Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Online Experience Sharing in Deep Reinforcement Learning

Author: Federico Bottoni
806944

Supervisor: Giuseppe Vizzari

Co-supervisor: Ivana Dusparic

Online Experience Sharing

CONTEXT

- Reinforcement Learning (RL)
  - Deep Q-Learning
- Transfer Learning (within RL)
  - No previous experience
  - Online

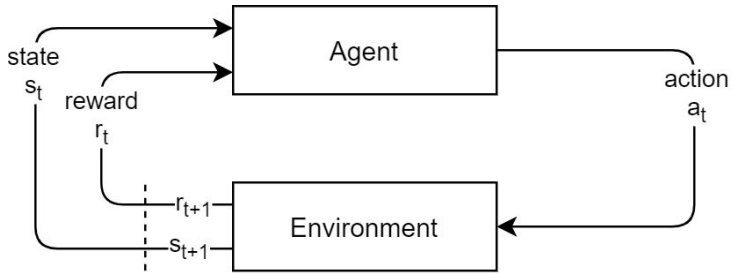
OUTLINE

1. Context introduction
2. Algorithm design
3. Evaluation
4. Discussion
5. Future works

Reinforcement Learning

- “Trial and Error” approach to identify the best agent **policy** to solve a task
- Related process is a Markov model
- Policy function can be approximated by Artificial Neural Networks
- Replay Buffer

- Widely used in game theory, control, simulations and other fields



Transfer Learning (within Reinforcement Learning)

Generalize across data drawn from multiple distributions or domains, potentially reducing the amount of data required to learn a new task

PURPOSE

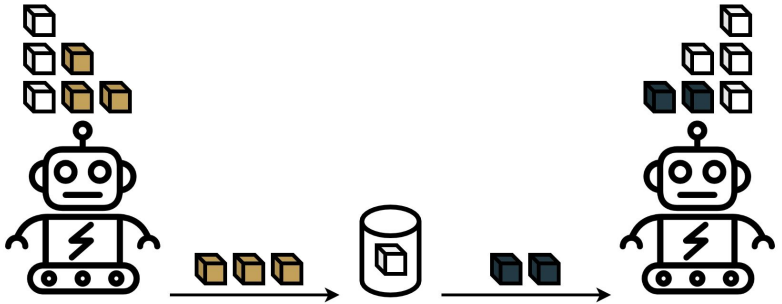
Improve learning agents' performance and learning speed exchanging data among them

INSTANCES

Inter-Task Mapping, Reusing Representations, Learning from demonstrations, Policy Transfer

Contribution

- Design online transfer systems and agents
- Enable transfers defining:
  - Transferred object
  - Transfer frequency and data size
  - Data selection methods
- Design confidence evaluation systems for states and state-action pairs
- Introduced hyperparameters tuning



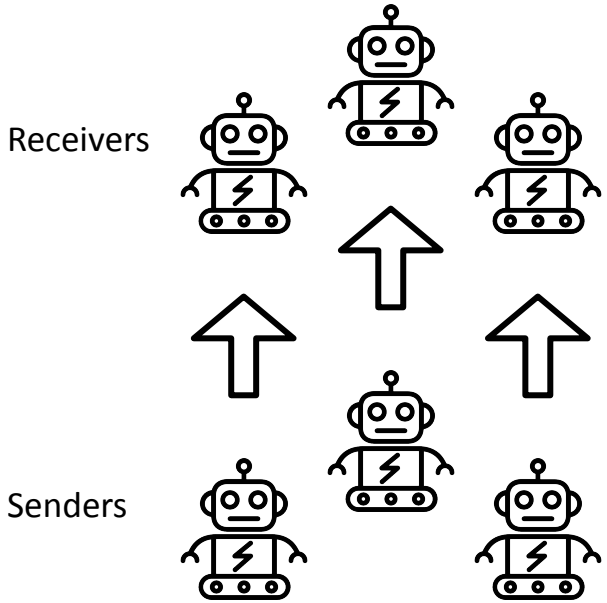
Online Experience Sharing

Parallel DeepRL Agents

SYSTEM

- Deep Q-Learning independent agents
- Agent's roles: *sender* or *receiver*
- System: two agents in 1-to-1 relationship

