THE IMPACT OF MODEL LEARNING LOSSES ON THE SAMPLE EFFICIENCY OF MUZERO IN ATARI

AUTHOR
Daniel Popovici (d.i.popovici-1@student.tudelft.nl)

SUPERVISORS
Frans Oliehoek, Jinke He

1. INTRODUCTION

- MuZero [1] achieves superhuman performance in Atari at the cost of **millions of environment interactions**
- The learned model is optimized to predict: rewards, values, and policies, **never explicitly trained to match true environment dynamics.**
- Previous research has augmented MuZero with different model learning objectives but they were **not tested for sample efficiency** [2].
- EfficientZero [3] adds a temporal-consistency loss to improve sample efficiency, but this loss: **is not tested in isolation, is not compared to alternatives (e.g. pixel reconstructions), different weights are not explored**

2. RESEARCH QUESTION

How do different model-learning losses impact the sample efficiency of MuZero, measured by scores achieved in Atari games after 100,000 environmental steps?

3. METHODOLOGY

We evaluate and compare:
- **Baseline MuZero Reanalyze (MZ)**: value-equivalent loss (policy, value, reward losses only)
- **MZ + Temporal-Consistency (TC)**: Latent state cosine similarity loss aligning predicted and encoded next states (Fig. 1)
- **MZ + Observation-Reconstruction (OR)**: Pixel-level MSE loss decoding latent states to reconstruct the next observation (Fig. 2)

Evaluation done on games (Fig. 3) from the **Atari100K** benchmark [4].

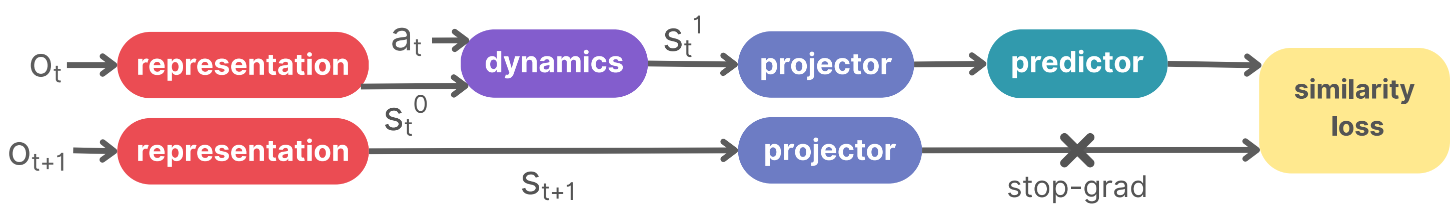

Fig. 1: Diagram of the SimSiam like architecture used to implement the temporal-consistency loss (illustrated for 1 unroll step).

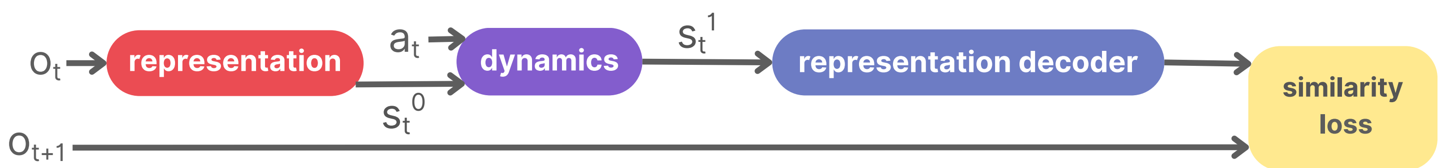

Fig. 2: Diagram of the Autoencoder-like architecture used to implement the observation-reconstruction loss (illustrated for 1 unroll step).

Fig. 3: Evaluated games: Pong, Breakout, MsPacman (in this order)

4. RESULTS

- **TC outperforms the baseline** and OR in Pong and Breakout. In MsPacman the **baseline remains strongest** (Fig. 4).
- In Pong, both TC and OR exhibit **non-monotonic performance** with the model-loss weight. Multiple performance peaks for both at different coefficients (Fig. 5).
- Applying Pong's best weights to Breakout and MsPacman we see **mixed results** (Fig. 6).

