Advances in Reliability, Safety and Security

ESREL 2024
Collection of Extended Abstracts

Advancing Artificial Intelligence Systems: ALife And Generative Model Applications

Carmen Mei-Ling Frischknecht-Gruber[a], Monika Ulrike Reif[b]

[a]*University of Bath, Bath, United Kingdom*
[a,b]*Zurich University of Applied Sciences, Winterthur, Switzerland*

Keywords: artificial intelligence systems, artificial life, robustness, reliability, testing environments

In the rapidly evolving field of AI, ensuring the robustness of Artificial Intelligence Systems (AIS) is critical. This study explores the application of Artificial Life (ALife) principles and Machine Learning techniques to develop challenging test scenarios, with a particular focus on edge cases, which are critical in revealing hidden vulnerabilities. Such scenarios present difficulties due to their rarity and unpredictability and are an essential tool for ensuring the robustness of AIS. Using Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and ALife optimisation algorithms such as Covariance Matrix Adaptation Evolution Strategy (CMA-ES) and Multi-dimensional Archive of Phenotypic Elites (MAP-Elites), this research proposes approaches to level generation for the game Super Mario Bros as a use case. The fidelity of generated levels to the original game design is then evaluated, highlighting the need for AIS adaptation in response to unforeseen challenges.

As the use of AIS increases, concerns about their reliability and safety are also rising. This has led to the development of comprehensive regulations and standards (EU, 2021; IEEE, 2022; DIN, 2023). Additionally, thorough testing is essential to ensure the robustness of these systems (ISO/IEC, 2021). This work focuses on developing synthetic environments for AIS to simulate rare and challenging scenarios, identify vulnerabilities, and adapt and improve AISs accordingly. One of the significant challenges in AIS testing is the limited availability of real-world data to fully capture the breadth of potential scenarios (Kalra and Paddock, 2016; Klischat and Althoff, 2019). This work explores the use of simulation-based environments as a solution, enabling the creation of diverse scenarios to test and reinforce system robustness.

The experimental setup employed Vanilla GAN, DCGAN (Volz, 2018), and Categorical VAE (González-Duque, 2022, 2023) in conjunction with MarioGPT (Sudhakaran, 2023b) for generating Mario levels. The efficacy of these models in level generation was assessed using mentioned architectures, as well as optimisation techniques such as CMA-ES (Hansen, 2016), MAP-Elites (Mouret and Clune, 2015), and Genetic Algorithms (GAs). CMA-ES dynamically adjusts the solution distribution based on fitness to optimise single-objective problems in continuous spaces and enhance exploration in latent space. This approach addresses non-linear problems without requiring domain-specific knowledge, resulting in levels with specific characteristics. In comparison, MAP-Elites generates a spectrum of high-quality solutions in different behavioural spaces. The technique categorises space into cells with distinct characteristics and then iteratively mutates and evaluates solutions to improve performance and facilitate optimal solution discovery. These techniques collaboratively refine the latent vector z, aiming to generate levels that are both challenging and feasible. Two fitness functions, F1 and F2 adapted from Volz et al. (2018) are employed, where F1 targets solvability and balance between challenge and playability, while F2 focuses on reducing the number of jumps, leading to simpler yet playable levels. The effectiveness of these functions is evaluated through latent space analysis for optimal level generation. Evaluation metrics include latent space representation and level playability using Robin Baumgarten's A* agent (Khalifa, 2009; Togelius, 2010). The levels are classified as 'difficult', 'normal', or 'easy' based on the agent's success rate across five runs. Diversity scores are calculated to measure solution variety. Furthermore, manual inspections are carried out to observe any unusual agent behaviour.

Out of 100 levels generated, DCGAN produced exclusively playable levels, while VAE variations and MarioGPT predominantly yielded playable outcomes. When paired with optimisation techniques, DCGAN GA, DCGAN MAP-Elites, and VAE GA all produced solely playable levels. In contrast, VAE MAP-Elites had a significant proportion of unplayable levels. The analysis of the latent space, which focused on the similarity of distribution and the proximity of tiles to the original designs, revealed that minimal categorical VAE and DCGAN outperformed MarioGPT in generating unique levels. Both MarioGPT and VAE showed a high capacity for generating unique levels. With optimisation, VAE GA, VAE MAP-Elites, and VAE CMA-ES achieved high diversity scores (> 0.96). Although MarioGPT and the original VAE created varying predominantly difficult levels, DCGAN consistently produced playable ones. When comparing the most promising models, DCGAN CMA-ES and VAE CMA-ES, DCGAN CMA-ES generated the most difficult levels but also included some unplayable ones.

