

## Using multistability to solve fading memory problems in reinforcement learning

**Auteur :** De Geeter, Florent

**Promoteur(s) :** Drion, Guillaume

**Faculté :** Faculté des Sciences appliquées

**Diplôme :** Master en ingénieur civil en informatique, à finalité spécialisée en "intelligent systems"

**Année académique :** 2020-2021

**URI/URL :** <http://hdl.handle.net/2268.2/11556>

---

### *Avertissement à l'attention des usagers :*

*Tous les documents placés en accès ouvert sur le site le site MatheO sont protégés par le droit d'auteur. Conformément aux principes énoncés par la "Budapest Open Access Initiative"(BOAI, 2002), l'utilisateur du site peut lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces documents, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale (ou prévue par la réglementation relative au droit d'auteur). Toute utilisation du document à des fins commerciales est strictement interdite.*

*Par ailleurs, l'utilisateur s'engage à respecter les droits moraux de l'auteur, principalement le droit à l'intégrité de l'oeuvre et le droit de paternité et ce dans toute utilisation que l'utilisateur entreprend. Ainsi, à titre d'exemple, lorsqu'il reproduira un document par extrait ou dans son intégralité, l'utilisateur citera de manière complète les sources telles que mentionnées ci-dessus. Toute utilisation non explicitement autorisée ci-avant (telle que par exemple, la modification du document ou son résumé) nécessite l'autorisation préalable et expresse des auteurs ou de leurs ayants droit.*

---

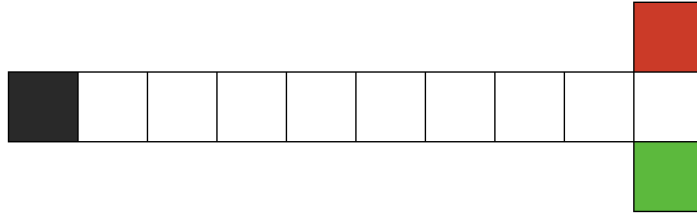


Figure 1: *T-maze* environment with a corridor length of 10. The agent is at the beginning of the corridor (black square) and the treasure is at the bottom of the junction (green square).

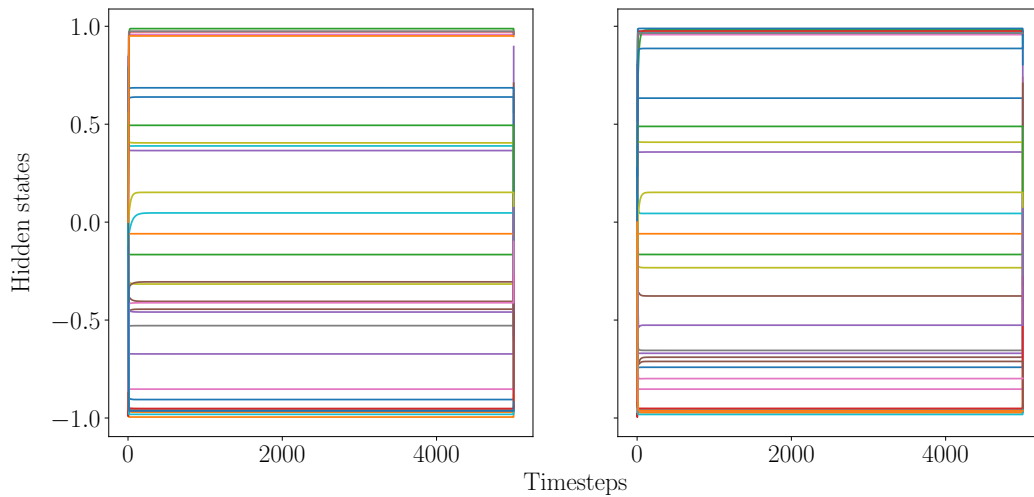


Figure 2: Evolution of the hidden states of a 32-units nBRC agent on two *T-maze* environments with a corridor length of 5000, one with the treasure at the top of the junction, and the other with the treasure at the bottom. The agent wins in both environments.

