Policy Consolidation for Continual Reinforcement Learning

Christos Kaplanis\textsuperscript{1}, Murray Shanahan\textsuperscript{1,2} and Claudia Clopath\textsuperscript{1}

\textsuperscript{1}Imperial College London, \textsuperscript{2}DeepMind

11th June 2019
Motivation

Catastrophic Forgetting in Artificial Neural Networks

Agents should cope with both discrete and continuous changes to data distribution without prior knowledge of when/how changes occur.

Test beds: alternating task, single task and multi-agent RL.
Motivation

- Catastrophic Forgetting in Artificial Neural Networks
Motivation

- Catastrophic Forgetting in Artificial Neural Networks

![Diagram showing alternating tasks between Task A and Task B]
Motivation

- Catastrophic Forgetting in Artificial Neural Networks

Task A  →  Task B  →  Task A

Questioning mindset icon
Motivation

- Catastrophic Forgetting in Artificial Neural Networks

![Diagram showing alternating tasks and single-task learning with question marks symbol]
Motivation

- Catastrophic Forgetting in Artificial Neural Networks

- Agents should cope with

![Diagram showing alternating tasks and single/multi-agent reinforcement learning test beds.](image-url)
Motivation

- Catastrophic Forgetting in Artificial Neural Networks

- Agents should cope with
  - Both discrete and continuous changes to data distribution
Motivation

- Catastrophic Forgetting in Artificial Neural Networks

- Agents should cope with
  - Both discrete and continuous changes to data distribution
  - No prior knowledge of when/how changes occur
Motivation

- Catastrophic Forgetting in Artificial Neural Networks

- Agents should cope with
  - Both discrete and continuous changes to data distribution
  - No prior knowledge of when/how changes occur

- Test beds: alternating task, single task and multi-agent RL
Policy Consolidation

\[ \pi_1 \ldots \pi_N \]

\[ \pi_1^{\text{old}} \ldots \pi_N^{\text{old}} \]

*Play game*
*Train agent*
*KL distillation loss*

*Store Policy*
*Recall Policy*
Alternating task experiments

[Walker2d-v2, Walker2dBigLeg-v0]

[HalfCheetah-v2, HalfCheetahBigLeg-v0]

[HumanoidSmallLeg-v0, HumanoidBigLeg-v0]
Single task experiments

- **[Walker2d-v2]**
- **[HalfCheetahBigLeg-v0]**
- **[RoboschoolHumanoid-v1]**

Each graph shows the reward over steps for different values of $\beta$ and clip values. The graphs illustrate the performance of various reinforcement learning algorithms across different tasks, highlighting the impact of hyperparameter tuning.
Multi-agent self-play experiments

(a) Final model vs. self history

(b) PC vs. baselines over training
Future work

- Prioritised consolidation
- Adapt for off-policy learning
Future work

- Prioritised consolidation
Future work

- Prioritised consolidation
- Adapt for off-policy learning