

Teacher-Student Curriculum Learning

Tambet Matiisen^{*,†}
University of Tartu

Avital Oliver^{*,‡}
OpenAI

Taco Cohen^{*}
University of Amsterdam

John Schulman
OpenAI

NIPS-2017

Introduction

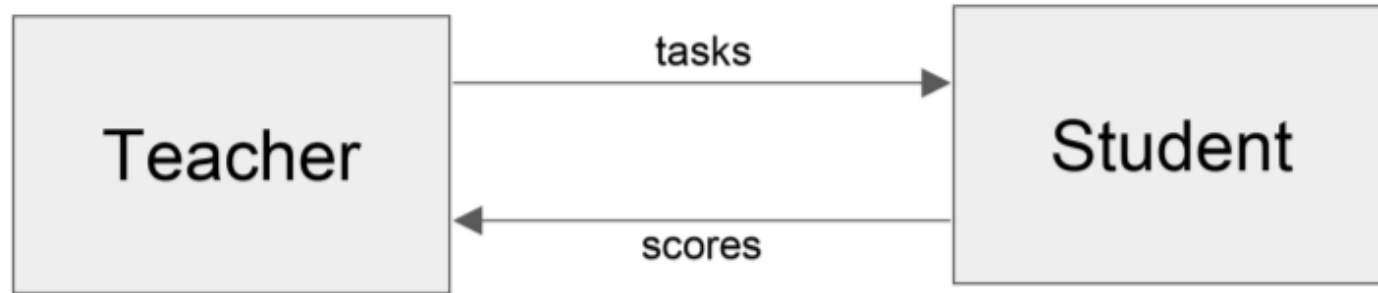
- Background
 - Some RL tasks are difficult. Learning is slow.
- Use the idea of curriculum learning. Broke the task into subtasks and learn the task from easy to hard.
- To use CL:
 - Manually order the difficulty of the subtasks.
 - need to decide a “mastery” threshold—when learning task stops.
 - requires prior knowledge.
- Goal: automatically select task for the model to learn.

RL review



Goal of RL: learn a policy—select action which can achieve maximum long-term reward at certain state.

Framework



- Teacher select task for the student to learn.
- Student returns scores on the task.

Formalize the process as MDP.

- Simple POMDP—select single task
- Batch POMDP—select tasks

Simple POMDP Formulation

s_t : state of the student(i.e neural network parameter, not observable)

a_t : select single task for student to learn.

o_t : the score of the task the student trained.

r_t : change of the score

Batch POMDP Formulation

s_t : state of the student(i.e neural network parameter, not observable)

a_t : probability over N tasks.

o_t : the scores of the tasks the student trained.

r_t : sum of the changes of the scores

Algorithms

Concrete algorithm based on two aspects: estimate learning performance and selecting task.

Estimate learning performance

- Native algorithm: train K times.
- Window algorithm: keep a buffer of K scores and timesteps of the score recorded.
Estimate the score by regression.

