The Intelligent Robot Learning Laboratory

WISEST
women in scholarship, engineering, science & technology

Syncrude Canada

AUGUST 2021

AGENT TAMER
THE SECRET LIFE OF ALGORITHMS

IRENE OLAYINKA
CALARINA MUSLIMANI
DR. MATTHEW TAYLOR

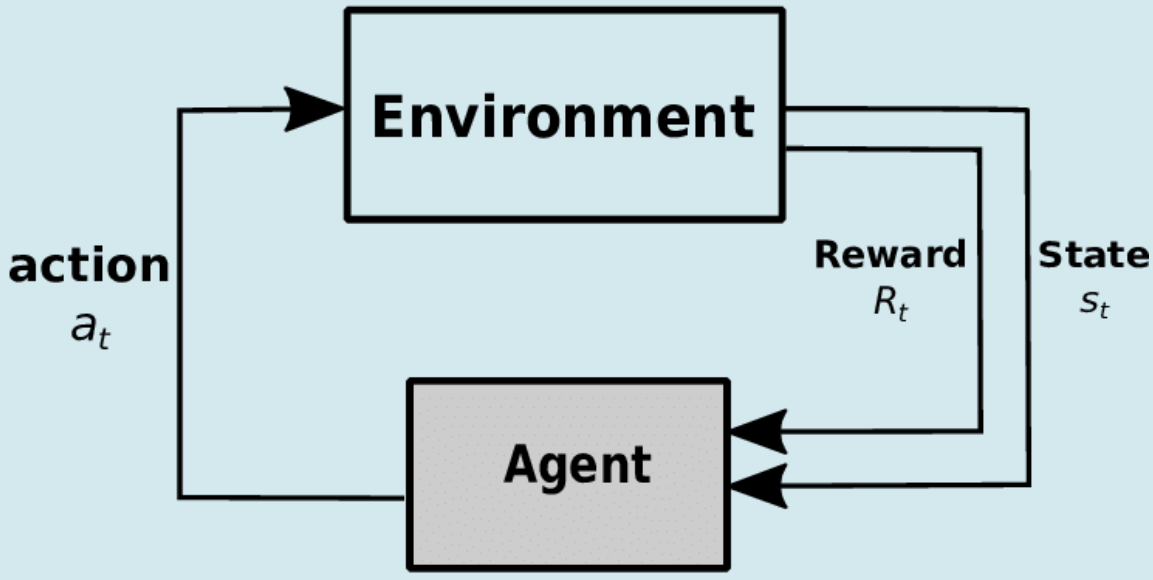
ABSTRACT

Although this report deals with the mechanisms of *artificially intelligent* rather than *intelligence* agents, the former is no less a subject of fascination. My research centred around an algorithm called Training an Agent Manually via Evaluative Reinforcement (TAMER), which incorporates human feedback into a reinforcement learning model. I ran several trials in the Mountain Car environment provided by the OpenAI gym library, altering the uniform value, credit assignment value, and budget of each to see which changes returned the best performance for the agent. Ultimately, lower credit assignment values and uniform values that are slightly better than those an average human trainer can provide are most effective in improving the performance of the agent, while the budget does not have a significant effect on the agent's efficiency.

UNIVERSITY OF ALBERTA

INTRODUCTION

Artificial intelligence is a field of research that has impacted every facet of digital life, and the concept of reinforcement learning (RL) has been crucial to the discipline. RL, a model of learning used to train AI agents, is often mentioned alongside supervised and unsupervised learning, all of which are subcategories housed under the roof of machine learning [Taylor et al. 2021]. While supervised learning agents have a trainer specifying which actions are correct, and unsupervised agents learn by looking for patterns in a given set of data, RL is independent of both. RL is conceptually intuitive because it mimics the human learning process; an agent performs an action in an environment, and in turn, the environment returns a reward signal and a new state to the agent. The goal of the agent is to maximize the cumulative value of the numerical rewards it receives [Sutton and Barto 2018].



Amiri, R., 2018. *A Machine Learning Approach for Power Allocation in HetNets Considering QoS.*

WHAT IS AN MDP?

To understand RL in a mathematical way, we can use a Markov Decision Process (MDP).

An MDP is a 5 tuple [Jagtap 2020] (S, A, p, r, γ), where:
S - states in the environment
A - actions the agent can perform
p - probability of reaching a given state after having taken an action
r - a reward function that returns the reward of reaching a given state
γ - a discount factor that determines how heavily future reward is weighted [Taylor et al. 2021].

WHAT IS TAMER?

