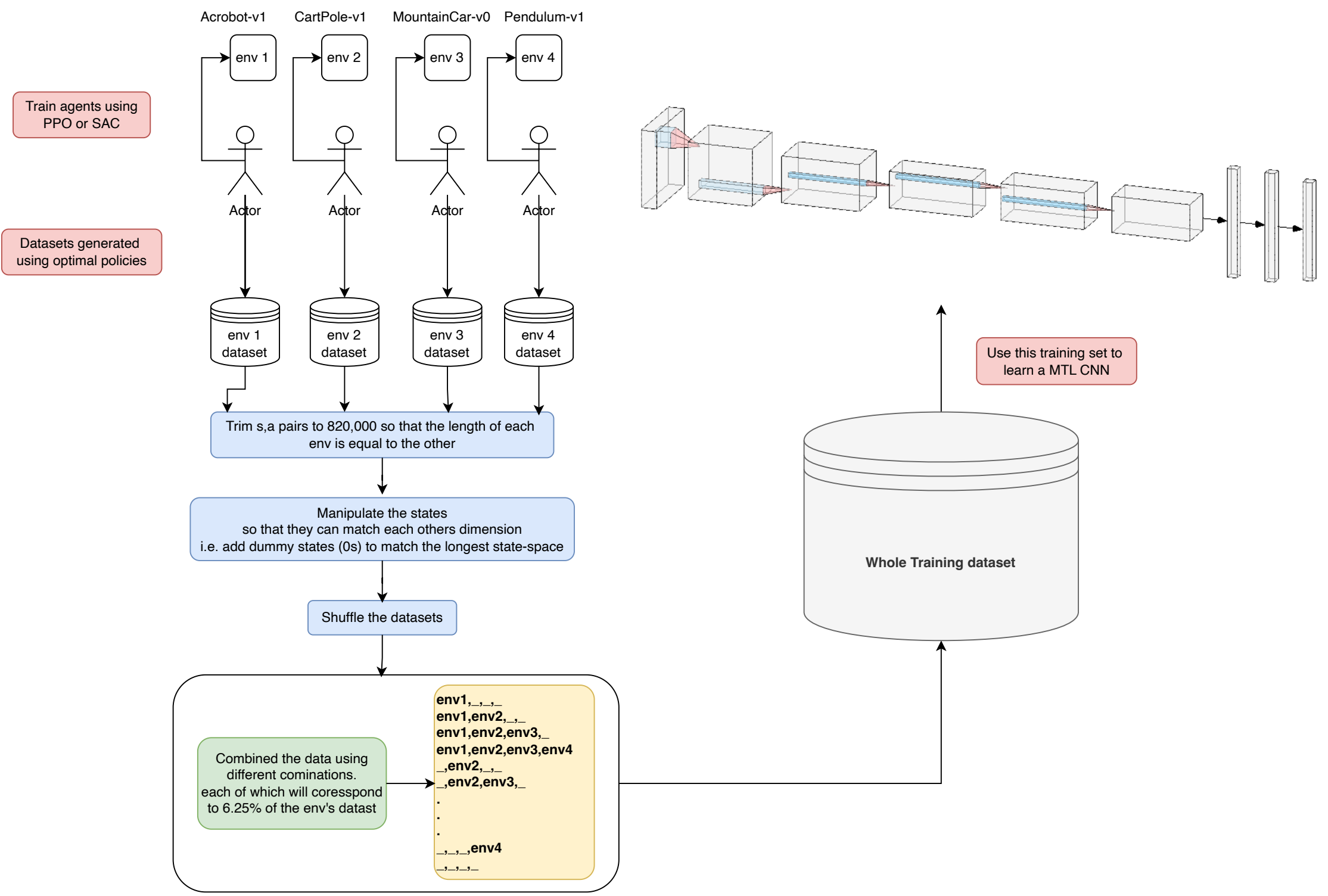
Acrobot-v1
CartPole-v1
MountainCar-v0
Pendulum-v1
env 1
env 2
env 3
env 4
Train agents using PPO or SAC
Actor
Actor
Actor
Actor
Datasets generated using optimal policies
env 1 dataset
env 2 dataset
env 3 dataset
env 4 dataset
Trim s,a pairs to 820,000 so that the length of each env is equal to the other
Manipulate the states so that they can match each others dimension i.e. add dummy states (0s) to match the longest state-space
Shuffle the datasets
Combined the data using different cominations. each of which will coresspond to 6.25% of the env's datast
env1,_,_,_
env1,env2,_,_
env1,env2,env3,_
env1,env2,env3,env4
,env2,,_
,env2,env3,
.
.
.
_,_,_,env4
_,_,_,_
Use this training set to learn a MTL CNN
Whole Training dataset