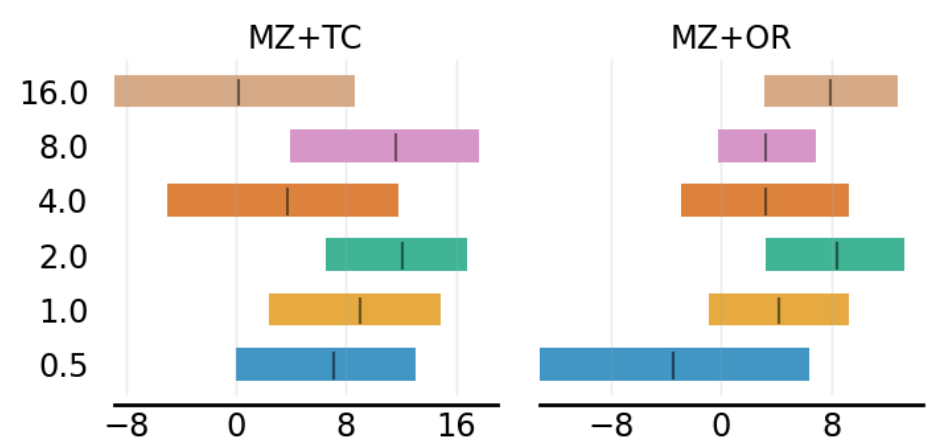

Fig. 5: The impact of weight coefficients on the loss augmented agents in Pong. Mean final evaluation scores shown with black lines and 95% confidence intervals with shaded boxes.

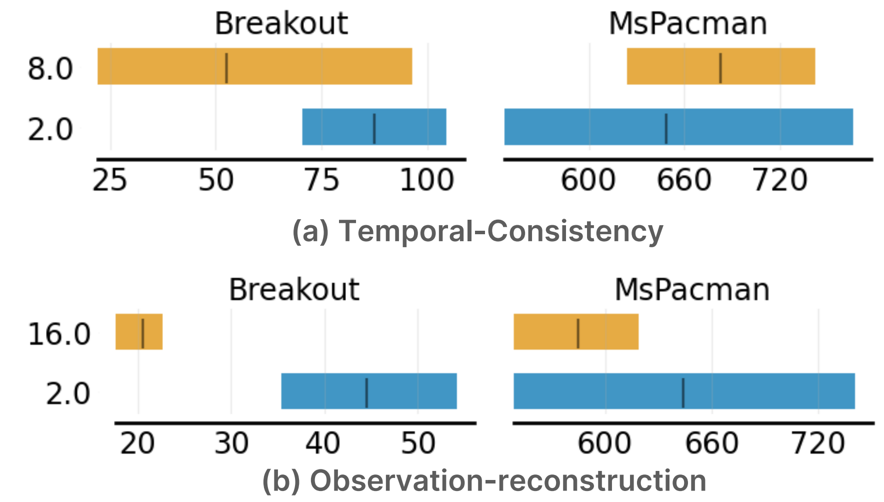

Fig. 6: The impact of weight coefficients on the loss augmented agents in Breakout and MsPacman. Mean final evaluation scores shown with black lines and 95% confidence intervals with shaded boxes.