Selecting tasks

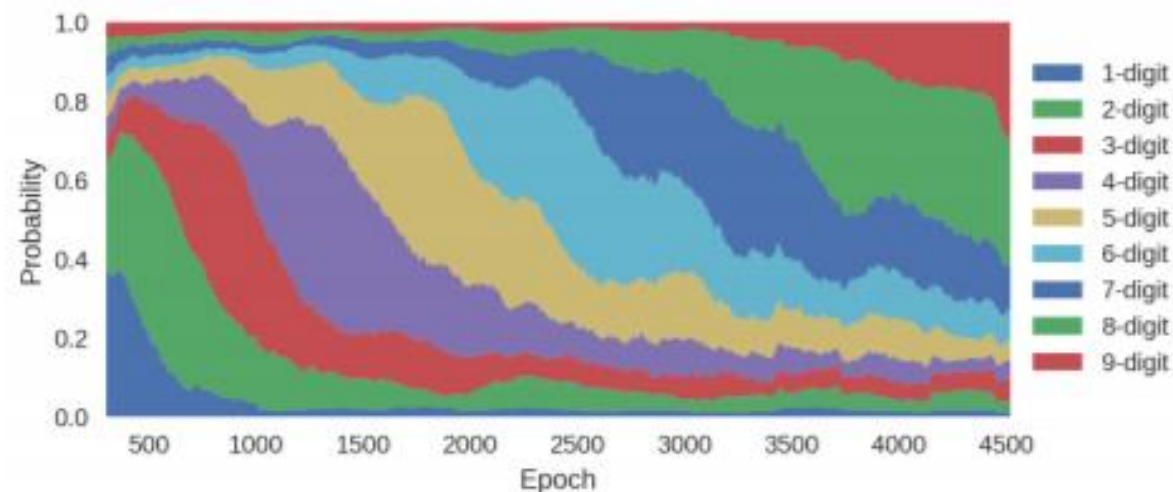
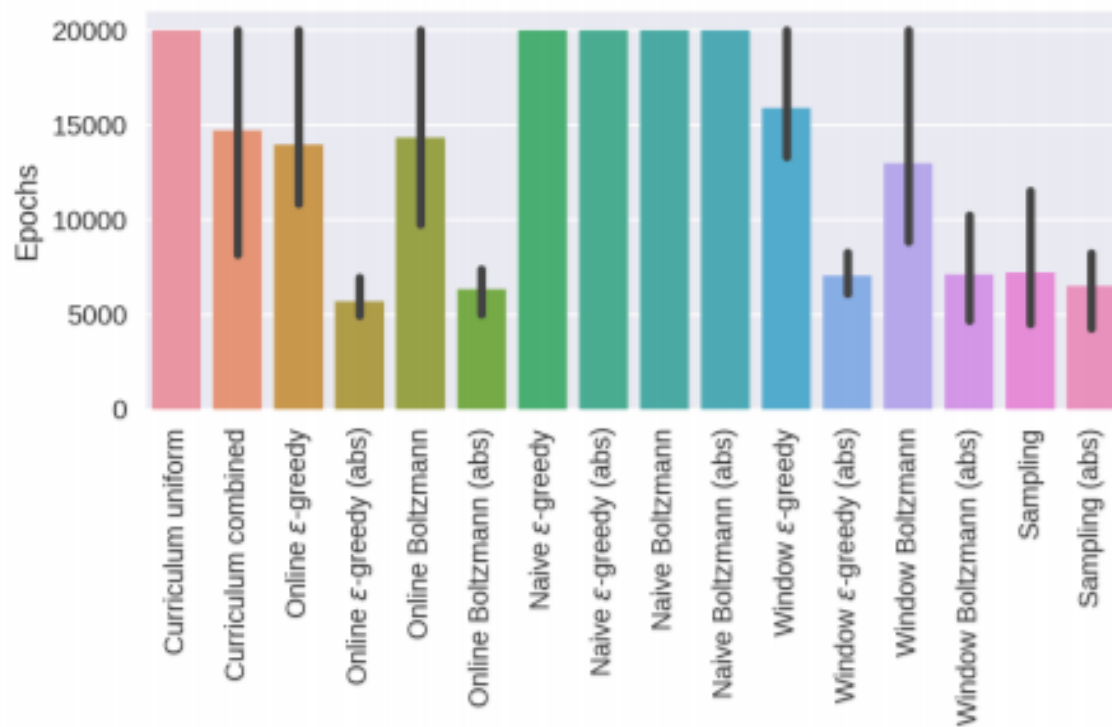
- Online algorithm: Select the task with ϵ -greedy or Boltzman distribution.
- Sampling algorithm: keeps a buffer of last K rewards for each task and sample from the buffer. Select task with highest sampled value.

Decimal Number Addition

Task: add two numbers.

Input: two numbers.

Output: sum of numbers.

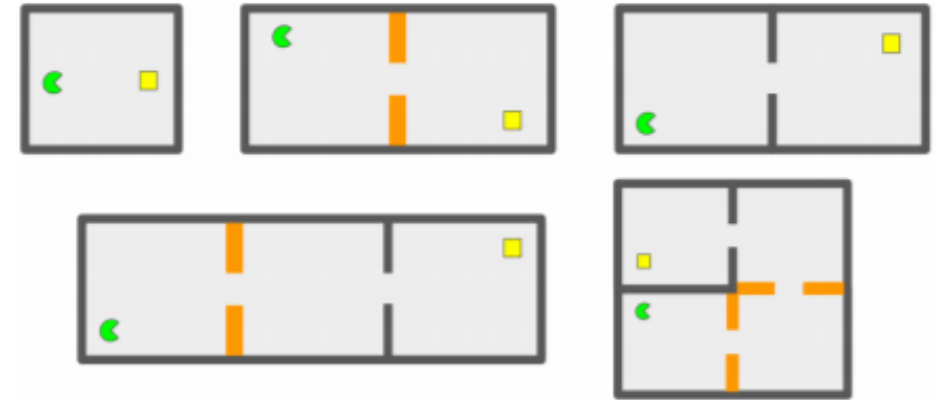


Minecraft

Minecraft is a video game. The agent has to cross a doorway in the wall. Or cross a bridge over lava to find the target.

We created a simple curriculum with 5 steps:

1. A single room with a target.
2. Two rooms separated by lava.
3. Two rooms separated by wall.
4. Three rooms separated by lava and wall, in random order.
5. Four rooms separated by lava and walls, in random order.



Minecraft Result

