Lifelong Hanabi: Continuous coordination as a realistic scenario for lifelong learning

Hadi

Akilesh

Aaron

Sarath

ICML 2021

Mila

Université de Montréal

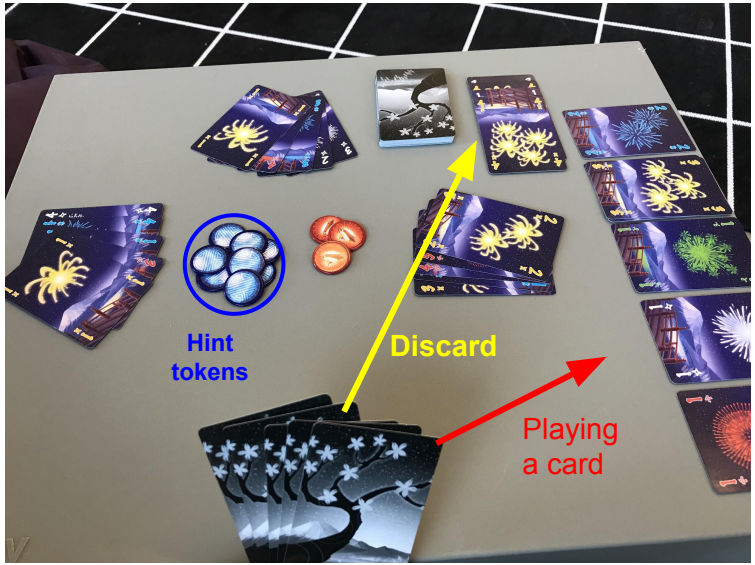
POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

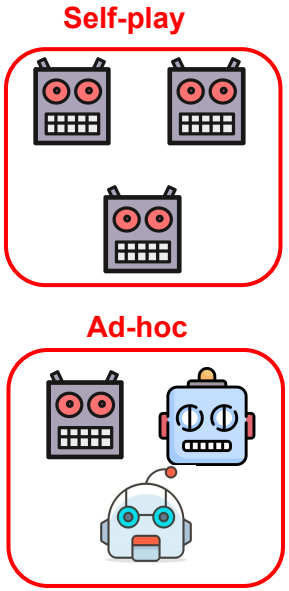
Talk Outline

- The Hanabi Challenge:
  - Self-play for Hanabi
  - The difficulty of the ad-hoc challenge (*Motivation 1*)
- Current Lifelong Learning benchmarks (*Motivation 2*)
- Lifelong Learning for MARL and MARL for Lifelong Learning
- Summary

The Hanabi Challenge: A New Frontier for AI Research

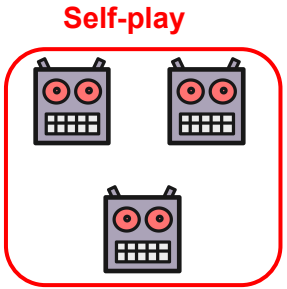
Nolan Bard[1,*], Jakob N. Foerster[2,*], Sarath Chandar[3], Neil Burch[1], Marc Lanctot[1], H. Francis Song[1], Emilio Parisotto[4], Vincent Dumoulin[3], Subhodeep Moitra[3], Edward Hughes[1], Iain Dunning[1], Shibl Mourad[1], Hugo Larochelle[3], Marc G. Bellemare[3], Michael Bowling[1]



