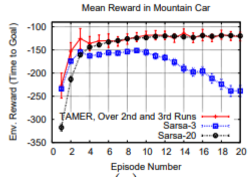
Social Robotics

Learning Using Human Feedback

Guennif/Ghaouel
3804567/3804527

Problem 1

Project	Algorithm	Feedback	Type	env	Learning	Results
Training an Agent Manually via Evaluative Reinforcement	Q-learning for the policy,The teacher gives a binary feedback or chooses to give none. These returns are a reward that is independent from the env rewards .They shape the behavior /policy of the agent .	keyboard (W = positive reward A= negative reward)	Evaluative	-gym car mountain -Tetris	RL	 The environment's reward converge to the same value -140 for mountain car which means that the task is successful
IRL imitation	Linear inverse reinforcement learning given Q-learned policy	we give expert behavior policy	imitation given an observed optimal behavior.	-gym cart pole -car mountain on the git	IRL adaptive learning	The learning speeds up with an adaptive teacher even if he has limited information. The results underlight the necessity of the agent representation of feature space and reward function.These two elements are used to shape the desired policy.

Learning-Behaviors-with-Uncertain-Human-Feedback	As trainers are more likely to give positive feedback to sub-optimal actions and negative feedback to good actions and to compensate that the researchers use the Expectation Maximization with the approximation of the expectation step using descent gradient to assume the feedback as hidden parameter to consider uncertainty	keyboard	Evalu ative	rat chasing	RL	The more rats we chase successfully per step the more we converge rapidly
Interactive Reinforcement Learning with Dynamic Reuse of Prior Knowledge	Target Learning Bootstrap	keyboard	Demonstration	gym cart pole	RL	For Cartpole , Teachers first let a trained agent demon strate 20 episodes (average number of steps : 821+-105) and if the prior knowledge is imperfect,the agent is forced by prior knowledge . The agent sometime repeat suboptimal actions.

Problem 2

Replicating the results

In the tamer project, the researchers chose to study only giving to the agent binary feedback occasionally. The tamer approach makes the evaluation using a human value function, the teacher can either give a feedback or not. We consider only past actions and we cannot give a feedback during the actions . Results depend highly on the teacher capacities . If the teacher decides to give feed backs too frequently, the agent explore less .
We evaluate during 20 episodes . We get an Average total episode reward over 20 episodes: -140.05 Which is a similar result than the one described on the paper .

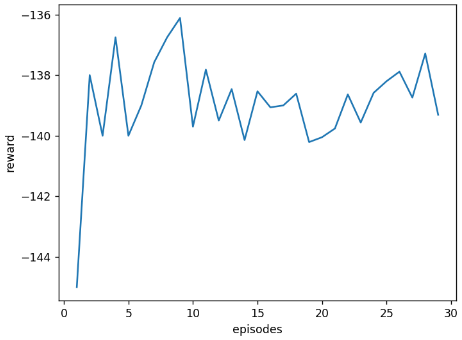


Figure 1: Evaluation's mean rewards for 20 episodes

Evaluative Learning

This type of learning requires that the teacher pays attention for a transition from the state to a next one and evaluate the quality of the action that the agent took. We replicated the same results as the researchers by achieving a reward that varies from -145 to -140 at average . The results depend mainly on the teacher, for example if we try to only go to the right which is not intuitive as we'd rather go backward to get a bigger impulse when going forward , we may succeed during training but the car will not succeed when trying for the evaluation.

Problem 2 continued on next page. . .

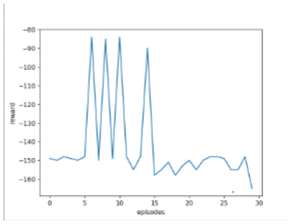
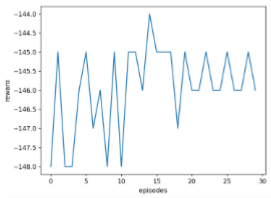
	
Average total episode reward over 30 episodes: -142.83	Average total episode reward over 30 episodes: -145.93

Figure 2: The teacher intervene only by giving a feedback

We can see in the precedent figures that using the Tamer method is successful for each episode as we get to the flag each time, the only variation is the time that it will take to get to it . The training for the left figure is done by giving feed backs that push the car to the left than go right. Results are way better when we do that compared to the right figure which is the rewards that we have when giving feedback only when the car go right and bad feed backs when it goes left so it does not learn to take an impulse.

