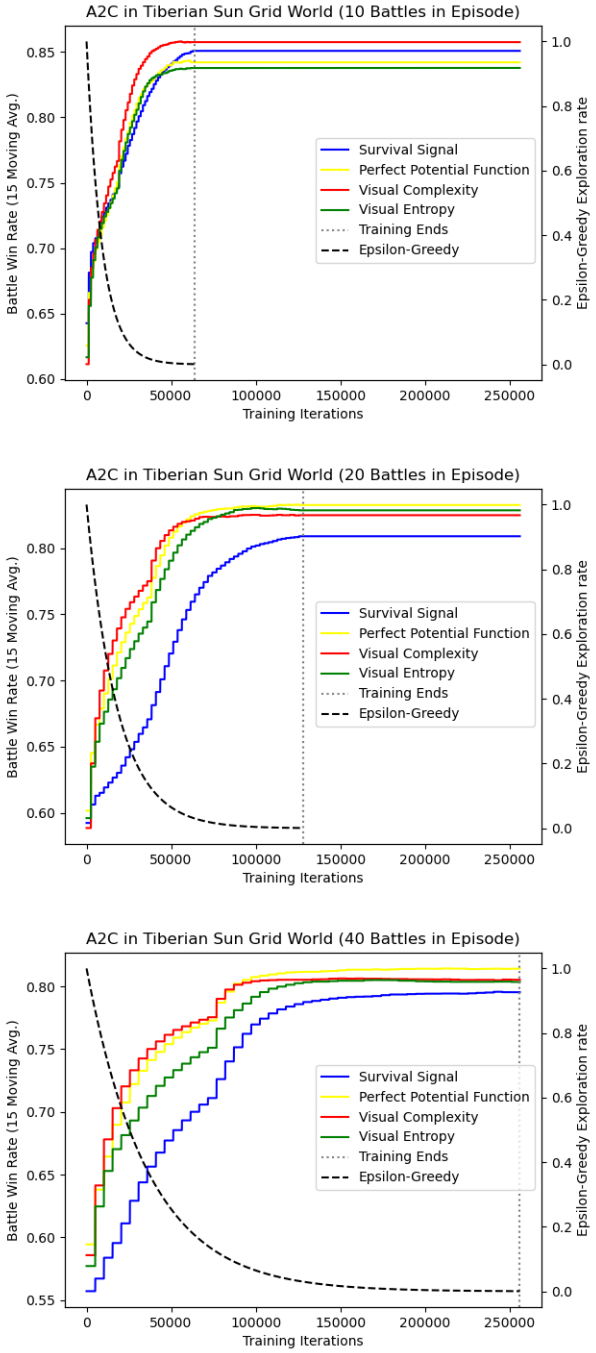

Figure 4: 10, 20 and 40 battles per episode in Tiberian Sun Grid World

5.2 Cartpole-v0 with Sparse Rewards

Cartpole-v0 is a classic RL benchmark where an RL agent must keep a pole on a cart upright for as long as possible, receiving a numeric +1 reward signal for every step upright. Cartpole's numeric rewards were mapped to the Tiberian Sun Grid World. Continued success in Cartpole had its numeric reward taken and visually converted to an advancement across the terrain of the Tiberian Sun Grid World, and the conversion of terrain resource into Tiberian Sun units and structures. This provided visual information of success in Cartpole for a visual measure such as Visual Entropy or Visual Complexity to detect and convert to reward.

Cartpole was run under similar parameters with the same reward signals of Survival Signal, Perfect Potential Function, Visual Complexity and Visual Entropy. Cumulative reward from an episode was held until the end for Survival Signal. A mean-average of 250 training sessions was graphed on the following graphs.

5.2.1 Visual Complexity in Cartpole-v0 with Sparse Rewards. A comparison between Survival Signal and Visual Complexity shows an even more pronounced performance difference when the underlying MDP is from Cartpole-v0. The difference from 50 to 200-step learning episodes is noticeable. Visual Complexity tracks closely with Perfect Potential Function, both achieving a similar lead in learning speed over Survival Signal alone. The underlying MDP appears to make no difference. Increases in the visual complexity of environmental information provides a similar reward to an ideal potential function across both testbeds.

5.2.2 Visual Entropy in Cartpole-v0 with Sparse Rewards. Visual Entropy clearly outperformed Survival Signal. It demonstrates improvements in learning speed over sparse rewards as episode length increases.

In terms of how Visual Entropy compares against Visual Complexity, with Cartpole we see a paradoxical contradictory result. We see Visual Entropy clearly outperform Visual Complexity. This contrasts with its underperformance against Visual Complexity in the Tiberian Sun Grid World. This performance gulf increases as episode length increases.