TAMER - an acronym for Training an Agent Manually via Evaluative Reinforcement - is a supervised learning algorithm that integrates human feedback into the RL model. TAMER uses a modified MDP, where human feedback replaces the environmental reward as the value of r [Muslimani 2021]. In traditional RL, the agent performs an action upon the environment, which returns a numerical reward signal and a new state for the agent to learn from. TAMER builds upon this basic concept. A human trainer provides feedback to a TAMER agent, which is composed of a credit assigner, a supervised learner, and an action selector. The role of the credit assigner is to mitigate the slight delay between the perceived input of feedback from a human trainer and the reception of the feedback from the agent (more discussion below). The credit assigner distributes human feedback among state-action pairs. For every state-action pair input, the credit assigner filters out an adjusted output of human feedback. This human feedback goes through a supervised learner, which uses the new feedback as input for a regression algorithm that outputs a reward model. The action selector consults the output feedback model to determine which action is likely to return the greatest human feedback [Knox 2020]. TAMER allows the human to submit both positive and negative feedback to the credit assigner, so the learning process of the agent is not dependent on trial and error.

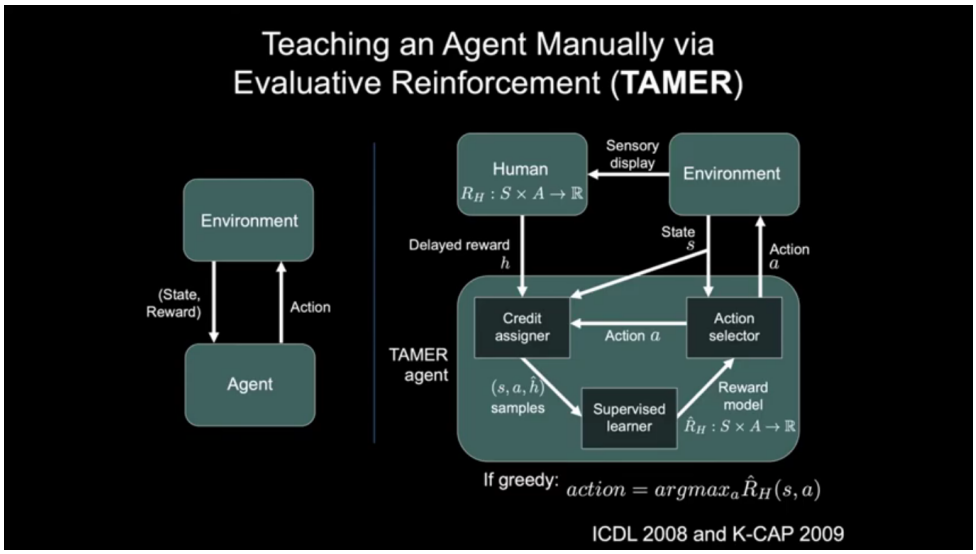
In its learning phase, TAMER can receive feedback from either a human trainer or an artificially intelligent trainer that has already been through the learning process. Our experiment will be using an AI trainer for two reasons. First, using an artificial teacher will allow for a more controlled experiment. Human trainers introducing factors like feedback delay and a gradually diminishing attention span and contribute to an increase in uncertainty, but an AI teacher circumvents these issues [Muslimani 2021]. Second, we assume that the AI agent has perfect knowledge of the environment and - unlike a human trainer - can provide a reward at every time-step. However, this experiment will simulate the experience of being taught by a human trainer by increasing the uniform value. A 2017 paper by Warner et al. "observed that human feedback is typically provided at a rate of approximately one signal every 25 time steps," so the initial uniform value will be set at 25 to approximate authentic human feedback.

Three variables are relevant to this problem: the credit assignment value, the uniform value, and the budget. The **credit assignment value** is the number of time steps across which the reward is spread. The **uniform value** is the interval at which the human reward is provided.

To illustrate the concept more clearly, we can imagine a scenario with eight state-action pairs. If the uniform value is four, every fourth state-action pair will receive a reward. If the credit assignment value is two, the reward will be spread across the two state-action pairs leading up to the state-action pair specified by the uniform value. In this case, the third and fourth SA pair will receive a reward, as will the seventh and eighth, given in the diagram by my supervisor Calarina Muslimani below:

(S1,A1) → (S2,A2) → ***(S3,A3)*** → ***(S4,A4)*** → (S5,A5) → (S6,A6) → ***(S7,A7)*** → ***(S8,A8)***

Finally, the **budget** is the total number of times the teacher can provide the agent with feedback before the episode is over [Muslimani 2021]. An episode is the sequence of events that occurs before the agent reaches a terminal state, and every event in an episode is separated by a time-step [Dernoncourt 2016].