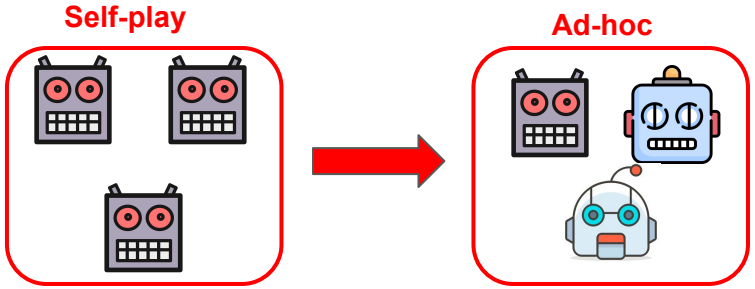


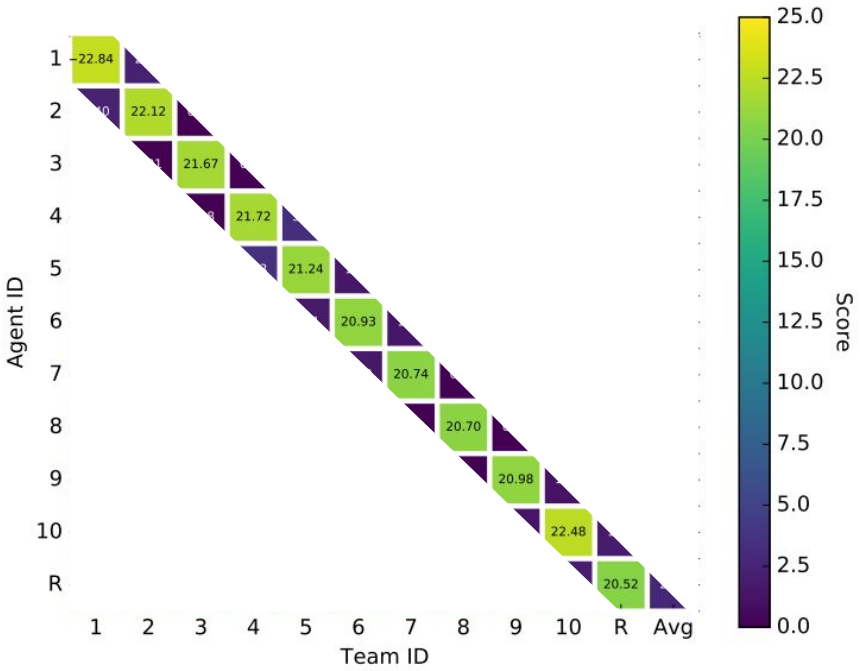
Previous work in self-play

- **Bayesian Action Decoder for Deep Multi-Agent Reinforcement Learning (BAD)**, (Foerster et al ICML 2019)
- **Improving Policies via Search in Cooperative Partially Observable Games**, (Lerer et al AAAI 2020)
- **Simplified Action Decoder for Deep Multi-Agent Reinforcement Learning (SAD)**, (Hu et al ICLR 2020)
- **Learned Belief Search: Efficiently Improving Policies in Partially Observable Settings**, (Hu et al AAAI 2021)
- ...



Ad-hoc/Zero-shot coordination challenge





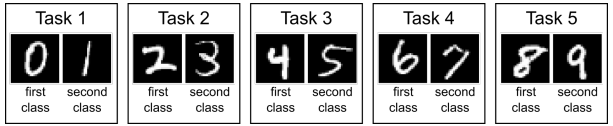
Talk Outline

- The Hanabi Challenge:
  - Self-play for Hanabi
  - The difficulty of the ad-hoc challenge (*Motivation 1*)
- Current Lifelong Learning benchmarks (*Motivation 2*)
- Lifelong Learning for MARL and MARL for Lifelong Learning
- Summary

Current Lifelong Learning benchmarks

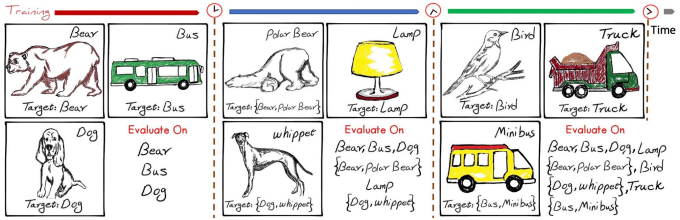
Supervised learning:

