Transformers are Meta-Reinforcement Learners

Luckeciano Melo

Microsoft, USA and
Center of Excellence in Artificial Intelligence, Brazil

Introduction

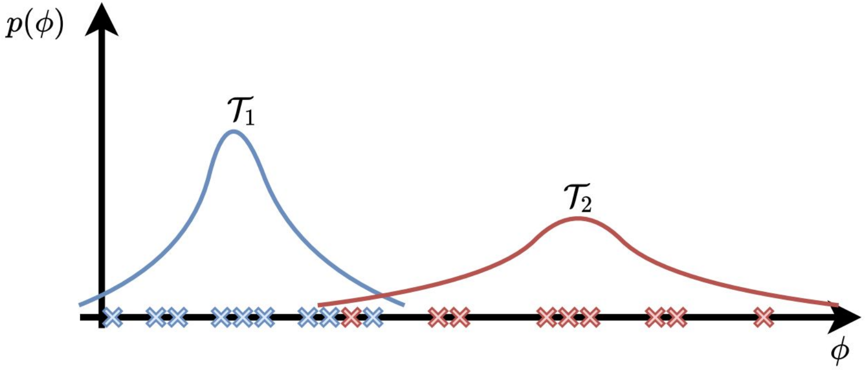
- Transformers as very powerful architectures for many Machine Learning tasks
  - Capability of handling long sequences
  - Presence of context-dependent weights from attention mechanism
- Hypothesis:
  - These two properties suit the central role of a meta-RL agent
    - Task Inference from a sequence of sampled episode trajectories
    - Policy Fast Adaptation Strategy → Self-Attention
- Proposal:
  - **Tr**ansformers for **M**eta-**R**einforcement **L**earning (TrMRL)

Transformers as a Memory System

- Task Representation: distribution over working memories:

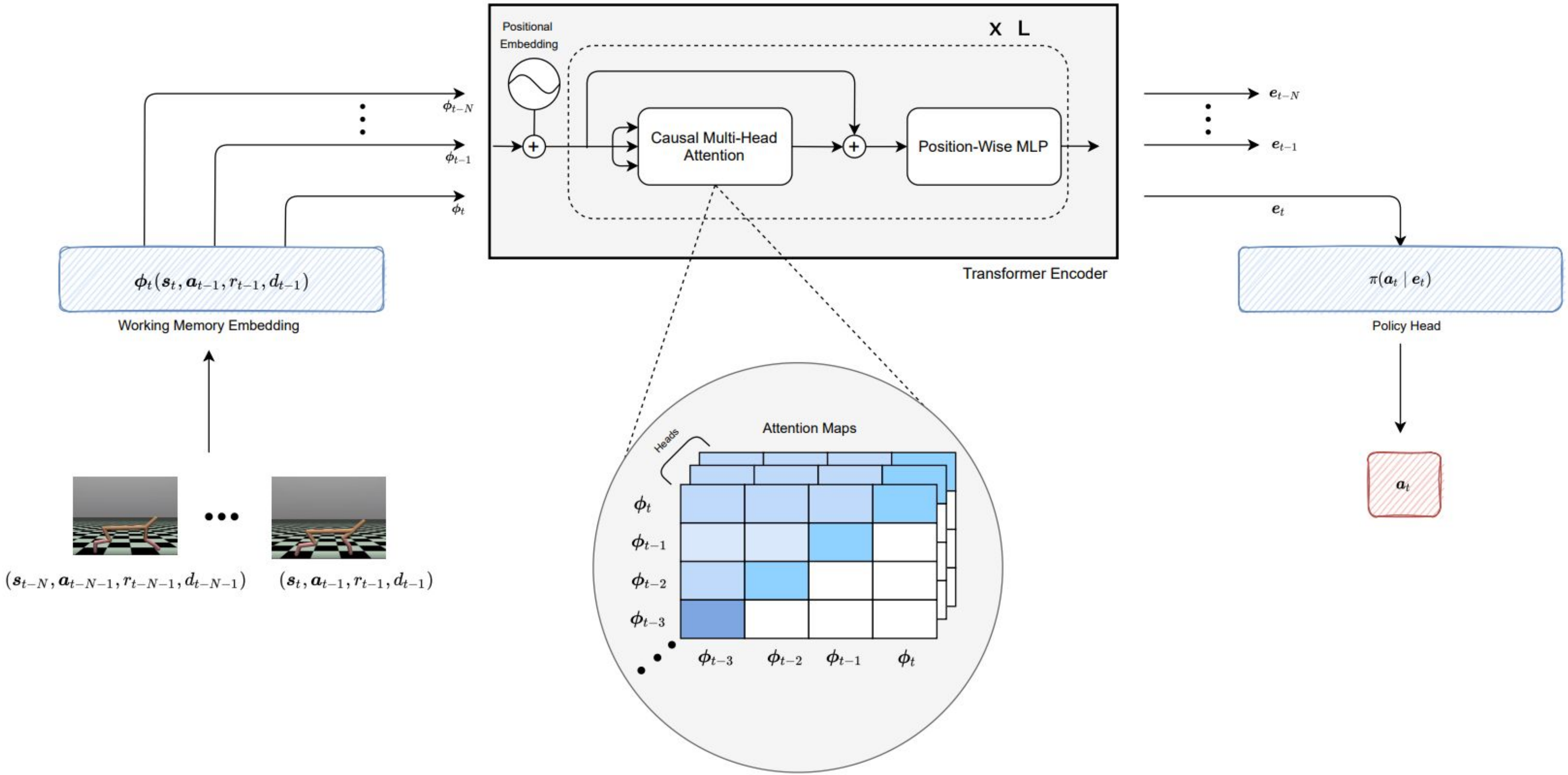
$$\mathcal{T}(\phi) : \Phi \to [0, \infty)$$