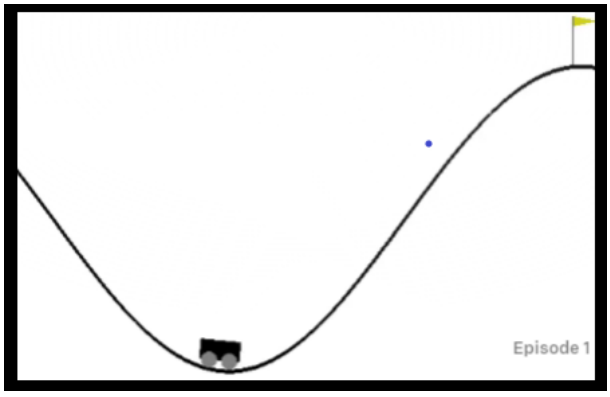
Knox, B., 2020. *Brad Knox at ICSR-13: Teaching a robot via human feedback (for class 9)*.

PROBLEM

A communication gap becomes apparent when human trainers try to assign credit to agents trained using RL techniques. The trainer, observing the behaviour of the agent, knows which event it wants to reward, but because of delays in response time, they cannot know the time that the agent receives this feedback with accuracy. Conversely, the agent is aware of when exactly it receives feedback, but it does not know which event the trainer intended to reward. This is called the credit assignment problem, and it raises the question - how do we bridge the chasm between an agent and its trainer?

METHODS

My experiment takes place in the Mountain Car environment provided by OpenAI Gym, a tool for evaluating machine learning algorithms (pictured below). The goal of the experiment is to see which changes to the uniform value, credit assignment value, and budget are most effective in improving the performance of the agent. The progress of the agent is monitored via a graph updated with every new trial. The x-axis maps the number of episodes the agent experiences, while the y-axis gives the number of time-steps it took to complete each episode. A lower value on the y-axis means the agent learned in less time, which shows an improvement in the agent's behaviour.



I began by changing the credit assignment value while keeping a constant uniform value and budget, to see which credit assignment value yielded the best results. I went on to change the uniform value while keeping the credit assignment value and budget constant, and then to manipulate the budget while controlling the credit assignment and uniform values. I ran approximately twenty trials for each experiment, but will only feature those that were relevant to my research.

