CS294 HW1

Heri Zhao

Section 2 Behavioral Cloning

1. With rollout = 100, we have

	mean	std
Ant-v2	4812.829515	259.567021
HalfCheetah-v2	4152.797563	90.270327
Hopper-v2	3778.708843	3.83649
Humanoid-v2	10308.81148	978.375883
Reacher-v2	-4.080761	1.887553
Walker2d-v2	5519.265739	72.801996

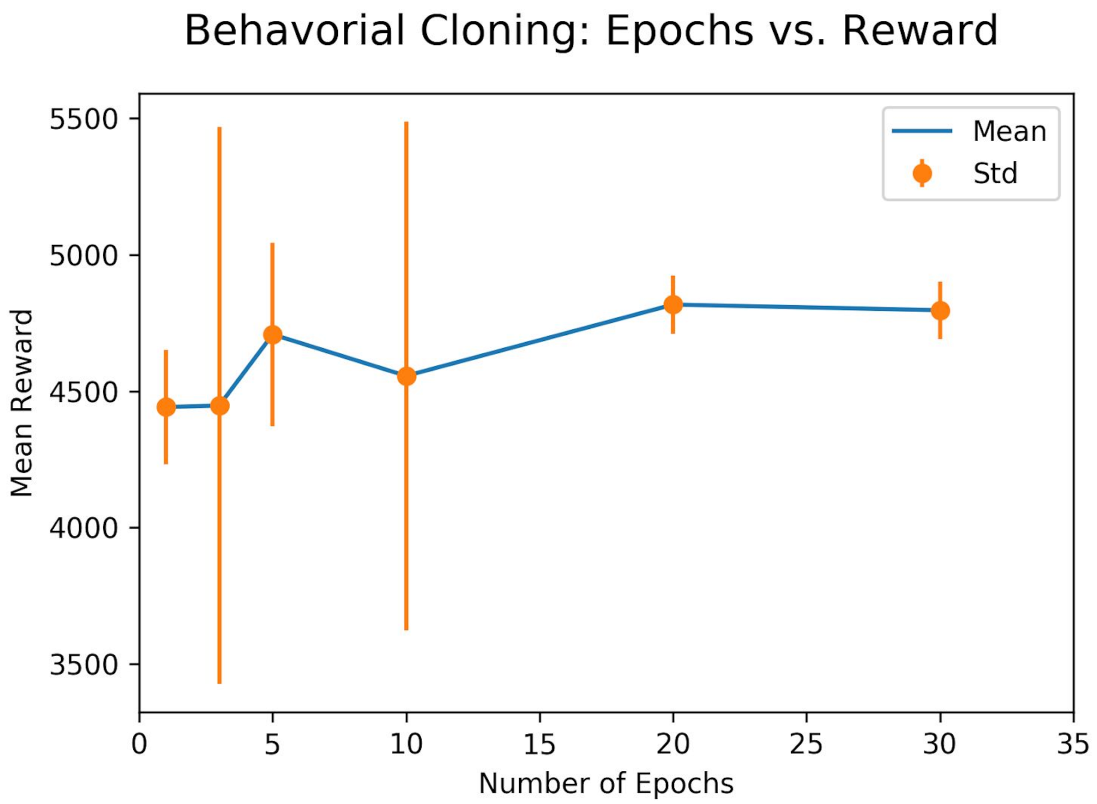
2. Use the same NN to train all the tasks.
3 fully connected layer network with number of hidden variables 128 for each layer.
Training data: 100-rollout data
Training epochs: 30

	rollout	Expert mean	Expert std	Bc mean	Bc std
Ant-v2	20	4832.54	73.56	4796.71	105.68
HalfCheetah-v2	20	4159.65	68.62	4134.81	95.14
Hopper-v2	20	3779.55	3.94	2817.24	505.22
Humanoid-v2	20	9961.02	1915.15	2923.62	2069.12
Reacher-v2	20	-3.37	1.62	-5.21	2.12
Walker2d-v2	20	5497.03	69.91	5518.19	40.36

We can see that except Hopper and Humanoid, behavioral cloning achieves comparable performance on the other tasks.