- Working Memory: parameterized function $\phi_t(\boldsymbol{s}_t, \boldsymbol{a}_t, r_t, \eta_t)$



- Starting from recent working memories, transformer implements **memory reinstatement for episodic memory retrieval**
  - A reminder procedure that reintroduces past elements in advance of a long-term retention test (Rovee-Collier, 2012)

TrMRL: Transformers for Meta-RL



Memory Reinstatement

- Transformer architecture recursively refines the episodic memory interacting memories retrieved from the past layer:

$$e_t^l = f(e_0^{l-1}, \ldots, e_t^{l-1})$$

- This refinement is guaranteed by a crucial property of the self-attention mechanism: it computes a **consensus representation across the input memories** associated to the sub-trajectory
  - Consensus representation is the memory representation that is closest on average to all likely representations (Kumar & Byrne, 2004), i.e., **minimizes the Bayes risk considering the set of memories**.

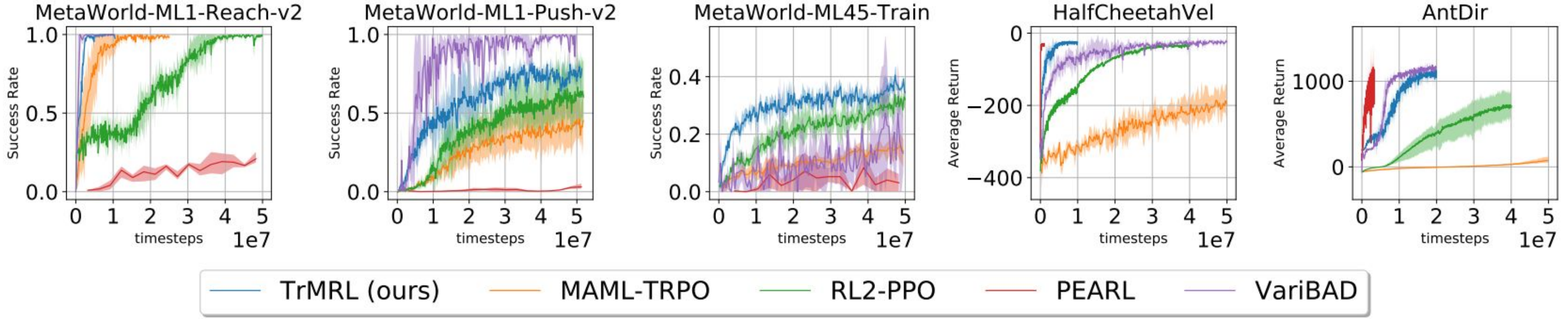
Stabilizing Transformer for Meta-RL

- Problem: Optimize transformers is often unstable
  - Especially in the context of **RL gradients**