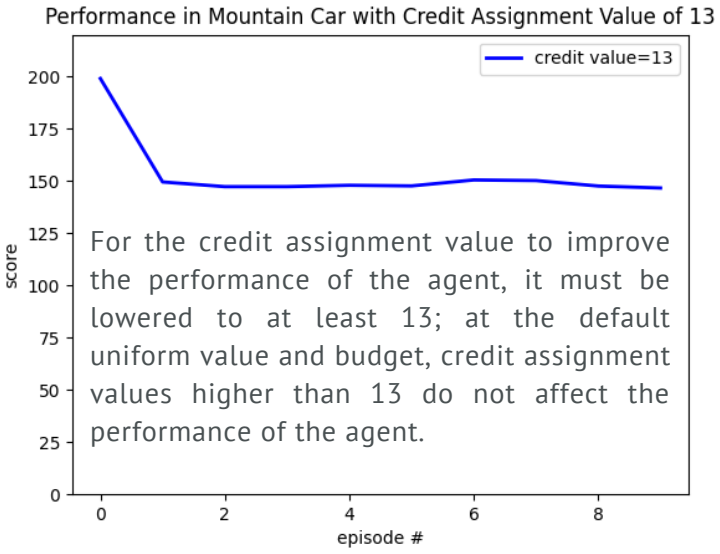
06

Best credit assignment value: ***1***

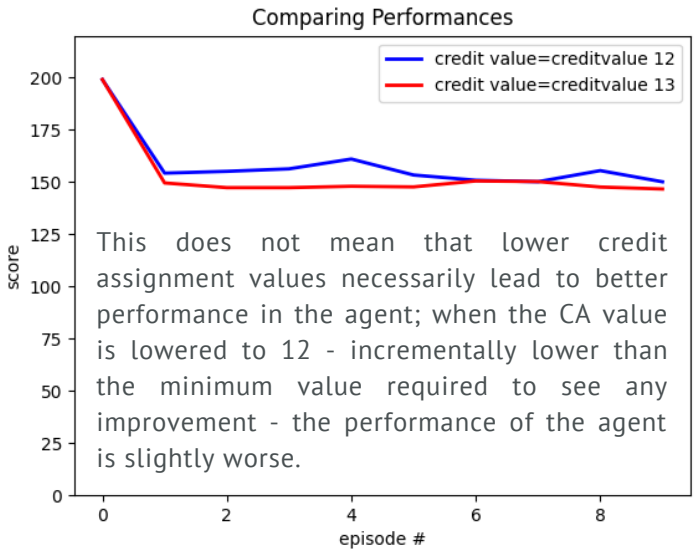
Score: ***114.45***

RESULTS & DISCUSSION

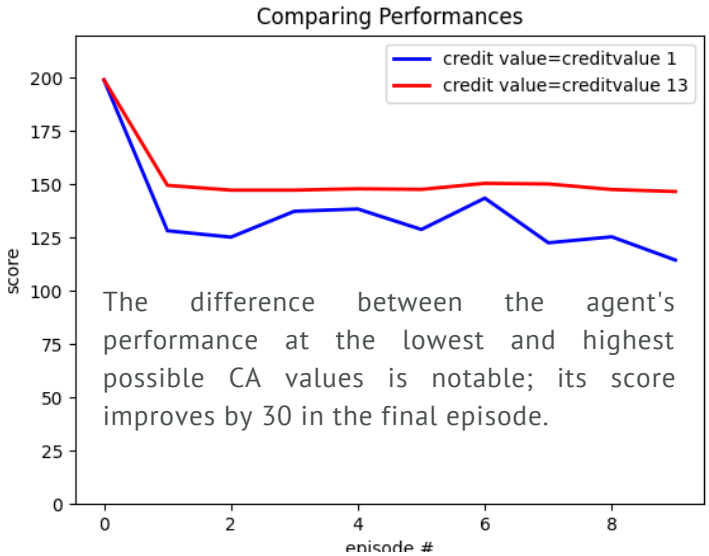
The following experiments were run with a constant uniform value (25) and budget (100), while the **credit assignment (CA) value** changes. The lower credit assignment value is blue, while the higher value is red



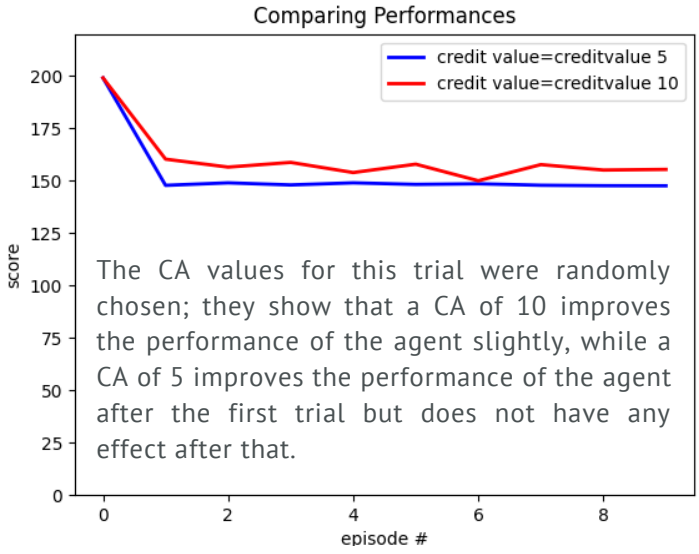
For the credit assignment value to improve the performance of the agent, it must be lowered to at least 13; at the default uniform value and budget, credit assignment values higher than 13 do not affect the performance of the agent.



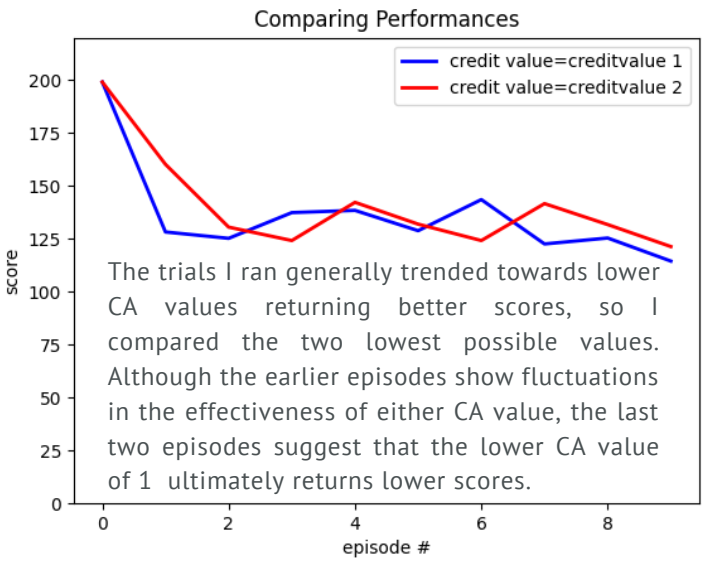
This does not mean that lower credit assignment values necessarily lead to better performance in the agent; when the CA value is lowered to 12 - incrementally lower than the minimum value required to see any improvement - the performance of the agent is slightly worse.



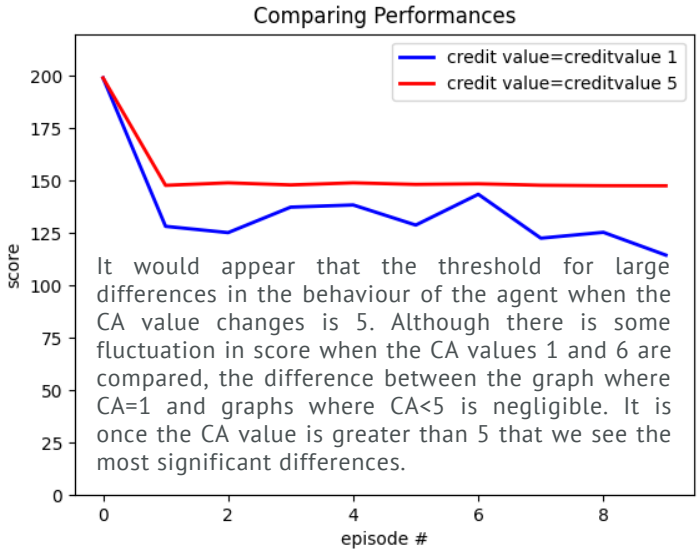
The difference between the agent's performance at the lowest and highest possible CA values is notable; its score improves by 30 in the final episode.