3. Experiment with hyperparameter - **number of training epoch**

	1	3	5	10	20	30
mean	4441.38	4447.10	4707.91	4555.75	4817.26	4796.71
std	210.81	1021.70	337.22	933.02	107.66	105.68



Task: **Ant-v2**
rollout = 20

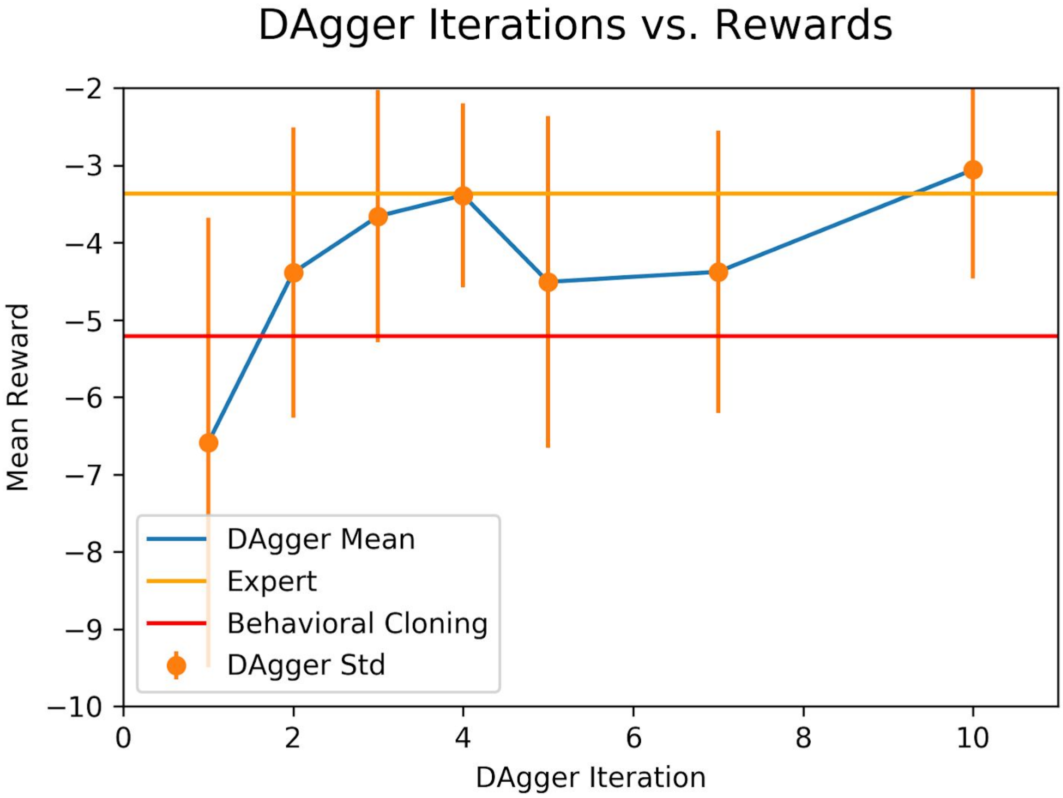
We can see that with training epochs increasing, the mean value is getting closed to expert's mean, and almost become stable when epoch size gets larger. We can also see that the std is getting smaller.

Retionale: Since our model is very simple, and it is behavioral cloning, it is very intuitive that if we train a model more epochs, the model should look like training data more.

Section 3 DAgger

We run DAgger on task **Reacher**.

	1	2	3	4	5	7	10
mean	-6.59	-4.39	-3.66	-3.39	-4.51	-4.38	-3.06
std	2.91	1.88	1.63	1.19	2.15	1.83	1.41



Task **Reacher**

3-layer fully connected NN with hidden variables 128 for each layer.

Training set: 100 rollout of expert's policy

rollout for every DAgger iteration: 100 rollout label

Training epochs for every iteration: 10

We can see that with DAgger, the result is much more better than behavioral cloning. When we increase the DAgger iteration, the mean is much closer to expert's mean. Since every iteration, we add 100-rollout of data into the training set, for 10 iteration almost 90% of data are labeled from expert, it should behave more like expert.

Section 4 Alternative Policy Architectures

Add one more 128 fully connected layer to original BC:
4 fully connected layer network with number of hidden variables 128 for each layer.
Training data: 100-rollout data
Training epochs: 30

	Alternative BC mean	Alternative BC std	Bc mean	Bc std
Hopper-v2	1908.45	601.31	2817.24	505.22

We add one more layer, but the result is worse.

Add one more 128 fully connected layer to original DAgger:
4 fully connected layer network with number of hidden variables 128 for each layer.
Training data: 100-rollout data
Training epochs: 10
DAgger iteration: 10

	Alternative DAgger mean	Alternative DAgger std	DAgger mean	DAgger std
Hopper-v2	-3.82	1.63	-3.06	1.41

There is no much difference.