Rotated/Permuted/Split MNIST
Goodfellow et al., 2013



Core50
Zenke et al., 2017

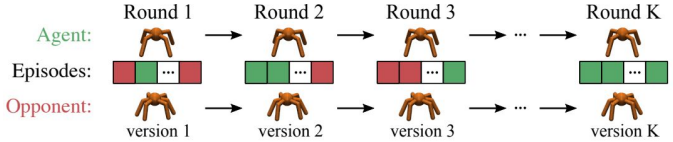
CRIB
Chaudhry et al., 2018b



IIRC
Abdelsalam et al. 2021

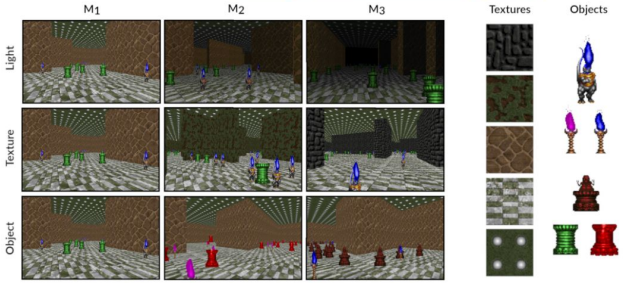
Reinforcement learning:

Robosumo
Al-Shedivat et al. 2017



Coinrun
Cobbe et al., 2019

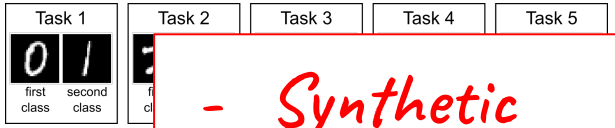
CRLMaze
Lomonaco et al. 2020



Current Lifelong Learning benchmarks

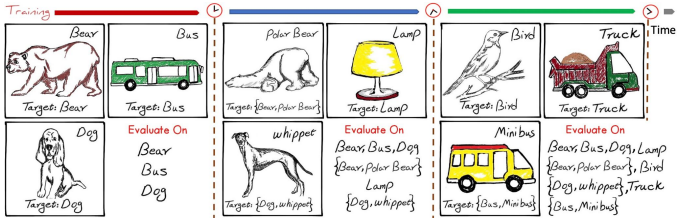
Supervised learning:

Rotated/Permuted/Split MNIST
Goodfellow et al., 2013



- Synthetic

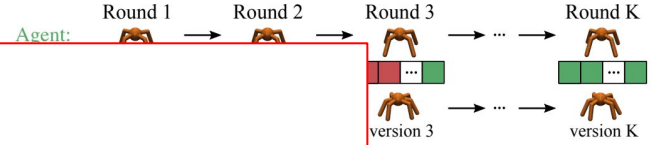
CRIB
Chaudhry et al., 2018b



IIRC
Abdelsalam et al. 2021

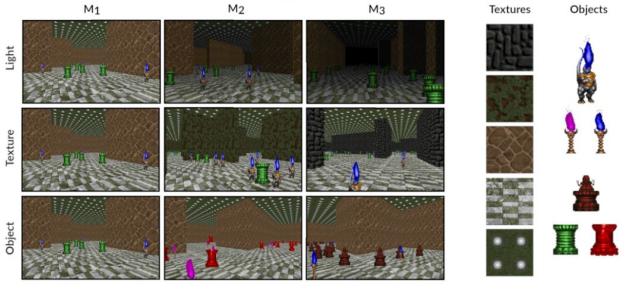
Reinforcement learning:

Robosumo
Al-Shedivat et al. 2017



Coinrun
Cobbe et al., 2019

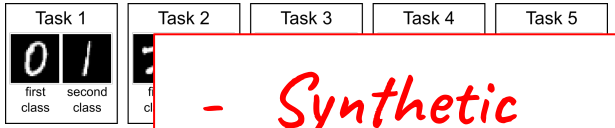
CRLMaze
Lomonaco et al. 2020



Current Lifelong Learning benchmarks

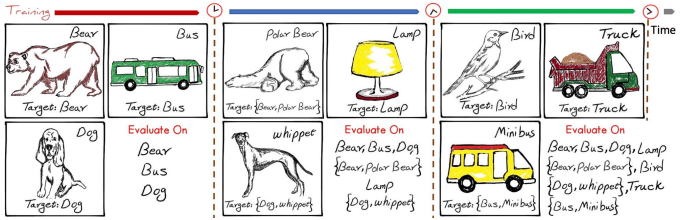
Supervised learning:

Rotated/Permuted/Split MNIST
Goodfellow et al., 2013



- Synthetic
- Lack a similarity metric

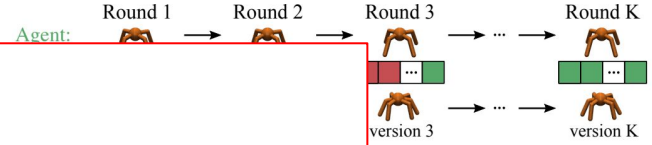
CRIB
Chaudhry et al., 2018b



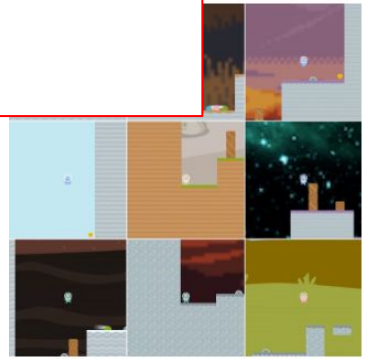
IIRC
Abdelsalam et al. 2021

Reinforcement learning:

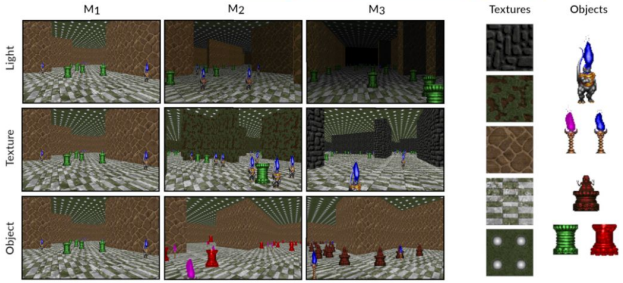
Robosumo
Al-Shedivat et al. 2017



Cobbe et al., 2019



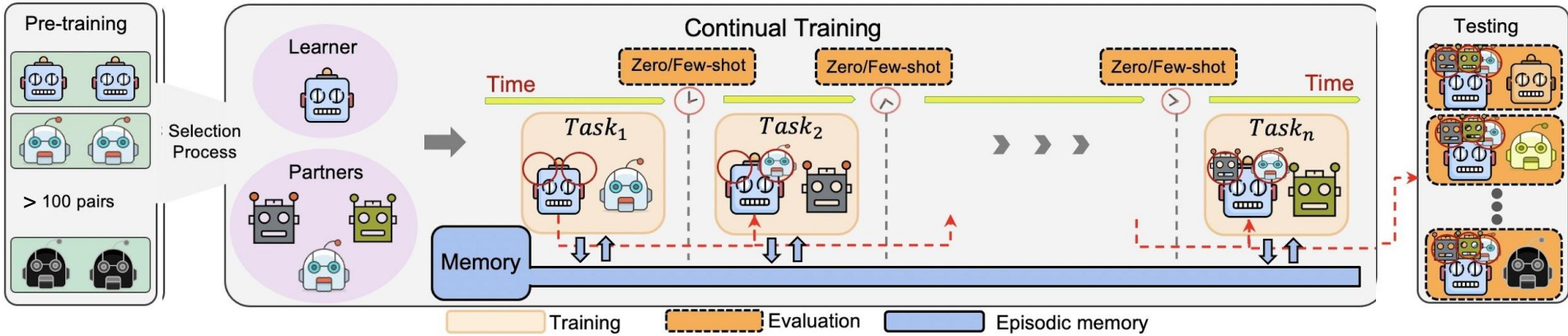
CRLMaze
Lomonaco et al. 2020



Talk Outline

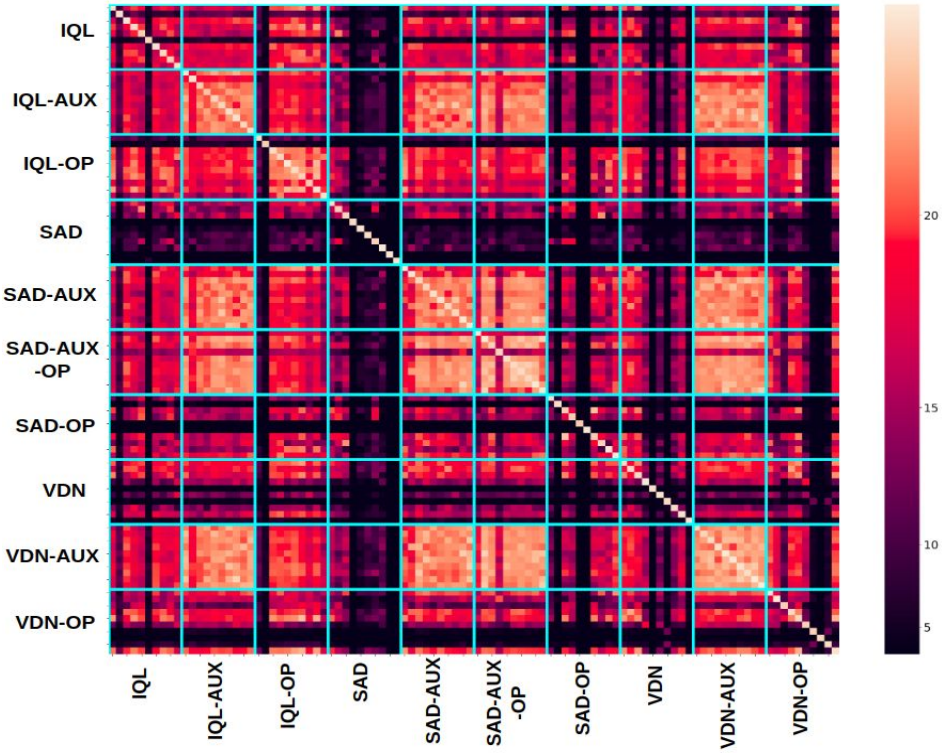
- The Hanabi Challenge:
  - Self-play for Hanabi
  - The difficulty of the ad-hoc challenge (*Motivation 1*)
- Current Lifelong Learning benchmarks (*Motivation 2*)
- Lifelong Learning for MARL and MARL for Lifelong Learning
- Summary

Lifelong-Hanabi setup



Two problems, one solution!

- Diverse set of strategies
  - 10 different MARL methods.
  - 5 different architectures with 2 seeds of each.
  - Easily extendable!