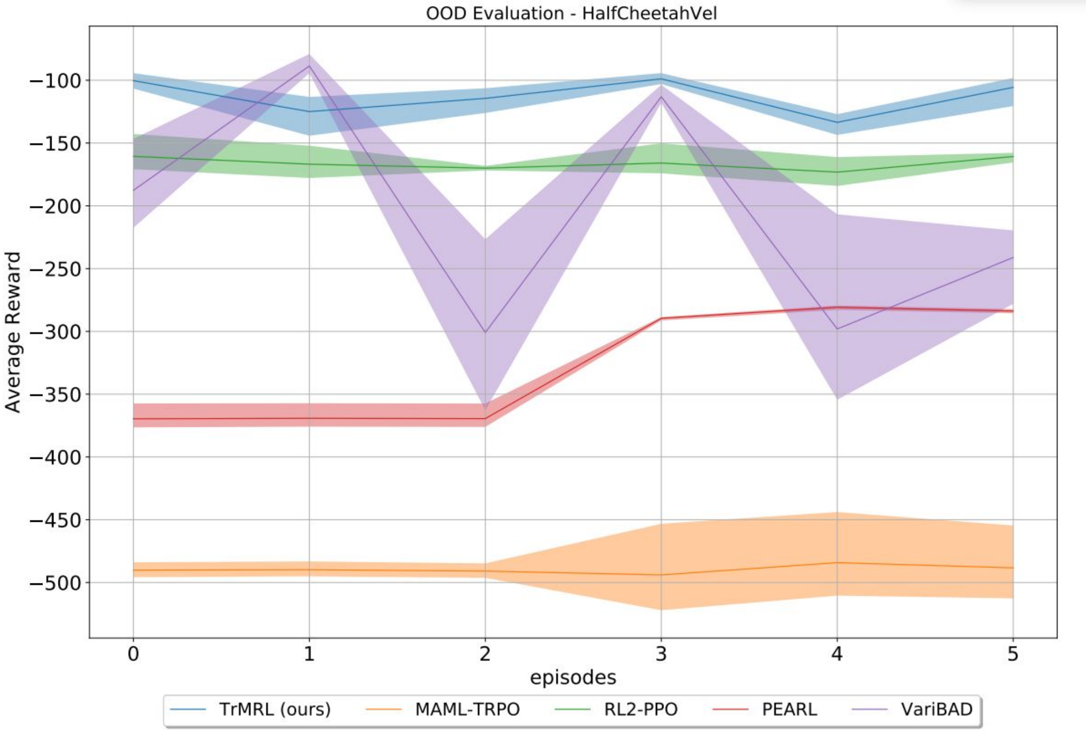
- For RL + Transformers, we need to reconcile initial exploration with the early stages of transformer training
  - Crucial for environments where initially learned behaviors must guide exploration to prevent poor policies

- Proposal: Apply a proper weight initialization scheme: **T-Fixup** (Huang et. al., 2020)

Results: Meta-Training

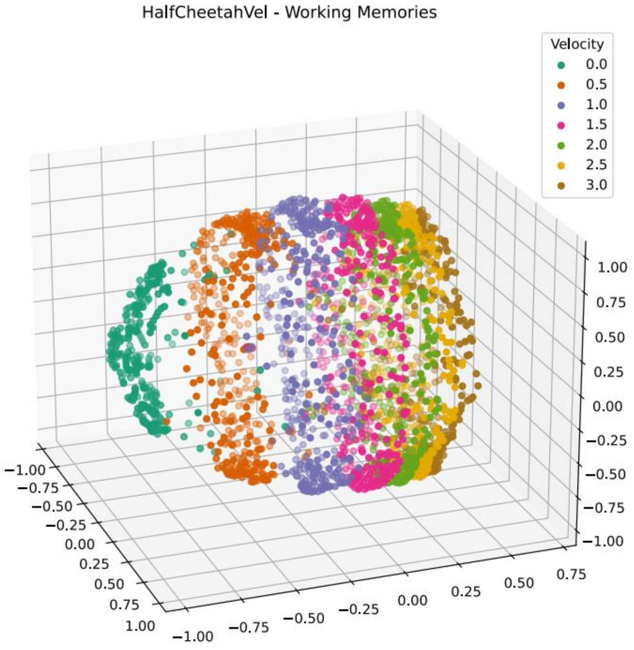


Results: Out-of-Distribution Evaluation

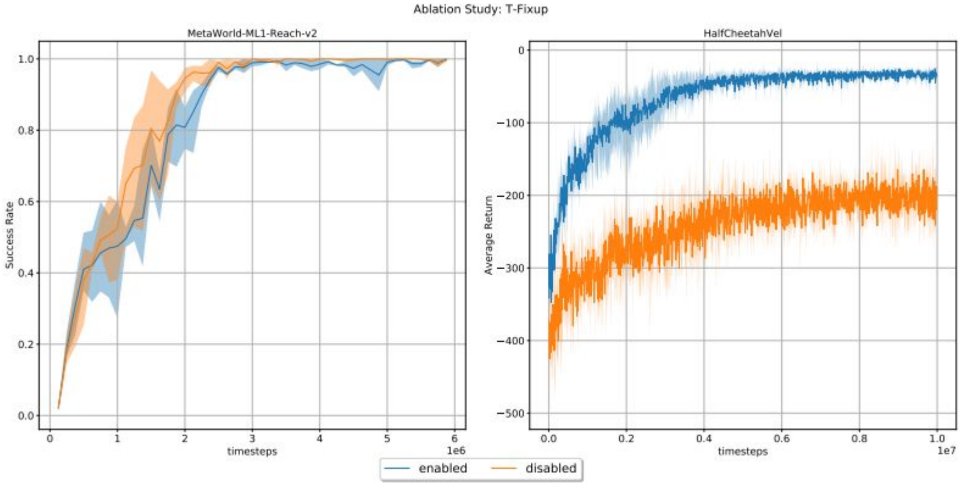


Additional Results

Latent Visualization:



Ablations:



… and many more!

Transformers are Meta-Reinforcement Learners

Source code: https://github.com/luckeciano/transformers-metarl

Luckeciano Melo

Twitter:
@LuckecianoMelo

CEIA
CENTRO DE EXCELÊNCIA EM
INTELIGÊNCIA ARTIFICIAL

Microsoft