5.3 Analysis

In terms of reliability Visual Complexity appears to be most robust, with low variance that consistently tracks with the perfect potential function in both RL environments. Visual Entropy on the other hand appears very temperamental, varying widely depending on the underlying MDP of the RL environment.

A clue for explaining Visual Entropy's out-performance of Visual Complexity in Cartpole is found in how it also over-performed against an ideal potential function. Success against even the Perfect Potential Function suggests that it was an improvement in exploration that was helping it learn faster. More successful exploration may be being helped by noise inherent in the Visual Entropy reward signal. It appears to only help in simple MDPs like Cartpole, but fails to do so in more complex MDPs like the Tiberian Sun Grid World Battle Simulator. Visual Complexity then appears superior if the exploration advantage of Visual Entropy in simple MDPs is not desired. However, further testing in other testbeds is required for a conclusive comparison against Visual Entropy.

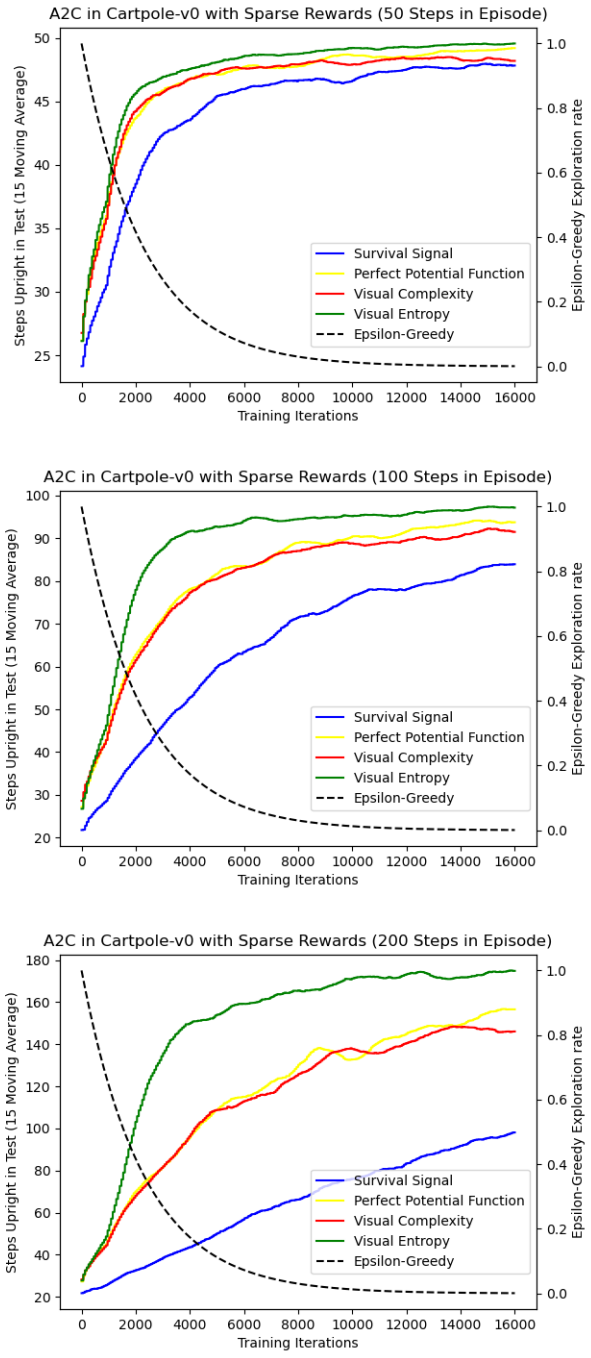

Figure 5: 50, 100 and 200 steps per episode in Cartpole-v0 with Sparse Rewards

6 DISCUSSION

6.1 Related uses of complexity

Complexity as a heuristic has already found use in evolutionary computation in the form of *novelty search*[15]. Through biasing solutions towards complexity, it demonstrated faster learning of high-fitness agent behaviour. The vast numbers of simplistic solutions in the search space tended to collapse into a smaller number of low complexity behaviours in the behaviour space, that were then avoided[16]. Unfortunately novelty search relied on subjectively chosen *behavioural characteristics* to measure complexity.

DIAYN went on to solve this issue for RL by providing an information theoretic basis for the selection of complex behaviour. Through the maximisation of *empowerment*, a diverse set of the most useful *skills* is learned for agent behaviour.