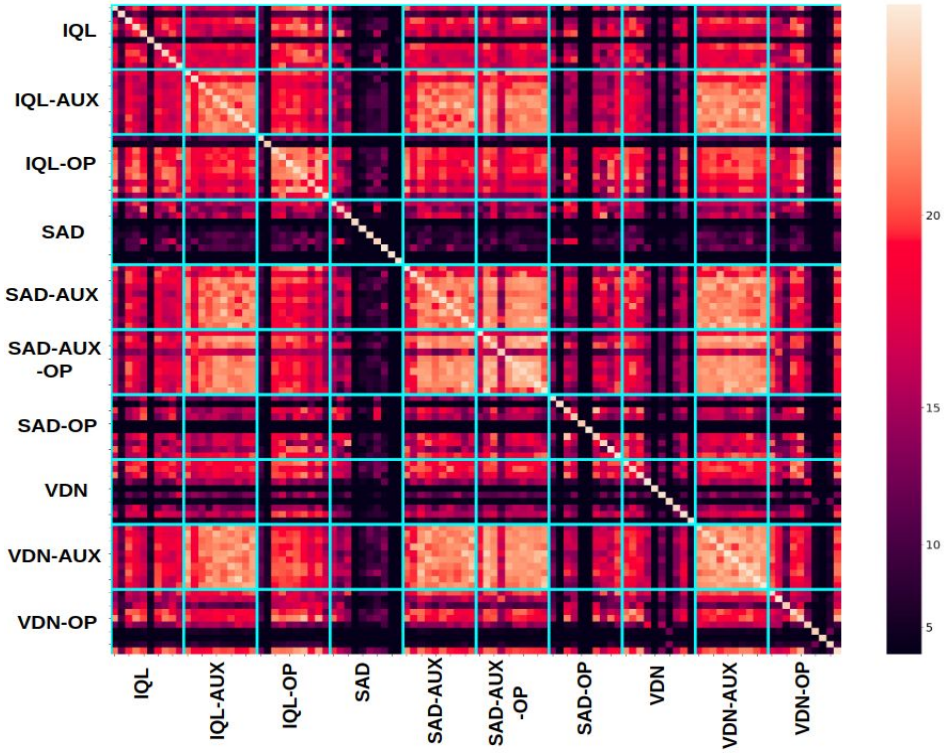


IQL: Tan et al., 1993.
VDN: Sunehag et al., 2017.
SAD: Hu and Foerster, 2019.
OP: Hu et al., 2020.

Two problems, one solution!

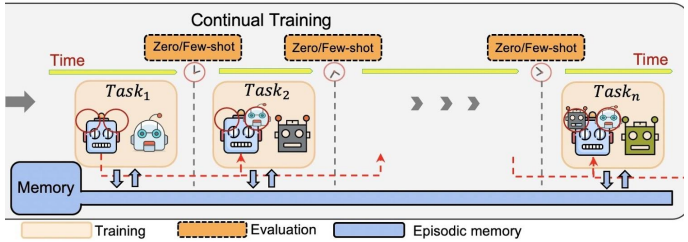