The CA values for this trial were randomly chosen; they show that a CA of 10 improves the performance of the agent slightly, while a CA of 5 improves the performance of the agent after the first trial but does not have any effect after that.



The trials I ran generally trended towards lower CA values returning better scores, so I compared the two lowest possible values. Although the earlier episodes show fluctuations in the effectiveness of either CA value, the last two episodes suggest that the lower CA value of 1 ultimately returns lower scores.

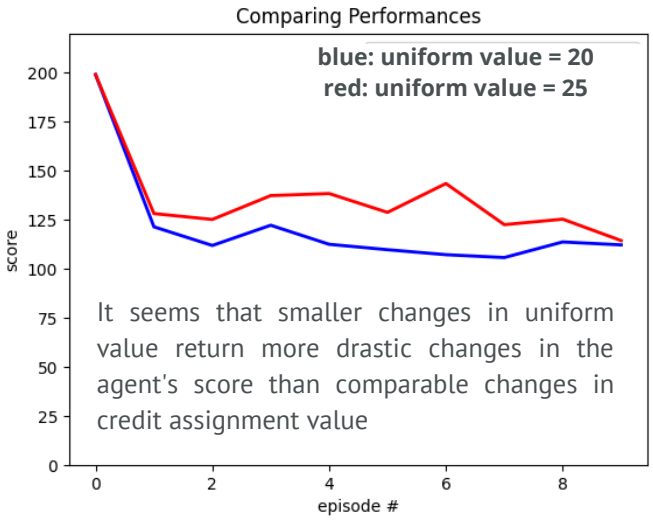


It would appear that the threshold for large differences in the behaviour of the agent when the CA value changes is 5. Although there is some fluctuation in score when the CA values 1 and 6 are compared, the difference between the graph where CA=1 and graphs where CA<5 is negligible. It is once the CA value is greater than 5 that we see the most significant differences.

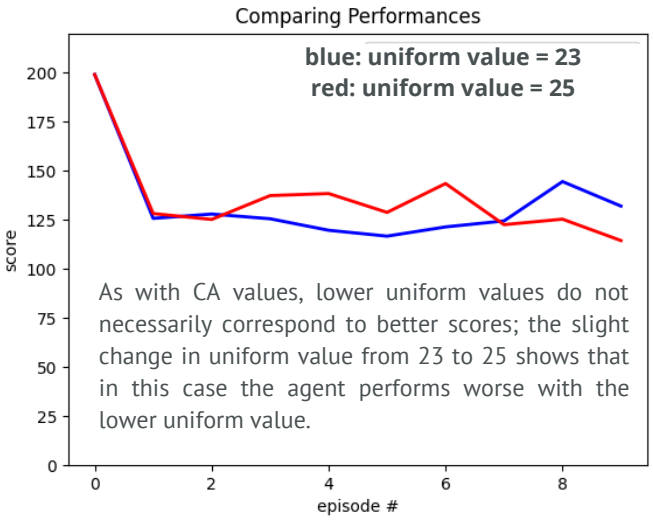
Best uniform value : ***18***
Score: ***135.5***

RESULTS & DISCUSSION (CONT.)

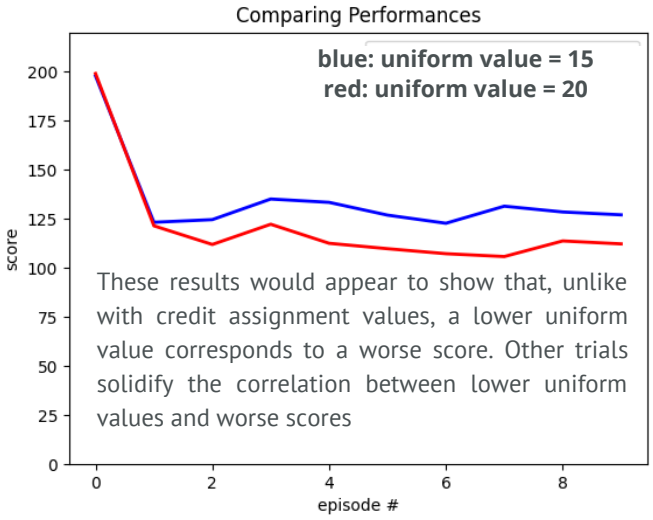
The following experiments were run with a constant credit assignment value (1) and budget (100), while the **uniform value** changes. The lower uniform value is blue, while the higher value is red.