DIAYN, though, is constrained to learning complex behaviour from only the action-space. It lacks the ability for a more holistic assessment of complexity from the environment. Behavioural activity outside of the action-space of the agent is not considered by agent empowerment. In situations where an agent's action-space merely represents signals that trigger behaviour in the wider environment, DIAYN would not take into account the behavioural activity of economic capital, or other agents or processes, responding to these signals. DIAYN is also unable to incorporate information on structural complexity in the environment. Conversely, visual information holds promise for objectively measuring both structural and behavioural complexity in an environment wider than the agent.

Tentatively, the results achieved so far demonstrate the value of maximising the visual complexity of static structure in certain environments. Currently though, this is useful only in thermodynamically constrained environments, such as RTS base-building games; Environments where resource gathering and material asset accumulation is required. In these environments, material asset accumulation presents as increases in the structural complexity of visual environmental information. Maximisation of Visual Complexity then provides a useful form of reward. But outside of these environments mere maximisation of structural complexity would not correlate with success. Other than RTS base-building games and city-building games, there exist few RL environments where increases in structural complexity across space correlate with reward. Most RL environments tend to focus on simple motion, such as robotics or the classic cartpole.

The measure of Visual Complexity proposed so far is currently not a competitive alternative to DIAYN for speeding up learning in general RL environments. At the moment it lacks an ability to measure the complexity of behavioural activity across the temporal dimension. Without this ability it is not applicable to classic RL benchmarks. Although, given the initial success of this approach with structural complexity across space, it is hoped to further develop it to take into account behavioural complexity across time.

With the addition of behavioural complexity, Visual Complexity holds promise for applicability to RL environments involving robotics or general motion. Future work will focus on developing Visual Complexity into a competitive alternative for general RL environments. If successful, Visual Complexity holds promise to provide a richer and more holistic heuristic for RL, with objective rewards generated from visual environmental information alone.

7 CONCLUSION

The maximisation of visual complexity has been proposed as an objectively shaped reward for RL. The use of visual information in the measure offers a reward that is frequently available to an agent, promising to offset learning delays associated with existing objective rewards that are encountered sparsely in the environment. Visual complexity also promises to be objectively sourced from visuals alone, avoiding the problems associated with subjective decisions involved in manual design of rewards.

A measure of visual complexity has been outlined in this paper that relies on the structural complexity of visual environmental information. The maximisation of visual structural complexity has demonstrated significant performance gains over existing sparse environmental rewards like survival in the chosen testbed scenarios. Furthermore, its performance tracks well with a potential function describing optimal asset accumulation in the economy of the base-building themed testbed. However, the sole use of structural complexity in the measure currently limits its use to domains where thermodynamic constraints necessitate diverse asset accumulation.

Given this limitation, the current implementation of the visual complexity measure renders it only appropriate for use in domains where thermodynamic constraints predominate. A measure that includes the complexity of behavioural activity would be required for generalisation to classic RL domains based on motion and robotics.

Future work hopes to take into account visual behavioural complexity. Visual observation promises to offer a measure of an agent's behavioural complexity that avoids the self-centred perspective of individual agent empowerment. Sourcing all necessary information for a complexity measure from the observation-space, and not the agent's own action-space, would allow a more holistic measure of complexity that takes into account the whole environment. If successful, simple visual observations would provide a measure encompassing the behavioural complexity of other agents, and all other economic and ecological processes within the environment. In tandem with the existing consideration of the structural complexity of economic assets, visual complexity could then offer a safer alternative to agent empowerment measures.

Unlike a self-centred machiavellian maximiser, visual complexity could offer a means of fundamentally considering and respecting the complexity of all other structure and behaviour in the surrounding environment. An agent equipped with visual complexity as a reward would be motivated to only further increase the complexity of its environment. Ultimately, an AGI equipped with this holistic perspective could help us avoid the horrors of a machiavellian maximiser. Elon Musk could then sleep soundly, with the nightmare of the Strawberry Picker itself put to bed.

ACKNOWLEDGMENTS

This publication has emanated from research supported in part by a grant from Science Foundation Ireland under Grant number 16/SPP/3296. For the purpose of Open Access, the author has applied a CC-BY-NC public copyright licence to any Author Accepted Manuscript version arising from this submission. This work is co-funded by Origin Enterprises Plc.

REFERENCES

[1] Arthur Aubret, Laetitia Matignon, and Salima Hassas. 2019. A survey on intrinsic motivation in reinforcement learning. *arXiv preprint arXiv:1908.06976* (2019).

[2] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. 2019. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680* (2019).