- Framework is flexible on mechanisms and methods



Enabling Transfer: Two Steps Selection

- Transferred object is the experience represented by transitions: **<s, a, r, s'>**
- Store transferred transitions: **Transfer buffer (Tb)**
- Senders select data to transfer and send them to receiver's Tb
- Receivers select a subset of Tb to fill partially their DQN batch

- Two selection algorithms handle the transfers of the two sides:
  - **Sender Selection Method (SSM)**
  - **Receiver Selection Method (RSM)**

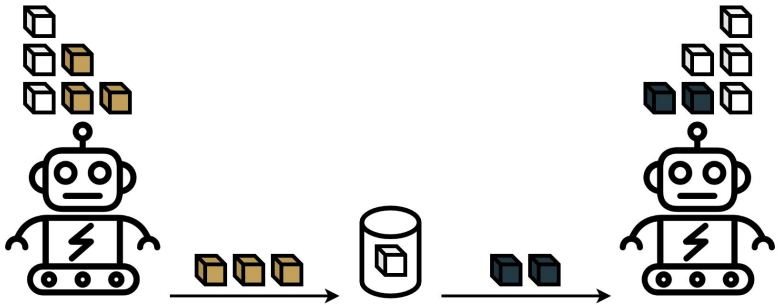
Transfer Frequencies and Transfer Sizes

FREQUENCIES

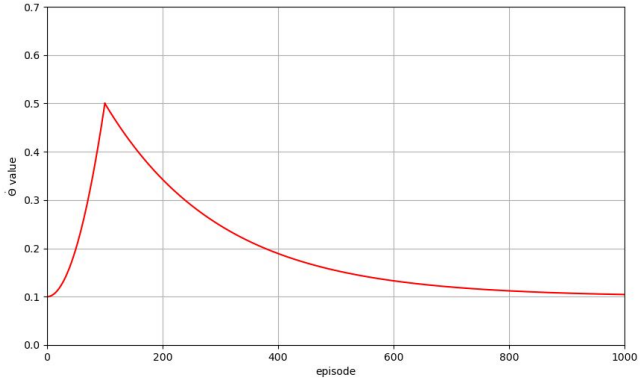
1. **Sender**: periodically, every **TransferInterval (TI)** steps, ($TI \in Z^+$)
2. **Receiver**: every single step

SIZES

1. **Sender**: constant **TransferSize (TS)** amount of transitions, ($TS \in Z^+$)
2. **Receiver**: **BatchSize** * **TransferBatchSize** which is result of θ *function* defined later



θ-function



GOAL

- Transfer little data at the very beginning of the learning process
- Encourage transfer during the final part of exploration phase
- Reduce transfer when policies are consolidated

PARAMETERS

- θ_{max}: global max value
- θ_{min}: global min value
- θ_{decay}: decay speed
- θ_{apex}: conjunction of the two functions

Result: **Transfer Batch Size** partition

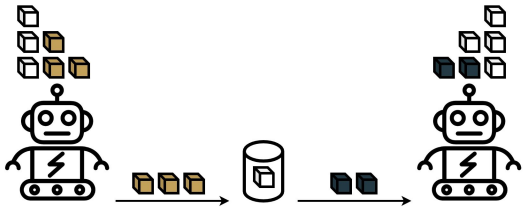
Agents model: “State Visit Table” (**VT**)