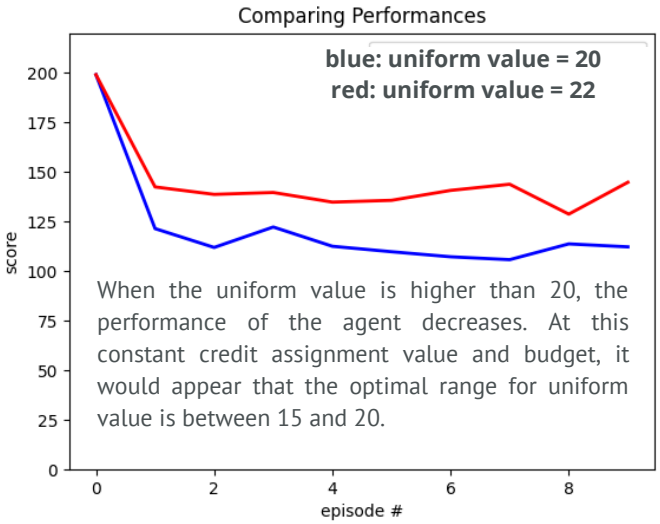
It seems that smaller changes in uniform value return more drastic changes in the agent's score than comparable changes in credit assignment value



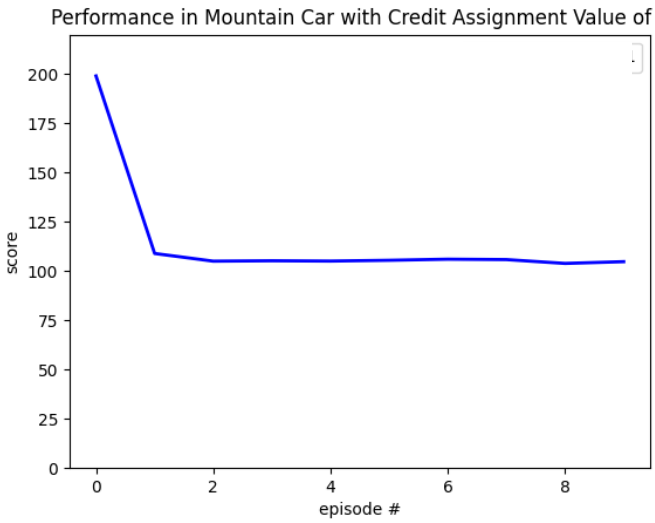
As with CA values, lower uniform values do not necessarily correspond to better scores; the slight change in uniform value from 23 to 25 shows that in this case the agent performs worse with the lower uniform value.



These results would appear to show that, unlike with credit assignment values, a lower uniform value corresponds to a worse score. Other trials solidify the correlation between lower uniform values and worse scores



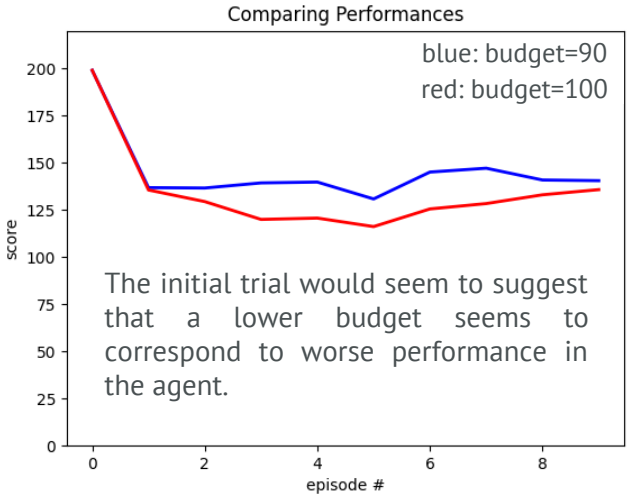
When the uniform value is higher than 20, the performance of the agent decreases. At this constant credit assignment value and budget, it would appear that the optimal range for uniform value is between 15 and 20.



