When Do Transformers Shine in RL? Decoupling Memory from Credit Assignment

Tianwei Ni, Michel Ma, Benjamin Eysenbach, Pierre-Luc Bacon

Nov 12, 2023

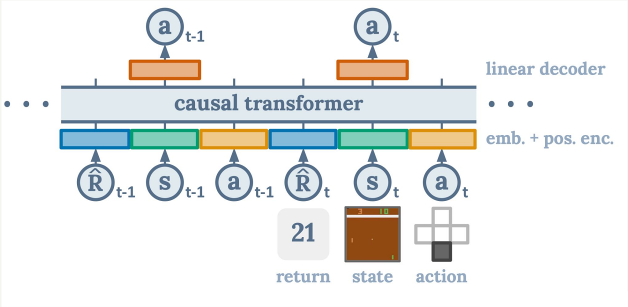
Accepted at NeurIPS 2023 (oral)

Paper is available at arXiv and code is open-sourced

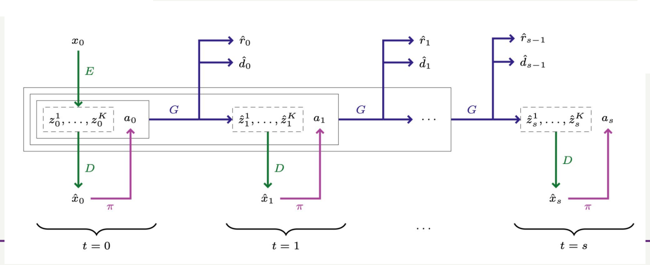
Mystery

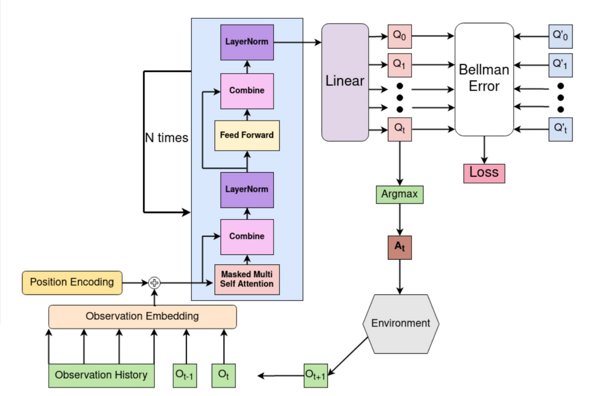
Transformers have been successful in **RL**

- **Offline RL:** Decision transformer (Chen et al., 2021)
- **Model-Based RL:** IRIS (Micheli et al., 2022)
- **Model-Free RL:** Deep Transformer DQN (Eslinger et al., 2022)



Why?





Why do Transformers shine in **SL**?
Excel at long-term dependencies

Attention mechanisms have become an integral part of compelling sequence modeling and transduction models in various tasks, allowing modeling of dependencies without regard to their distance in the input or output sequences [2, 19]. In all but a few cases [27], however, such attention mechanisms are used in conjunction with a recurrent network.

In this work we propose the Transformer, a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output. The Transformer allows for significantly more parallelization and can reach a new state of the art in translation quality after being trained for as little as twelve hours on eight P100 GPUs.

From *Attention Is All You Need*

Temporal dependency: *memory*?