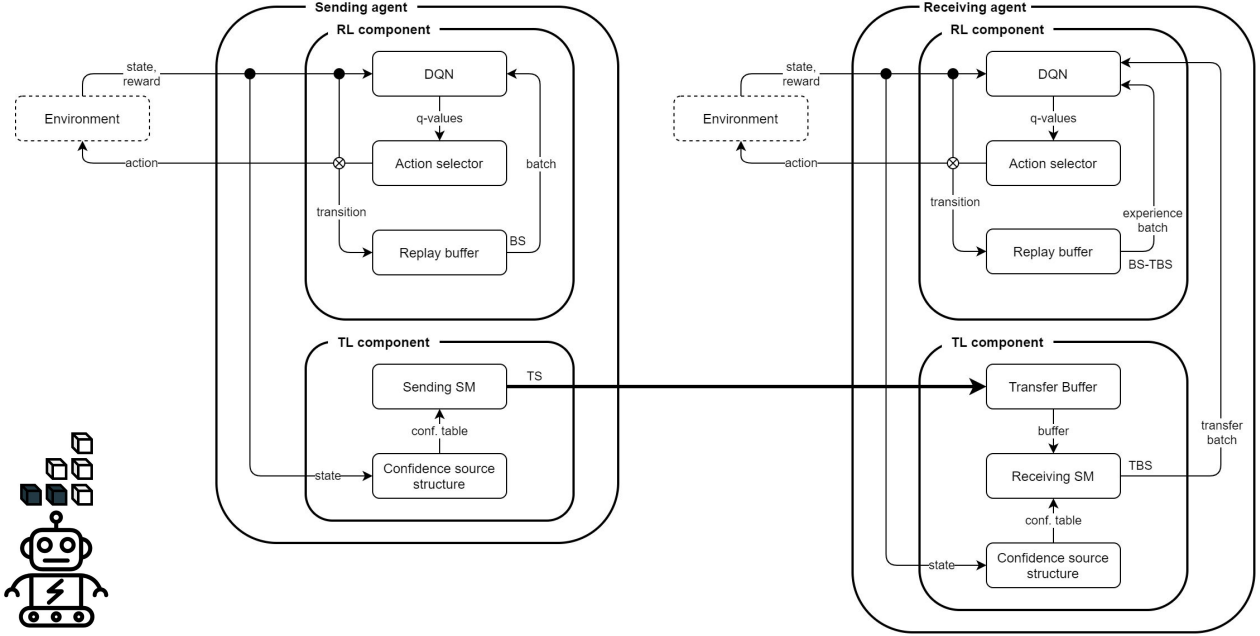
Sender SM

➔ Most Visited (**Mv**)

Receiver SM

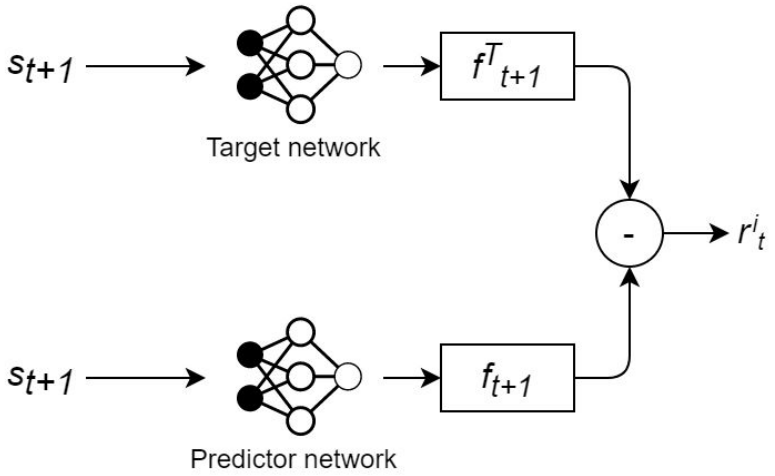
➔ Least Visited (**Lv**)





Random Network Distillation

- Curiosity mechanism
  - Target network
  - Predictor network
- Uncertainty as Mean Squared Error
- Hyperparameter: *encoding_size*