Problem 3

Alternative methods

We decided to try two methods :

Descriptive Learning

Instead of giving a feed back for the state s of the car, we give an action a that we think is better for the training.We plot the average reward for 20 episodes at evaluation we get :

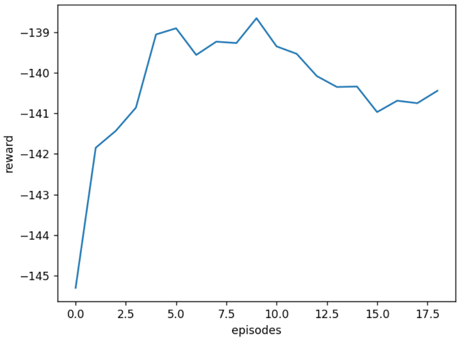


Figure 3: Evaluation's rewards for 20 episodes

The average result for the reward over the 20 episodes we got -140.8 which is similar than the result gotten by feed backs (Evaluative learning) . We only give occasionally an action that we think the agent should take . Results :

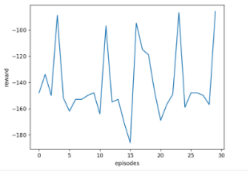
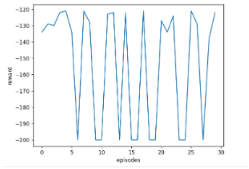
	
Average total episode reward over 30 episodes: -141.70	Average total episode reward over 30 episodes: -153.40

Figure 4: The teacher intervene only by forcing the car to an action

In the figure on the left we train the car by encouraging it to go left than go right. On the other one, we encourage it to go mainly on the right without trying to back first .The movement is not really successful on the first episode on training but it is for the second. We get an average reward of -153 but this result is not satisfactory as we clearly see that nearly half of the episodes the car fails to achieve the flag .

Problem 3 continued on next page. . .

When we see that on the left train when the teacher focuses on training using a more intuitive policy to the car, it never fails to achieve the task. The car is also faster to achieve the task by nearly 12 steps.

Evaluative and Descriptive Learning

In this part, the teacher lets the car do an action, he either gives a feedback or changes the action occasionally . We get for 20 episodes :

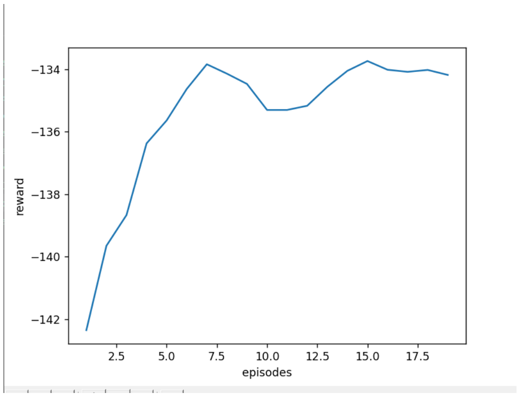


Figure 5: Evaluation's rewards for 20 episodes

We plot then the rewards :

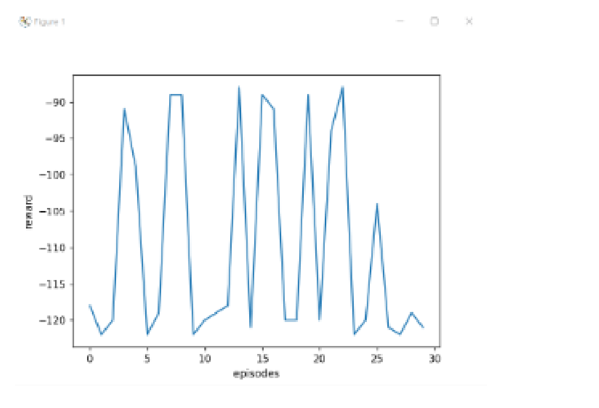
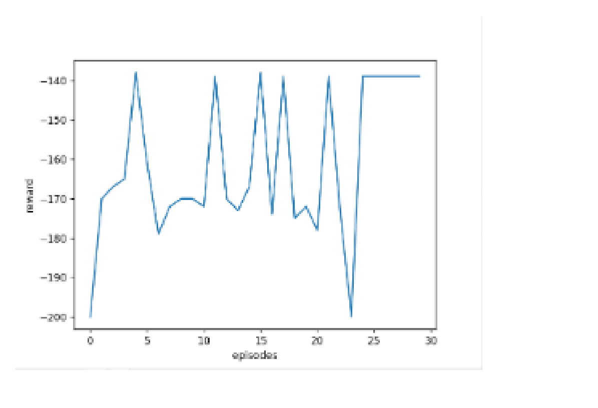
	