Long-range dependence

文A 1 language

Article Talk Read Edit source View history

From Wikipedia, the free encyclopedia

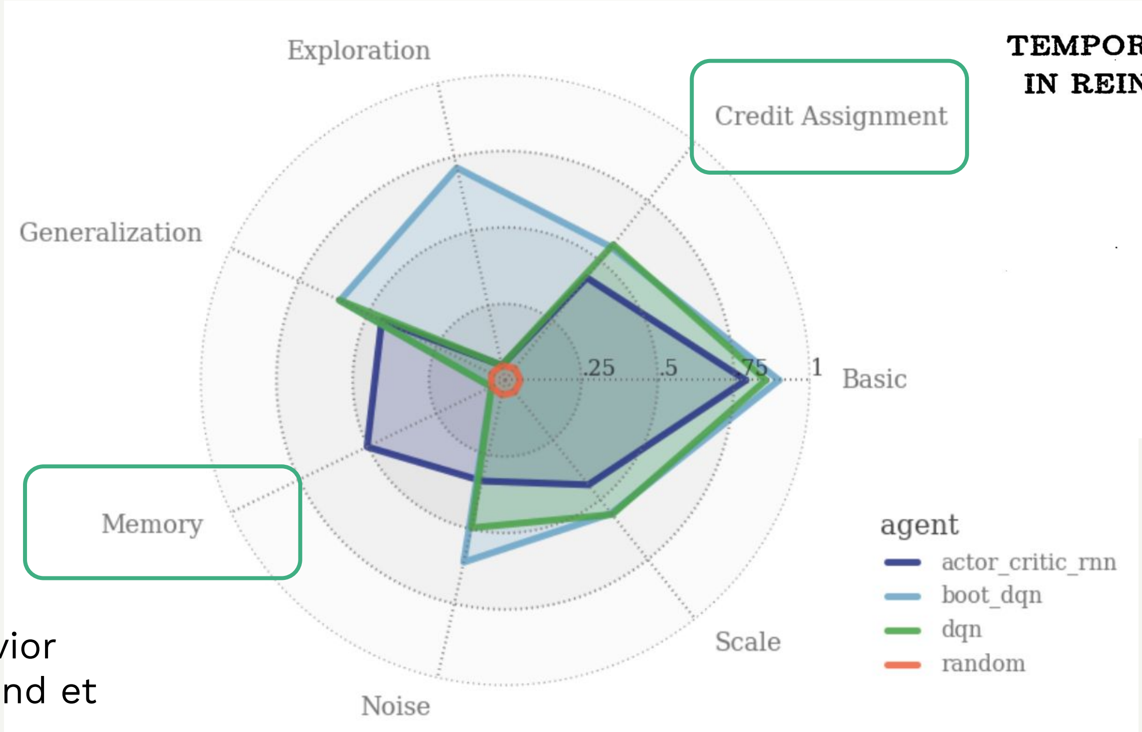
Long-range dependence (**LRD**), also called **long memory** or **long-range persistence**, is a phenomenon that may arise in the analysis of spatial or time series data. It relates to the rate of decay of statistical dependence of two points with increasing time interval or spatial distance between the points. A

The Problem of Learning Long-Term Dependencies in Recurrent Networks

Yoshua Bengio†, Paolo Frasconi‡, and Patrice Simard†
†AT&T Bell Laboratories
‡Dip. di Sistemi e Informatica, Universitá di Firenze

Supervised learning perspective

Capabilities in RL: memory and credit assignment



TEMPORAL CREDIT ASSIGNMENT IN REINFORCEMENT LEARNING

A Dissertation Presented

By

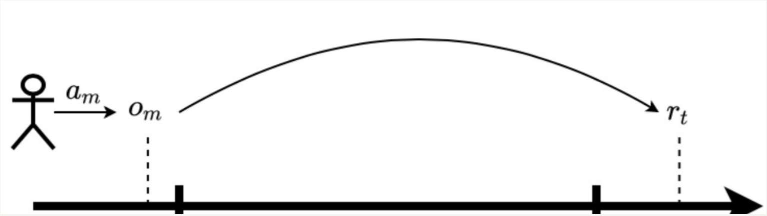
Richard S. Sutton

From behavior suite (Osband et al., 2019)

Memory and Credit Assignment: they are distinct!

Memory
The ability to recall distant past events

Temporal credit assignment
The ability to determine *when* the actions that deserved credit occurred (Sutton, 1984)



We have intuition!