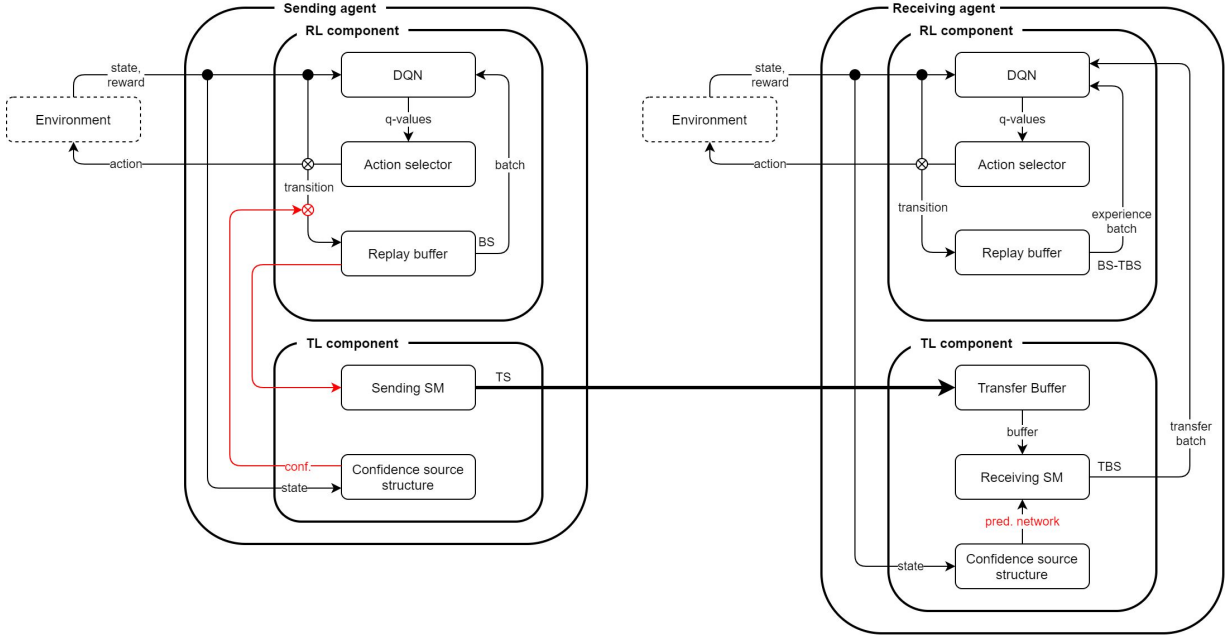
Agents model: “State RND” (**SRND**)

Sender SM

➔ Lowest Uncertainty (**Lu**)

Receiver SM

➔ Highest Uncertainty Difference (**Hud**)



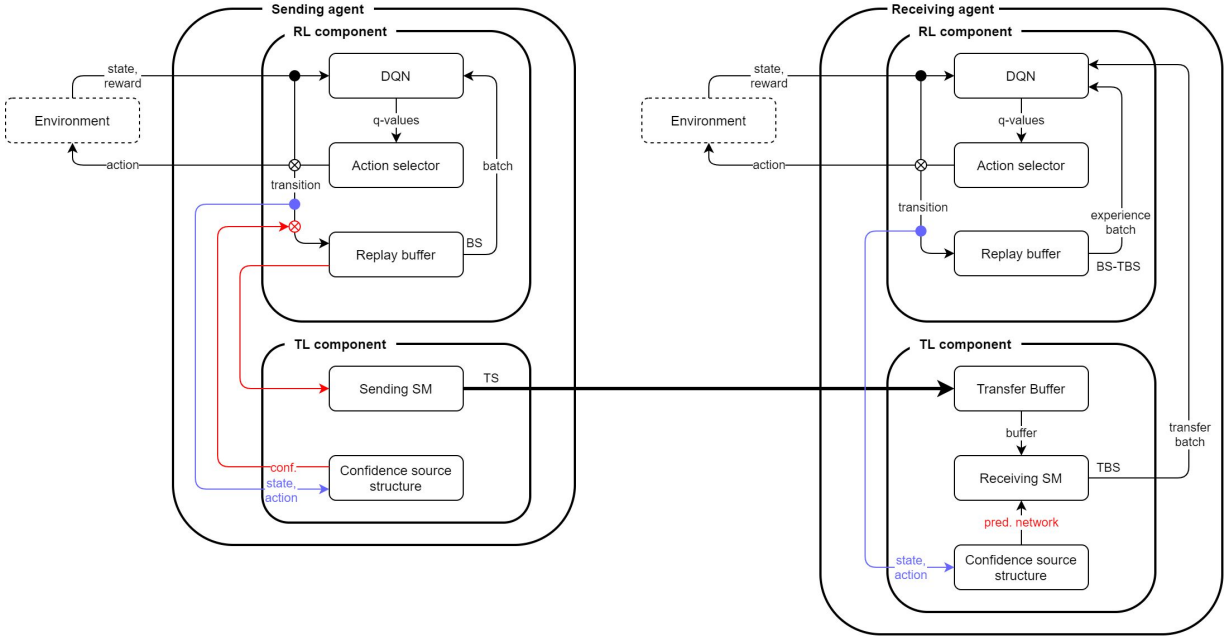
Agents model: “State-Action RND” (**QRND**)