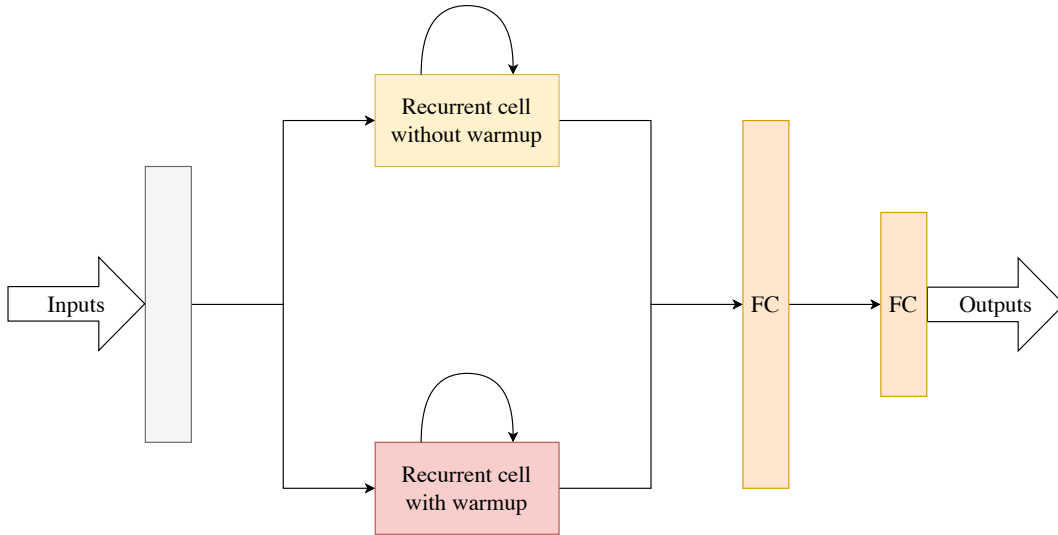


Figure 3: Double layer model used in the experiment. The inputs are passed to two recurrent cells including one that has been pretrained with the warmup. The outputs of these two cells are then concatenated and passed to two fully connected layer.