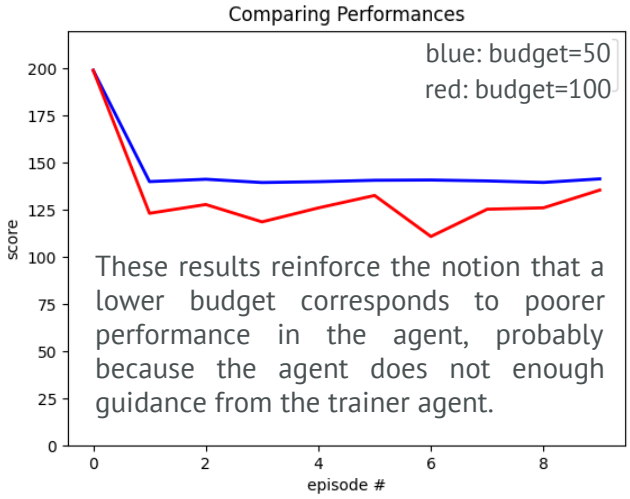
Uniform value = 18
The uniform value of 18 shows a better final score than one of 20, but the agent's performance does not change significantly after the first episode. On the other hand, uniform values lower than 18 show a decrease in the agent's performance. This implies that at our default credit assignment value and budget the optimal uniform value is 18.

RESULTS & DISCUSSION (CONT.)

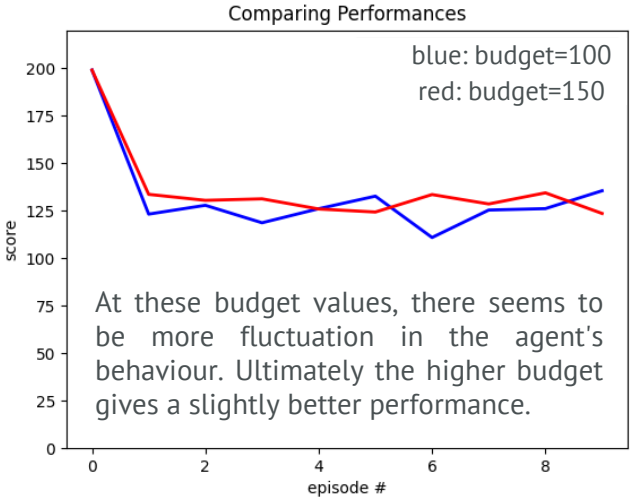
The following experiments were run with a constant credit assignment value (1) and uniform value (18), while the **budget** changes. The lower budget is blue, while the higher budget is red.



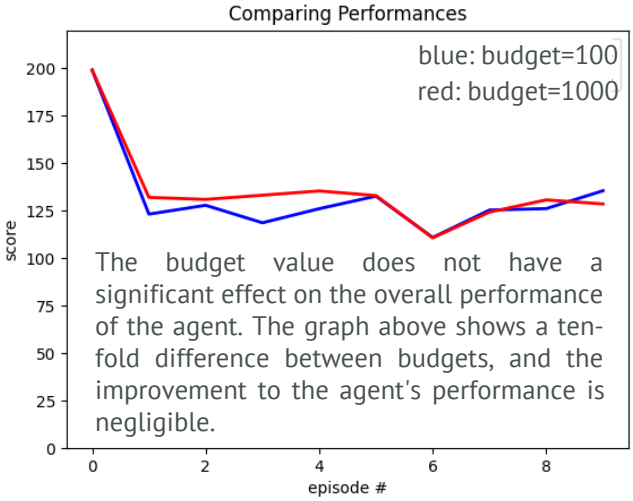
The initial trial would seem to suggest that a lower budget seems to correspond to worse performance in the agent.



These results reinforce the notion that a lower budget corresponds to poorer performance in the agent, probably because the agent does not enough guidance from the trainer agent.



At these budget values, there seems to be more fluctuation in the agent's behaviour. Ultimately the higher budget gives a slightly better performance.



The budget value does not have a significant effect on the overall performance of the agent. The graph above shows a ten-fold difference between budgets, and the improvement to the agent's performance is negligible.

TAKEAWAYS

The **credit assignment** has an optimal value at 1, but between 1 and 5 the fluctuations in the agent's score are insignificant. This shows that the agent performs better when reward is spread among the most recent state-action pairs.

The optimal **uniform value** is 18, but the range of fluctuation in the agent's score when the uniform value is between 16 and 21 is not significant. This is fairly close to the approximate uniform value at which a human trainer would have been able to provide feedback, which shows that above-average human trainers are capable of training an agent to perform well.

The **budget** does not have a notable impact on the agent's score. As long as it is above 90, the agent's performance will not be significantly affected by the budget no matter how great the budget is. This implies that the agent does not need a significant amount of training to function at an optimal level.

09 CONCLUSION

With more time, I would have widened the scope of my project. I would like to have seen how the uniform value and credit assignment value interact by simultaneously changing the values. If the credit assignment value that yielded the lowest score was objectively the best possible value, any subsequent changes to the uniform value and budget should have given values lower than 114.45. This was not the case, showing that the uniform and credit assignment values work in tandem in ways that I did not explore over the course of this project.

The learning processes of the machines that integrate themselves into our public psyche is a fascinating subject, and to take an active role in the winding path technology charts, we must begin with an understanding of the unseen forces that shape our lives. This project is the first step towards a holistic understanding of how artificial intelligence agents work - and thus, a first step towards acting as a participant rather than an observer in an increasingly data-driven world.