Sender SM

➔ Lowest Uncertainty (**Lu**)

Receiver SM

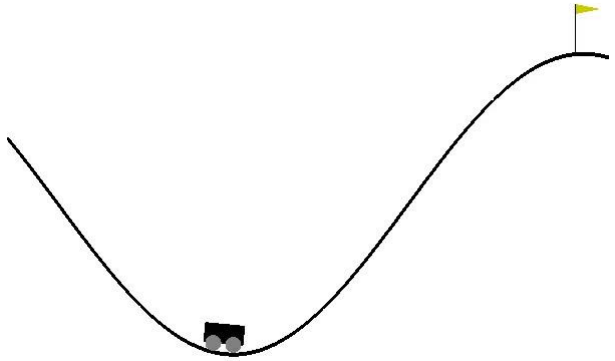
➔ Highest Uncertainty Difference (**Hud**)



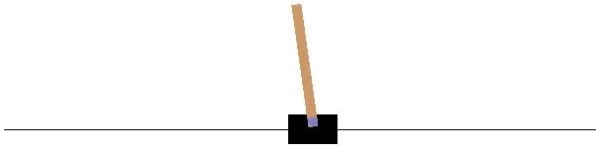
Evaluation

Experimental Settings

Mountain Car



Cart Pole



Benchmark single-agent systems with different desired task

1. Mountain Car (1000 episodes)
2. Cart Pole (150 episodes)

Experiments

Baselines (**BL**):

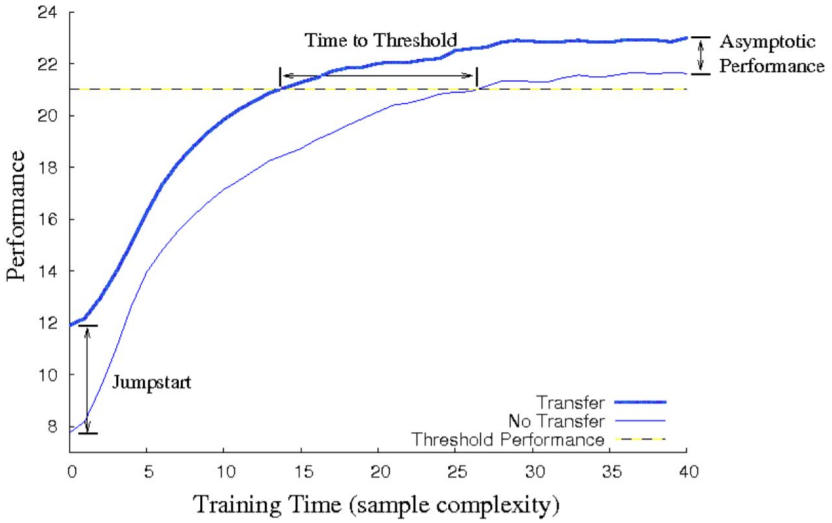
1. No Transfer (**NT**): regular DQL process
2. Random Transfer (**RT**):
   a. senders transfer the most recent experience
   b. receivers select randomly from Tb
3. Top Transfer (**TT**): agents share the same brain

Experiments for each scenario:

1. senders apply SSM,
   receivers select randomly from Tb (**R**)
2. senders select the most recent (**L**),
   receivers apply the RSM
3. sender apply SSM,
   receivers apply RSM

Evaluation Metrics

- Asymptotic performance
- Learning speed
  - Total Reward
  - (Time to Threshold)
- Jumpstart
- Learning stability