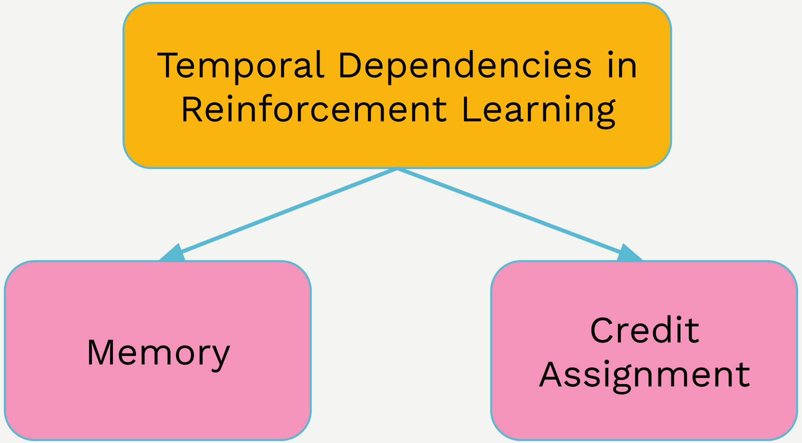
Scenario 1: Alice *remembers her passcode* set a month earlier, and opens a safe full of money.

Memory

Scenario 2: Bob *picks up a key* (then he can always see the key), and a month later he opens a safe full of money.

Credit Assignment

Why do Transformers shine in RL? Memory or Credit Assignment? Or Both?



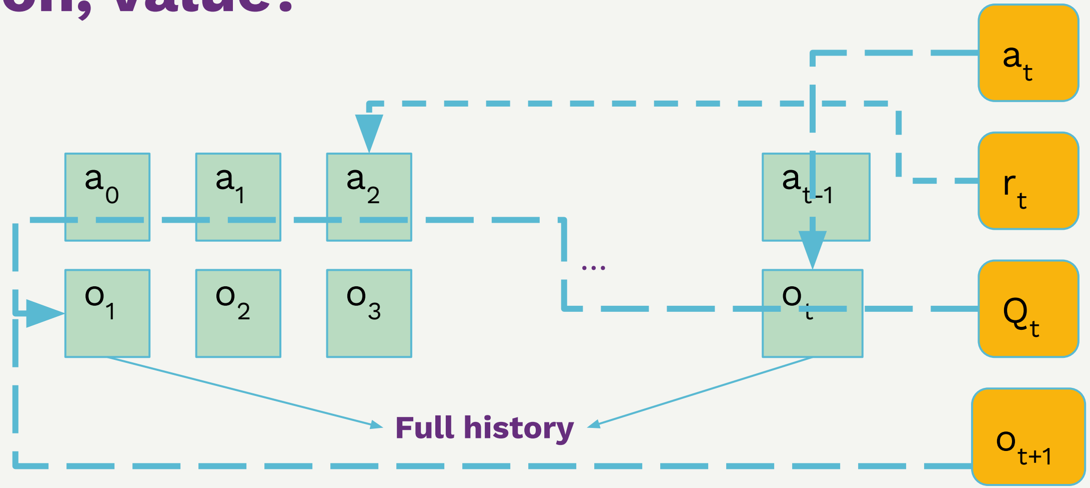
Although we have intuition, we don't have clear mathematical definitions.

This prevents us from understanding RL.