- Diverse set of strategies
  - 10 different MARL methods.
  - 5 different architectures with 2 seeds of each.
  - Easily extendable!

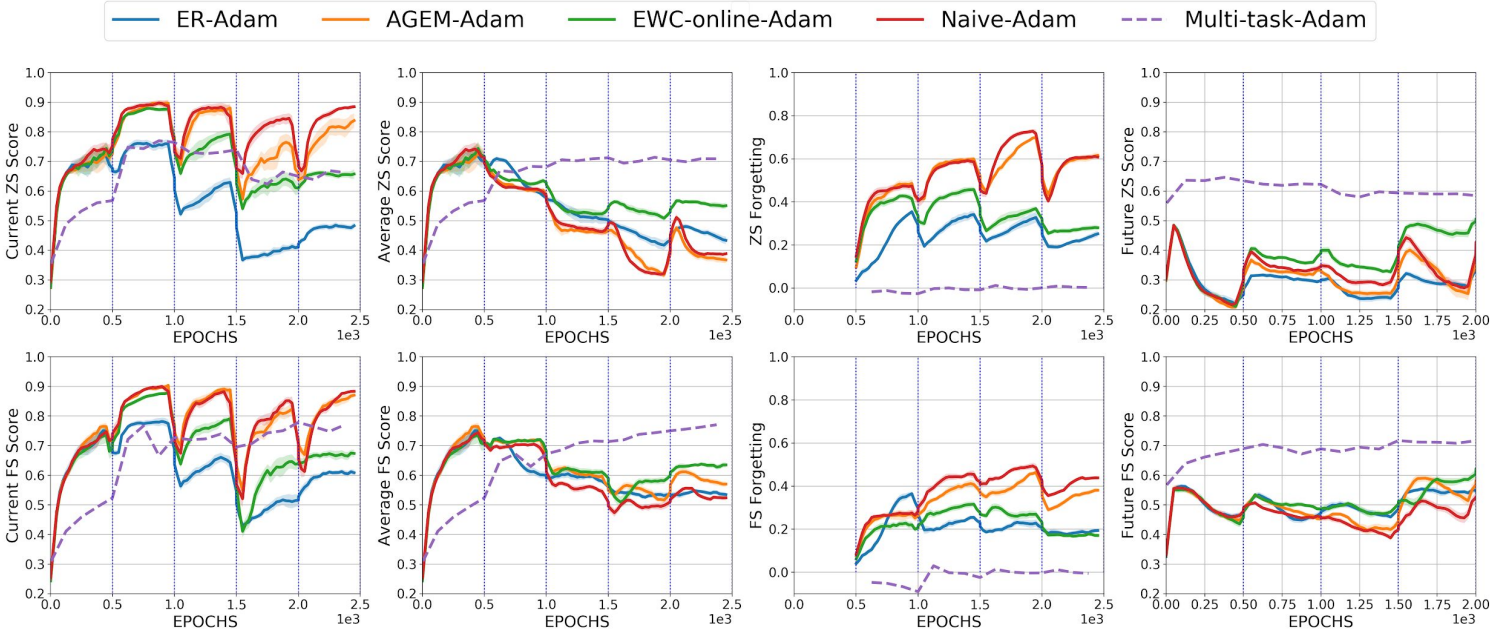
- Cross-play matrix as a proxy for how similar the tasks are.