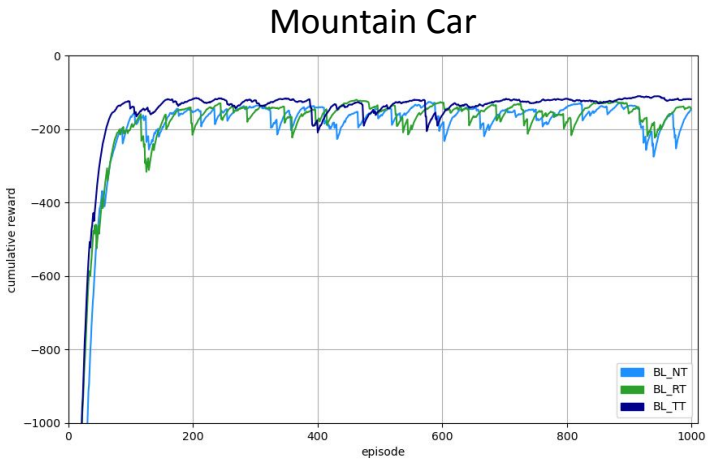
Baselines

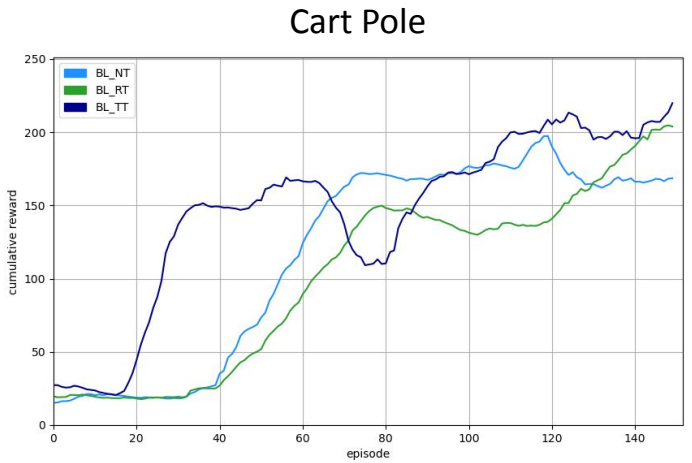
Mountain Car

- RT is faster the NT but slows down earlier, then is similar to NT

Cart Pole

- RT is slow and increases slowly





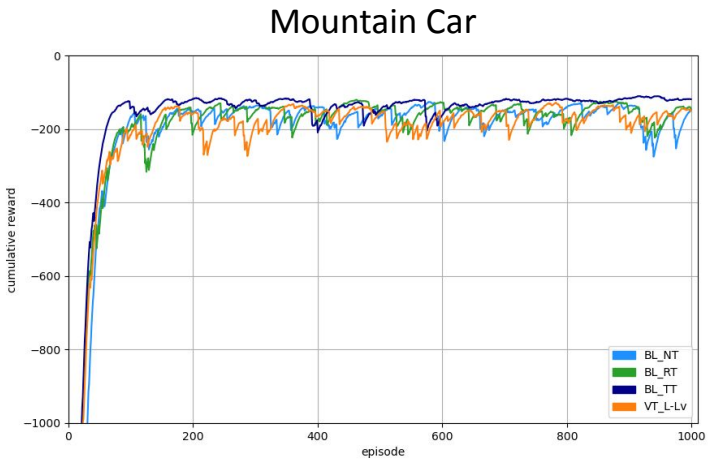
Visit Table (VT)

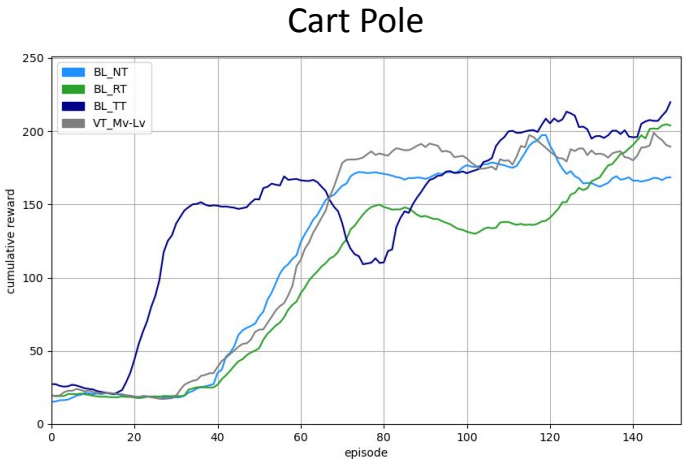
Mountain Car

- Does not bring significant improvements

Cart Pole

- Important **asymptotic improvement** of about 25 reward units





State-RND (SRND)