Measuring Temporal Dependencies

Memory lengths (intuition)

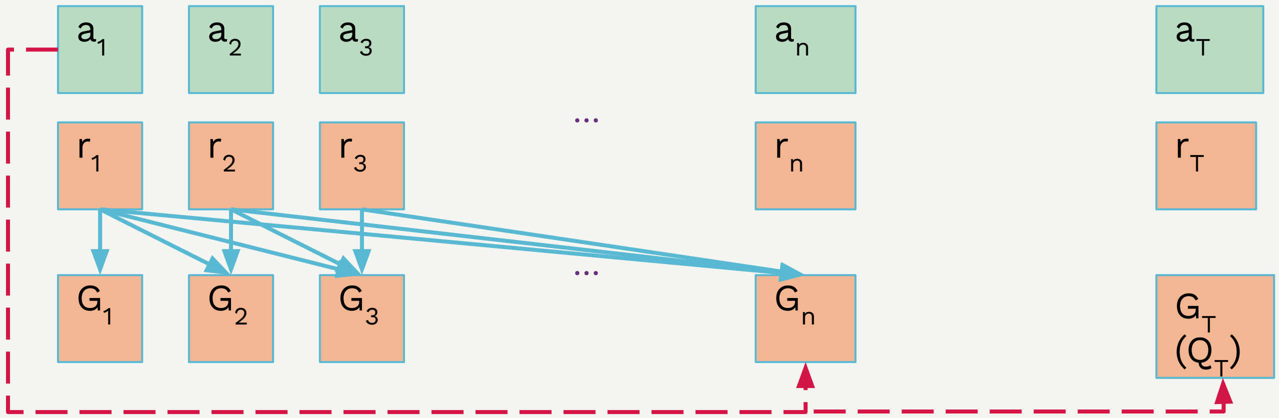
- History: all previous observations and actions
- **How long is the minimal history required to predict / generate current reward, observation, action, value?**



Mila

Credit assignment length (Intuition)

How long does it take for a greedy action to see its benefits regarding its *n-step rewards* (G_n)?



Mila

Examples: decoupling memory and credit assignment

Scenario 1: Alice remembers her passcode set a month earlier, and opens a safe full of money.

- Credit assignment length = 1 day
- Memory length = 1 month

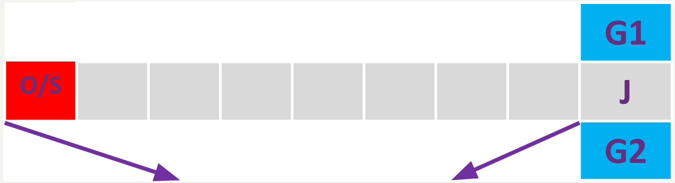
Scenario 2: Bob picks up a key (then he can always see the key), and a month later he opens a safe full of money.

- Credit assignment length = 1 month
- Memory length = 1 day

Mila

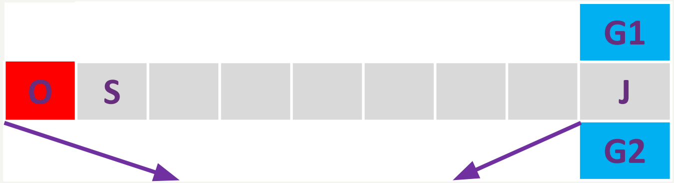
Proposing configurable toy tasks: Passive and Active T-Mazes

Passive T-Maze



Corridor Length T

Active T-Maze



Corridor Length T

Mem len: T
CA len: 1

O: Oracle (to get the info of G, randomized in each episode)
S: Start; J: Junction
G1, G2: Goal candidates (to get bonus)