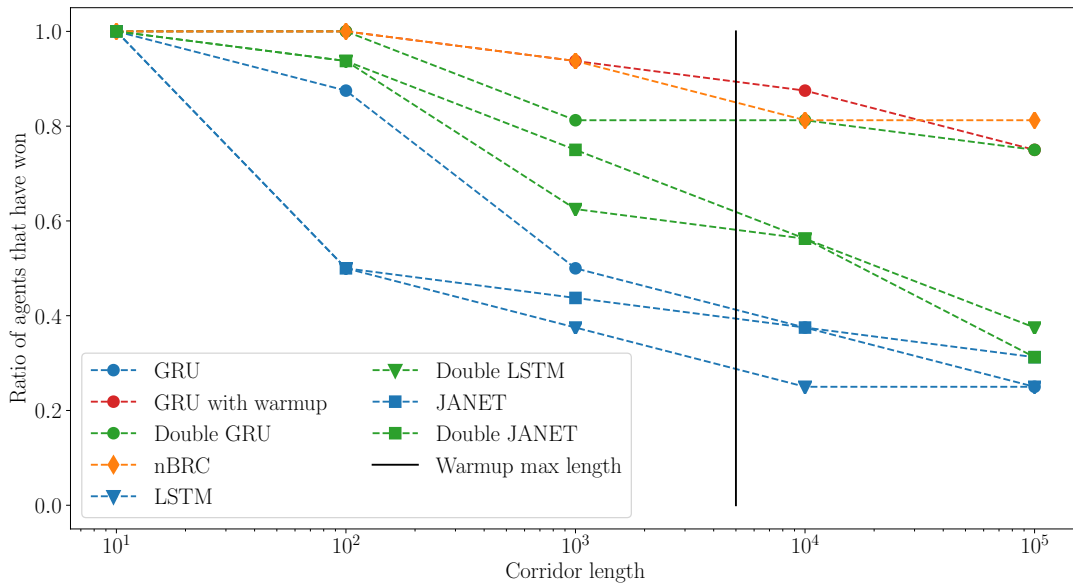
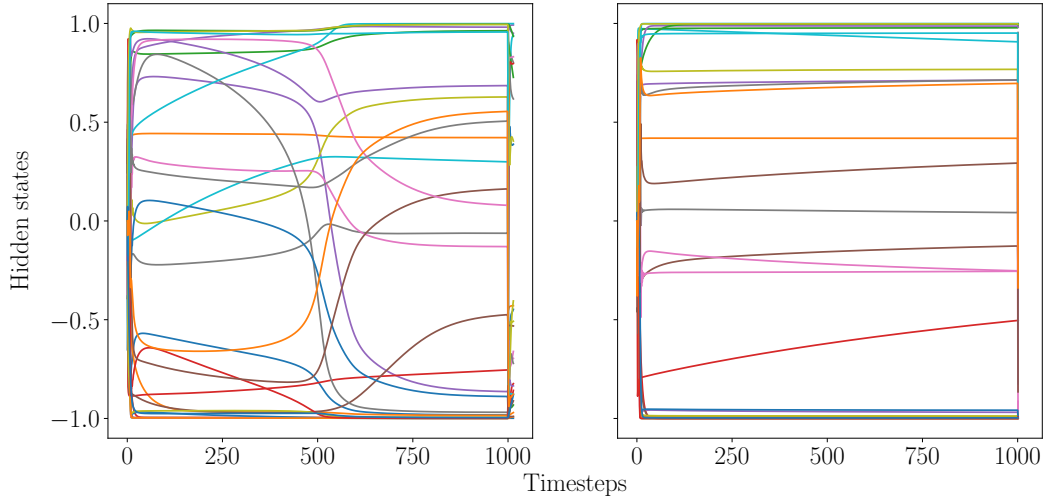
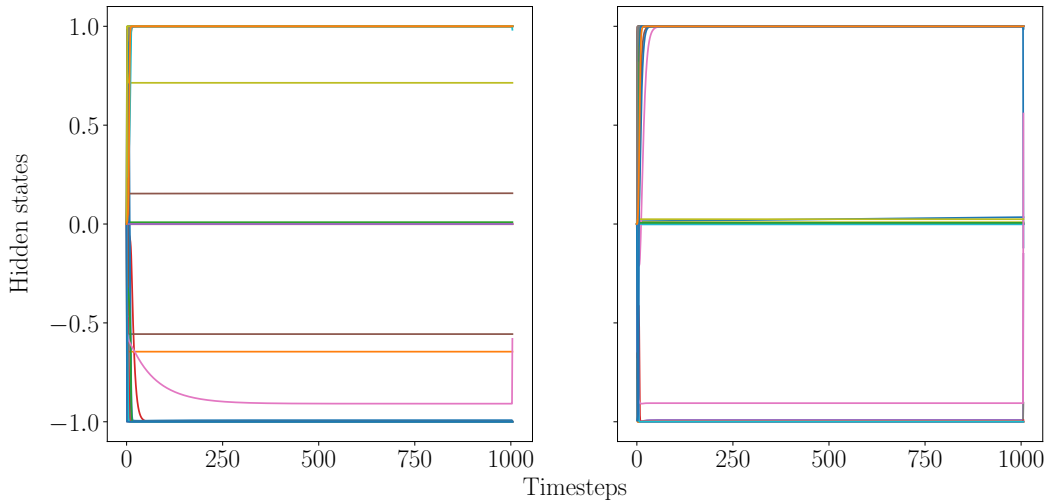


Figure 4: Evolution of the ratio of the agents that always win in the  $T$ -maze environment to the total number of agents trained (16) with respect to the corridor length. The circles are related to GRU, the diamonds to nBRC, the triangles to LSTM and the squares to JANET. The green lines correspond to the double layer models, while the blue ones correspond to the simple layer version with no warmup.



(a) Simple GRU agent, loses when treasure is at the bottom (left graph).



(b) Double GRU agent, wins in both cases.

Figure 5: Evolution of the hidden states of a simple GRU agent and a double GRU agent.