IQL: Tan et al., 1993.
VDN: Sunehag et al., 2017.
SAD: Hu and Foerster, 2019.
OP: Hu et al., 2020.

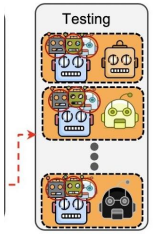
Lifelong learning benchmarks





ER: Chaudhry et al., 2019.
A-GEM: Chaudhry et al., 2018.
EWC: Kirkpatrick et al., 2017 ; Schwarz et al., 2018.

Zero-Shot Coordination



Evaluated with other agents trained with same method

Evaluated with other agents trained with different methods

	Training Method	SP	Intra-CP	Inter-CP	Limitations
	SAD	23.85 ± 0.03	7.70 ± 0.69	14.60 ± 0.24	C + GA
	SAD + AUX	23.57 ± 0.03	20.97 ± 0.80	18.51 ± 0.23	C + GA + L
	SAD + OP	24.14 ± 0.03	10.10 ± 0.87	16.09 ± 0.25	C + Sym + GA
	SAD + AUX + OP	23.40 ± 0.04	21.23 ± 0.25	17.77 ± 0.23	C + Sym + L + GA
Ours	IQL + ER	20.91 ± 0.05 ($\downarrow$ 2.98)	15.73 ± 0.39 ($\uparrow$ **7.06**)	16.32 ± 0.21 ($\uparrow$ **8.09**)	P
Ours	IQL + AUX + ER	22.34 ± 0.06 ($\downarrow$ 1.46)	20.90 ± 0.06 ($\downarrow$ 0.15)	$\mathbf{19.17 \pm 0.22}$ ($\uparrow$ **1.33**)	L + P
Ours	IQL + Multi-task	20.93 ± 0.09 ($\downarrow$ 2.96)	16.05 ± 0.30 ($\uparrow$ **7.38**)	$\mathbf{17.88 \pm 0.17}$ ($\uparrow$ **9.65**)	UP

Talk Outline

- The Hanabi Challenge:
  - Self-play for Hanabi
  - The difficulty of the ad-hoc challenge (*Motivation 1*)
- Current Lifelong Learning benchmarks (*Motivation 2*)
- Lifelong Learning for MARL and MARL for Lifelong Learning
- Summary

Summary

- Lifelong Learning to improve zero-shot coordination
- MARL for designing a Lifelong Learning benchmark

Next steps:

- Evaluate our agents with human partners
- Applying Few-shot adaptation methods (MAML, ...)

Thank you!

Code and all pre-trained models:
https://github.com/chandar-lab/Lifelong-Hanabi