ACKNOWLEDGEMENTS

I would like to thank my principal investigator, Dr. Matthew Taylor, and my supervisor, Ms. Calarina Muslimani for their continual support and wisdom. The research that I have done this summer would not have been possible without them, and I am grateful for all of their time and kindness. Thank you to Nick Nissen for his patience and humour during the grueling hours it took to install Ubuntu on my poor Asus, and thank you to the IRL lab.

Thank you to the WISEST coordinators for making this summer a life-changing experience. I will never forget the invaluable six weeks of growth that AJ, Bridget, Deb, and Hannah have facilitated. The coordinators have shown me patience, care, and compassion, and all of my gratitude goes to them.

Huge thanks to the University of Alberta for providing the space to host the SRP, and to Syncrude for their financial support.

Adam Bignold et al. 2020. A Conceptual Framework for Externally-influenced Agents: An Assisted Reinforcement Learning Review. arXiv:2007.01544v1. Retrieved from: https://arxiv.org/pdf/2007.01544.pdf

Calarina Muslimani et al. 2021. Comparing Feedback Distributions in Limited Teacher-Student Settings. University of Alberta, Edmonton, AB. https://ala2021.vub.ac.be/papers/ALA2021_paper_37.pdf

Franck Dernoncourt (https://stats.stackexchange.com/users/12359/franck-dernoncourt), What is the difference between episode and epoch in deep Q learning?, URL (version: 2017-04-13): https://stats.stackexchange.com/q/250955 (Stack Exchange "what is an episode?")

Garret Warnell et al. 2018. Deep TAMER: Interactive Agent Shaping in High-Dimensional State Spaces. arXiv:1709.10163. Retrieved from https://arxiv.org/pdf/1709.10163.pdf

Interactive Machine Learning. 2020. Brad Knox at ICSR-13: Teaching a robot via human feedback (for class 9). Video. (29 September 2020). Retrieved July 2021 from https://www.youtube.com/watch?v=fb0IEhDLTPo

Jason Brownlee. 2019. A Gentle Introduction to Probability Density Estimation. Retrieved July 2021 from https://machinelearningmastery.com/probability-density-estimation/

Marco Wiering and Martijn van Otterlo. 2012. Reinforcement Learning. Adaption, Learning, and Optimization, Vol. 12. Nanyang Technological University, Singapore. DOI 10.1007/978-3-642-27645-3 https://link.springer.com/content/pdf/10.1007%2F978-3-642-27645-3.pdf

Matthew E. Taylor. 2021. Improving Reinforcement Learning with Human Assistance: An Argument for Human Subject Studies with HIPPO Gym. arXiv:2102.02639. Retrieved from https://arxiv.org/pdf/2102.02639.pdf

Richard S. Sutton and Andrew G. Barto. 2018. Reinforcement Learning (2nd. ed.). http://incompleteideas.net/book/RLbook2020.pdf

Rohan Jagtap. 2020. Understanding Markov Decision Process (MDP). Retrieved from https://towardsdatascience.com/understanding-the-markov-decision-process-mdp-8f838510f150

W. E. Hockley. 1984. Analysis of response time distributions in the study of cognitive processes. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10(4), 598. https://psycnet.apa.org/fulltext/1985-24150-001.pdf?auth_token=f15b72c59529b0b855321548eb4e6929d1345705

William Bradley Knox and Peter Stone. 2010. Combining Manual Feedback with Subsequent MDP Reward Signals for Reinforcement Learning. https://bradknox.net/public/papers/aamas10-knox.pdf

William Bradley Knox. 2021. Learning from Human-Generated Reward - W. Bradley Knox's 2012 PhD Defense on the TAMER Framework. Video. (10 March 2021). Retrieved July 2021 from https://www.youtube.com/watch?v=erUzgrUVPlQ

William Bradley Knox and Peter Stone. 2009. Interactively shaping agents via human reinforcement: the TAMER framework. In Proceedings of the fifth international conference on Knowledge capture (K-CAP '09). Association for Computing Machinery, New York, NY, USA, 9–16. DOI:https://doi.org/10.1145/1597735.1597738

William Bradley Knox. 2012. Learning from Human-Generated Reward. Ph. D. Dissertation. The University of Texas, Austin, TX. https://www.bradknox.net/wp-content/uploads/2013/06/thesis-knox.pdf

William Bradley Knox and Peter Stone. 2012. Reinforcement Learning from Simultaneous Human and MDP Reward. https://bradknox.net/public/papers/aamas12-knox.pdf

William Bradley Knox and Peter Stone, "TAMER: Training an Agent Manually via Evaluative Reinforcement," 2008 7th IEEE International Conference on Development and Learning, 2008, pp. 292-297, DOI: https://doi.org/10.1109/DEVLRN.2008.4640845.