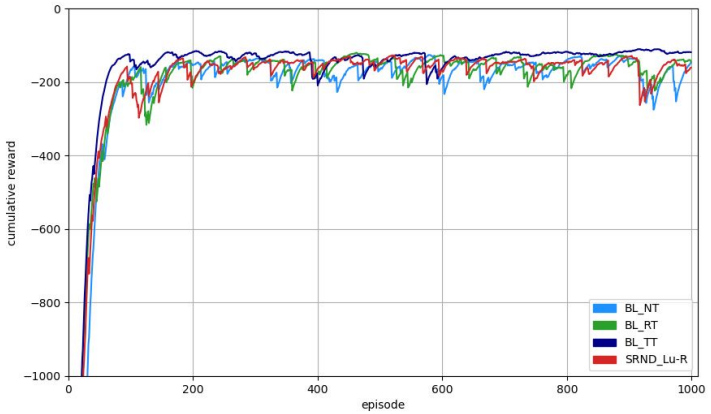
Mountain Car

- Anticipates baselines (**speed-up**) and in general brings **stability improvements**

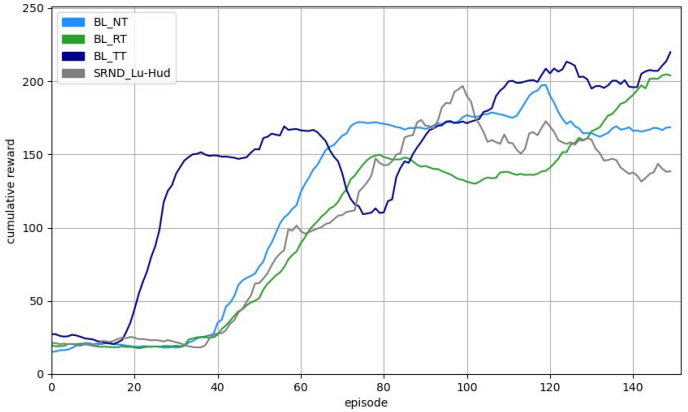
Cart Pole

- Reaches high reward before every baseline
- Does not bring significant improvements

Mountain Car



Cart Pole



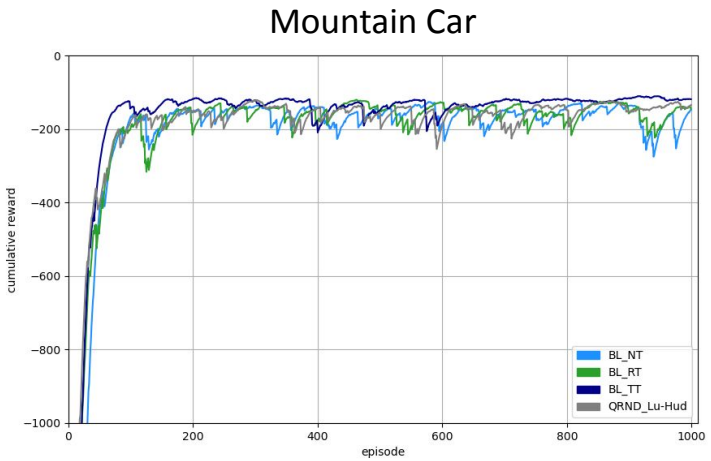
State-Action RND (QRND)

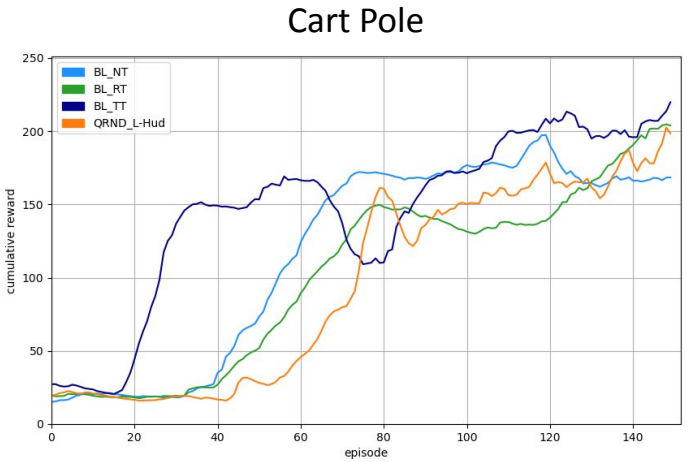
Mountain Car

- Interesting **stability improvement, speed-up**
- Particularly stable in last 100 episodes