[3] Tim Brys, Anna Harutyunyan, Halit Bener Suay, Sonia Chernova, Matthew E Taylor, and Ann Nowé. 2015. Reinforcement learning from demonstration through shaping. In *Twenty-fourth international joint conference on artificial intelligence*.

[4] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. 2018. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355* (2018).

[5] Sam Michael Devlin. 2013. *Potential-based reward shaping for knowledge-based, multi-agent reinforcement learning.* Ph.D. Dissertation. University of York.

[6] Sam Michael Devlin and Daniel Kudenko. 2012. Dynamic potential-based reward shaping. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*. IFAAMAS, 433–440.

[7] Ben Dickson. [n.d.]. DeepMind scientists: Reinforcement learning is enough for general AI. https://bdtechtalks.com/2021/06/07/deepmind-artificial-intelligence-reward-maximization/

[8] Maureen Dowd. 2017. Elon Musk's billion-dollar crusade to stop the AI apocalypse. https://www.vanityfair.com/news/2017/03/elon-musk-billion-dollar-crusade-to-stop-ai-space-x

[9] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. 2018. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070* (2018).

[10] Murray Gell-Mann and Seth Lloyd. 1996. Information measures, effective complexity, and total information. *Complexity* 2, 1 (1996), 44–52.

[11] Anna Harutyunyan, Sam Devlin, Peter Vrancx, and Ann Nowé. 2015. Expressing arbitrary reward functions as potential-based advice. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[13] Geoffrey E Hinton and Ruslan R Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *science* 313, 5786 (2006), 504–507.

[14] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[15] Joel Lehman and Kenneth O. Stanley. 2011. Abandoning Objectives: Evolution Through the Search for Novelty Alone. *Evolutionary Computation* 19, 2 (2011), 189–223. https://doi.org/10.1162/EVCO_a_00025

[16] Joel Lehman, Kenneth O Stanley, et al. 2008. Exploiting open-endedness to solve problems through the search for novelty.. In *ALIFE*. Citeseer, 329–336.

[17] Ming Li, Paul Vitányi, et al. 1997. *An introduction to Kolmogorov complexity and its applications*. Springer.

[18] James W McAllister. 2003. Effective complexity as a measure of information content. *Philosophy of Science* 70, 2 (2003), 302–307.

[19] Daniel W McShea. 2001. The hierarchical structure of organisms: a scale and documentation of a trend in the maximum. *Paleobiology* 27, 2 (2001), 405–423.

[20] Daniel W.. McShea and Robert N Brandon. 2010. *Biology's first law: The tendency for diversity and complexity to increase in evolutionary systems*. University of Chicago Press.

[21] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous Methods for Deep Reinforcement Learning. *CoRR* abs/1602.01783 (2016). arXiv:1602.01783

[22] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).

[23] Andrew Y Ng, Daishi Harada, and Stuart Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, Vol. 99. 278–287.

[24] Scott Niekum, Lee Spector, and Andrew Barto. 2011. Evolution of reward functions for reinforcement learning. In *Proceedings of the 13th annual conference companion on Genetic and evolutionary computation*. 177–178.

[25] Howard T Odum. 2007. *Environment, power, and society for the twenty-first century: the hierarchy of energy*. Columbia University Press.

[26] AI Open. 2018. AI and Compute. https://openai.com/blog/ai-and-compute/

[27] Jette Randløv and Preben Alstrøm. 1998. Learning to Drive a Bicycle Using Reinforcement Learning and Shaping.. In *ICML*, Vol. 98. 463–471.

[28] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. *nature* 550, 7676 (2017), 354–359.

[29] David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. 2021. Reward is enough. *Artificial Intelligence* (2021), 103535.

[30] Herbert A. Simon. 1962. The Architecture of Complexity. *Proceedings of the American Philosophical Society* 106, 6 (1962), 467–482.

[31] Tom Simonite. 2016. Moore's Law Is Dead. Now What? https://www.technologyreview.com/s/601441/moores-law-is-dead-now-what/

[32] Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*. IEEE, 1–5.

[33] Alexander Trott, Stephan Zheng, Caiming Xiong, and Richard Socher. 2019. Keeping your distance: Solving sparse reward tasks using self-balancing shaped rewards. In *Advances in Neural Information Processing Systems*. 10376–10386.

[34] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

[35] Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. 2017. A survey of preference-based reinforcement learning methods. *The Journal of Machine Learning Research* 18, 1 (2017), 4945–4990.

[36] Haosheng Zou, Tongzheng Ren, Dong Yan, Hang Su, and Jun Zhu. 2019. Reward shaping via meta-learning. *arXiv preprint arXiv:1901.09330* (2019).