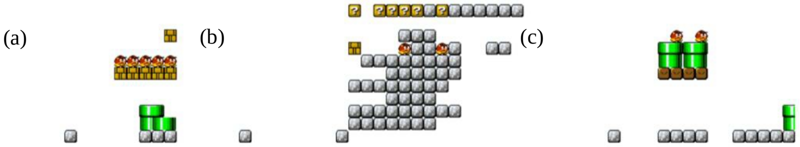

Fig. 1. Difficult considered levels where the agent succeeded after five attempts: (a) DCGAN CMAES; (b) VAE CMAES; (c) MarioGPT.

Despite the limited dataset, a significant number of playable and challenging levels were generated. DCGANs performed exceptionally well in creating new playable levels, while ALife-based variants such as VAE-CMAES and DCGAN-CMAES demonstrated their potential by producing diverse and challenging environments. The existence of broken elements in different levels indicates their potential use in testing AIS beyond their operational design domain, thus enhancing robustness in scenarios that deviate from the typical distribution, as well as in edge cases. However, it is also important to explore an automated validation for level design acceptance to ensure compliance, if necessary, with architectural rules. Targeting specific behaviours, such as jumping in Mario level generation, can sometimes result in unsolvable levels. This highlights the trade-offs involved in creating challenging environments. These levels also revealed atypical AIS behaviours, such as retreating or failing to jump, indicating that unconventional environments are valuable for identifying and improving unusual AIS behaviours. Future research could explore the synergies between GAs, Evolutionary Strategies, and Reinforcement Learning to refine latent vectors. Approaches that involve co-creation of agents and levels could result in a cyclical interplay, fostering the development of increasingly complex levels. Integrating approaches for continuous data and exploring large language models such as GPT-Neo and alternative GPT models for MarioGPT, currently using DistilGPT-2 (Sanh, 2019), are also promising avenues to investigate.

Acknowledgements

This paper is based on the master's thesis 'Enhancing Artificial Intelligence Systems through ALife Principles and Generative Models' and was completed at the University of Bath under the supervision of Ben Ralph.

References

Council of European Union. 2021. Laying down harmonized rules on artificial intelligence com (2021), 206 final.

DIN, DKE. 2023. Artificial intelligence standardization roadmap. Available from: https://www.dke.de.

González-Duque, M. et al. 2022. Mario plays on a manifold: Generating functional content in latent space through differential geometry. 2022 IEEE COG, 385-392.

González-Duque, M. 2023. Minimal implementation of a Variational Autoencoder on Super Mario Bros (v.0.1). Available from: https://github.com/miguelgondu/minimal_VAE_ on_Mario.

Hansen, N. 2016. The CMA Evolution Strategy: A Tutorial. arxiv preprint arxiv:1604.00772.

ISO/IEC TR 24029-1:2021. 2021. Artificial Intelligence (AI): Assessment of the robustness of neural networks: Part 1: Overview. ISO, vol. 1.

Kalra, N., Paddock, S.M. 2016. Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?. Transportation Research Part A: Policy and Practice 94, 182-193.

Khalifa, A. 2009. The mario ai framework. Available from: https://github.com/ amidos2006/Mario-AI-Framework.

Klischat, M., Althoff, M. 2019. Generating critical test scenarios for automated vehicles with evolutionary algorithms. In 2019 IEEE Intelligent Vehicles Symposium (IV), 2352-2358.

Poretschkin, M. et al. 2021. Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz (KI-Prüfkatalog).

Mouret, J. B., Clune, J. 2015. Illuminating search spaces by mapping elites. arXiv preprint arXiv:1504.04909.

Sanh, V. et al. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. Neurips emc2workshop.

Sudhakaran, S. et al., 2023. Mariogpt: Open-ended text2level generation through large language models.

Togelius, J., Karakovskiy, S. and Baumgarten, R. 2010. The 2009 mario ai competition. IEEE CEC, 1-8.

Volz, V. et al. 2018. Evolving Mario Levels in the Latent Space of a Deep Convolutional Generative Adversarial Network.