Cart Pole

- Does not bring significant improvements
- Multiple similarities to RT baseline





Total reward improvement analysis

Mountain Car

Scenario	SenderSM	ReceiverSM	% earned
VT	MostVisited	Random	-0,38
VT	Lasts	LastVisited	0,03
VT	MostVisited	LastVisited	-0,26
SRND	LowestUncertainty	Random	0,27
SRND	Lasts	HighestUncertaintyDiff	0,27
SRND	LowestUncertainty	HighestUncertaintyDiff	-0,1
QRND	LowestUncertainty	Random	-0,14
QRND	Lasts	HighestUncertaintyDiff	0,29
QRND	LowestUncertainty	HighestUncertaintyDiff	**0,49**

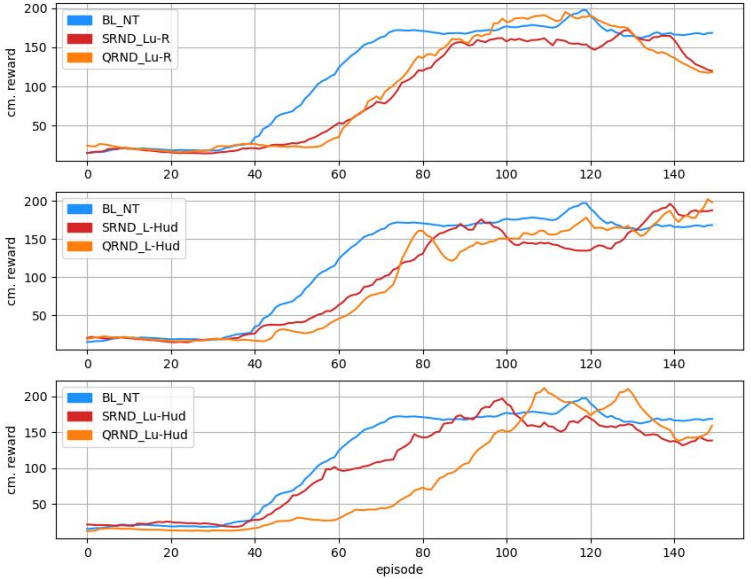
Cart Pole

Scenario	SenderSM	ReceiverSM	% earned
VT	MostVisited	Random	-0,02
VT	Lasts	LastVisited	0,03
VT	MostVisited	LastVisited	**0,22**
SRND	LowestUncertainty	Random	-0,96
SRND	Lasts	HighestUncertaintyDiff	-0,6
SRND	LowestUncertainty	HighestUncertaintyDiff	-0,53
QRND	LowestUncertainty	Random	-0,77
QRND	Lasts	HighestUncertaintyDiff	-0,66
QRND	LowestUncertainty	HighestUncertaintyDiff	-1,03

SRND - QRND Cart Pole comparison

RND scenarios compared

- Confirm Cart Pole's poor speed
- Demonstrate how considering action in confidence evaluation pushes the agent to better policies



Transfer Parameters Tuning

TRANSFER PARAMETERS TUNING

[Bayesian optimization]

- Transfer Size
- Transfer Interval
- Transfer Buffer Size
- θ_{decay}
- θ_{apex}

VT TUNING

[Bayesian optimization]

- VT Dimension Discretizer

SRND - QRND TUNING

[Manual sampling]

- RND Encoding Size

Conclusions

Discussion

- Performances strictly dependent on environment and task to learn
  - Mountain Car is improved by SRND and QRND scenarios in **learning speed** and **stability**
  - Cart Pole is improved by VT scenario in particular in **asymptotic performance**
- High variability due to neural networks usage
- Considering state-action pairs instead of state to evaluate transition confidence can provide better outcome

Future Works

- Experiments in multi-agents systems to study agents' interactions too
- Analysis of the effect of multi-sender and multi-receivers transfers
- Analysis of a online role-swap mechanism

- Change of transfer object, eg. sub trajectories of states, neural network data like weights
- Study the transfer clear effects letting the receiving agent chose only through transfers after a first period of state-space self-exploration