Mem len: T
CA len: T

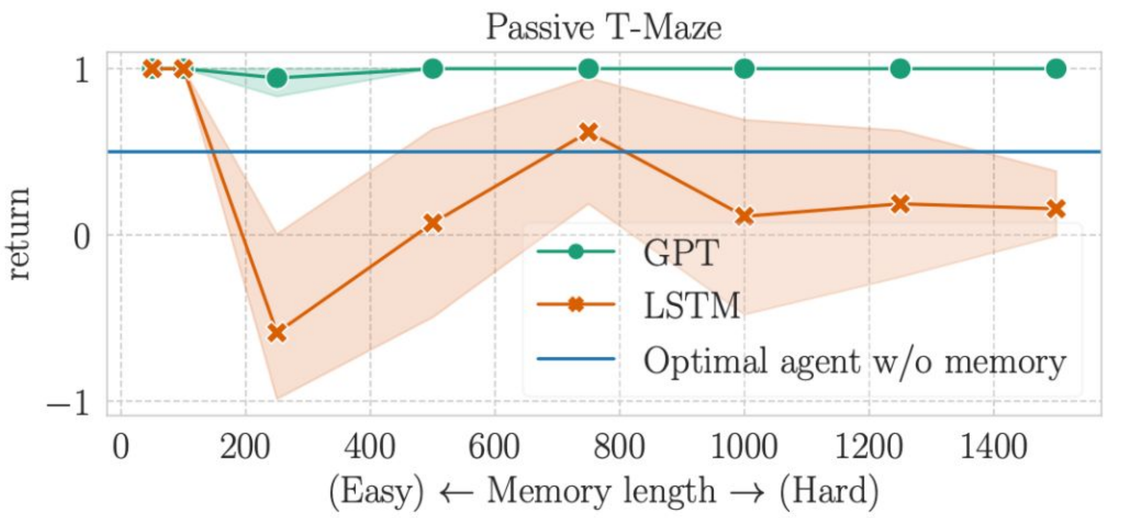
Going to O is only credited when the agent reaches G

Mila

Evaluating Transformer-based RL

Transformer-based RL excel at long-term memory

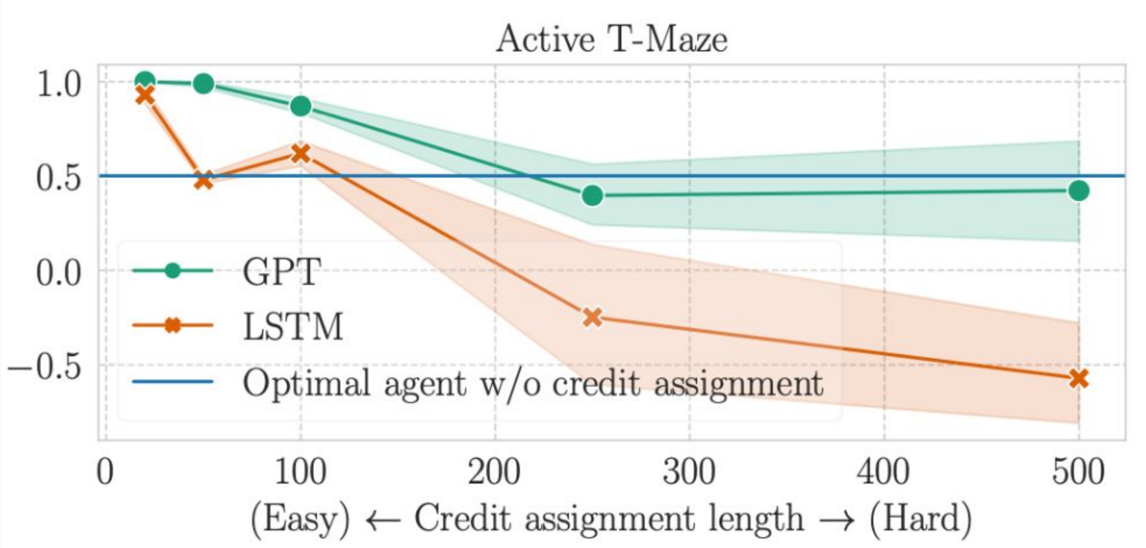
RL algorithm: DDQN w/ eps greedy
Sequence model: LSTM or Decoder-only Transformer (GPT-2)



Perfectly solved by GPT-2

Mila

Transformers cannot help long-term credit assignment in RL



In Active T-Maze, Transformers can reach the Junction.
Transformers helps, but degrades severely when the CA length >= 250

Future work

- Developing scalable RL systems for high-dim long-term memory tasks
- Searching sequence architectures or RL algorithms for long-term credit assignment

Thank you for watching!