Average total episode reward over 30 episodes: -109.90	Average total episode reward over 30 episodes: -161.17

Figure 6: The teacher intervene by forcing the car to an action or by giving a feedback

During the train for which we plotted the reward for each episode on the left, We get a very good reward on average. We clearly see that even the worst results for this method (-120) are better than the ones we got for the other methods . For the right figure, the teacher gives a feedback and changes the actions more often but giving bad instructions. We clearly see a degradation of the results even though the car mostly achieves its task(except twice), it's too slow.

Speech recognition

The function that we wrote get_sound() records an audio using methods of the Pyaudio library, the teacher gives a feedback,we read the recording using the wave library, then we use the google recognizer from the speech recognizer library to get a transcription of what was said in the recording .

We make the teacher give a vocal feedback to the agent . We chose the following words : ”yes”,”no”,”back”,”forward”. These words have been chosen because they were easily recognised by the google recognizer.

Then we either add a reward or remove it depending on the feed-back inside get_scalar_feedback function.

Gestures recognition

We get the feedback by showing a thumbs up to a good action and thumbs down for a bad action we chose the stop for backward mouvement and the smile for going forward . We use the code of TechVidvan using the MediaPipe project more specifically the hand recognition algorithm . We only detect one hand at time. We use the tensoreflow pre-trained model ‘my_hand_gesture’. It can detect a lot of hand moves but for now we only need 4 gestures .

Problem 4

Conclusion

Graphics

This task never succeed if there is not human intervention. Giving few feed backs but chosen well is more successful than giving not well chosen feed backs or actions. The mean reward can be a good indicative but if we train the agent for example during 2 episodes, for the first we give a good policy for the second we do not do well, the agent will have excellent rewards half of the time during inference and will fail half of the time . That is why we chose to plot for each method not only the mean of the reward but also the reward to see how well the car is doing each time .

Methods

The evaluative learning and the descriptive method have similar results . But they do not take fully advantage of the teacher expertise . Using both gives to the teacher freedom to the teacher to either leave the freedom to the agent to do an action and give him a feedback or to force him to an action .

Teaching quality

The results are deeply impacted by the choices of the teacher. These methods should not be used if the teacher is not an expert .

Feed backs channels

The teacher can use 3 ways to give information, keyboard,speech or gestures. The keyboard method is fast , easy to implement but is more challenging for the user as he needs to previously understand which key affect the agent .While using the keyboard the teacher can use it more than the gestures or the speech as there is almost no delay between the action and the moment when he can give a feedback .

As for the Speech, the recording can be problematic if the teacher does not pronounce well the chosen words . There is also the adding of a delay in which the recognizer needs to understand the content of the recording than transcript it to be able to use it in our application.It makes the process less fluid .

Lastly, the use of the gestures recognizer is not adviced as it slows drasticlly the process .

For this application, we'd rather keep a simple way of communication that does not slow the process even if it is not intuitive to the user which is keyboard .

References

W. Bradley Knox and P. Stone, ”TAMER: Training an Agent Manually via Evaluative Reinforcement,” 2008 7th IEEE International Conference on Development and Learning, 2008, pp. 292-297, doi: 10.1109/DEVLRN.2008.4640845.

Parameswaran Kamalaruban, Rati Devidze, Volkan Cevher, Adish Singla: Interactive Teaching Algorithms for Inverse Reinforcement Learning. CoRR abs/1905.11867 (2019)

Xu He, Haipeng Chen, Bo An: Learning Behaviors with Uncertain Human Feedback. CoRR abs/2006.04201 (2020)

Zhaodong Wang and Matthew E. Taylor. 2019. Interactive reinforcement learning with dynamic reuse of prior knowledge from human and agent demonstrations. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19). AAAI Press, 3820–3827.