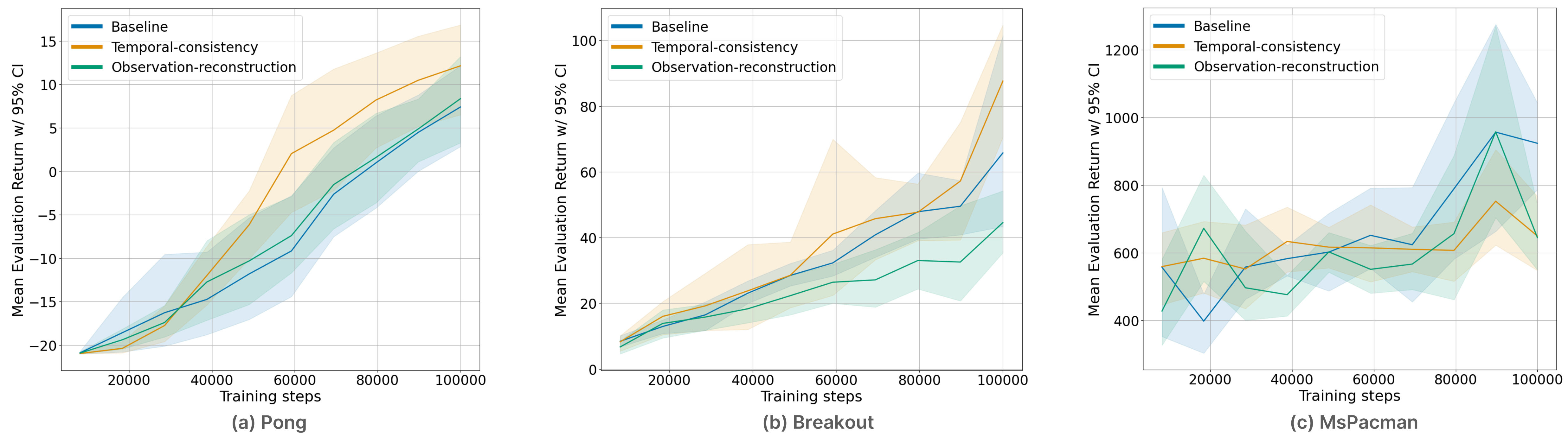

Fig. 4. Training curves of the agent variants. All losses use the weight = 2.0. Mean plotted with 95% CIs as shaded regions Pong results averaged over 10 runs; others averaged over 5 runs.

5. DISCUSSION

- **Simple games like Pong benefit strongly** from temporal-consistency, but in **complex tasks such as MsPacman** the value-equivalent loss performs best.
- **Pixel-level reconstruction fails** to improve compared to the baseline in all environments, proving **too demanding** in low-data regimes.
- Loss **weight needs careful tunning**. Indicates complex and unpredictable interactions with the other parts of the algorithm.
- Optimal loss **weights do not generalize** between environments
- The benefit of model-learning losses **remains unknown for other environments** than Atari or for bigger data budgets.
- **Large differences from published scores** (e.g. -6.7 vs 7.4 in Pong scores from EfficientZero compared to ours) show how small, seemingly unrelated, **changes in other parts of the algorithm** can majorly change low-data performance.

6. LIMITATIONS

This study is limited by the high **computational resources needed for training**. Therefore we focus on only **three Atari games** and **two auxiliary model-loss types**, as well as a modest number of independent runs (5–10). Also other alternative model-learning losses remain unexplored as well as various possible modifications and hyper-parameter choices.

7. FUTURE WORK

We recommend future work to explore **image augmentations** on observations, **other model-loss objectives** (e.g., contrastive losses) alternative **rollout lengths**, and **other environments**.

REFERENCES

[1] Julian Schrittwieser et al. "Mastering Atari, Go, chess and shogi by planning with a learned model". en. In: Nature 588.7839 (Dec. 2020). Publisher: Nature Publishing Group, pp. 604– 609. issn: 1476-4687. doi: 10.1038/s41586- 020- 03051- 4.

[2] . Anand, J. Walker, Y. Li, E. Vértes, J. Schrittwieser, S. Ozair, T. Weber, and J. B. Hamrick, "Procedural Generalization by Planning with Self-Supervised World Models," Nov. 2021. arXiv:2111.01587 [cs].

[3] Weirui Ye et al. Mastering Atari Games with Limited Data. arXiv:2111.00210 [cs]. Dec. 2021. doi: 10.48550/arXiv.2111.00210.

[4] Lukasz Kaiser et al. Model-Based Reinforcement Learning for Atari. arXiv:1903.00374 [cs] version: 5. Apr. 2024. doi: 10.48550/arXiv.1903.00374.