Investigating the Understanding of Spatio-temporal Information in World Models

Hiroshi Nonaka[*1] Asato Imadate[*2] Tatsuya Ogawa[*3]

[*1] Soka University of America, California USA
[*2] National Institute of Technology (KOSEN), Suzuka College [*3] Kobe University

Recent studies have revealed that language models encode internal representations of space and time, suggesting that world models may emerge within the language space during training. In this study, we aim to investigate whether world models acquire spatiotemporal representations similar to those of language models, and if so, how and where within their architecture such representations are acquired. Specifically, we examine DreamerV3 to explore the location and characteristics of neurons encoding meaningful spatiotemporal representations. Using the "Crafter" benchmark, a task rich in spatiotemporal features, we pre-trained models and performed probing experiments to evaluate their internal representations. We measure the ability of each model to reconstruct spatial (e.g., coordinates) and temporal (e.g., timestamps) information using linear probing and analyze the linearity and distribution of spatiotemporal representations across model layers.

1. Introduction

World models have gained researchers' interest due to their training cost-effectiveness in reinforcement learning, video generation, and high-level modeling of the decision-making of living systems. In contrast, large language models (LLMs) demonstrate sophisticated language reasoning, becoming a landmark of artificial intellingence development. Some scholarly works have delved into the internal representations of LLMs using probing techniques, revealing the emergence of spatiotemporal understanding of the world inside the models. However, few works have discussed whether and where world models contain spatiotemporal representations. Our research examined the presence of such meaningful features, and, if so, where in the models we can discover them.

2. Related Research

2.1 Spatiotemporal understanding of language models

Among studies suggesting a relationship between world models and language models, Gurnee et al. demonstrated that LLMs can learn representations of spatial and temporal information through language data about geography and history [Gurnee 23]. By analyzing the internal activations of Llama-2 [Touvron 23] across its layers, they trained linear regression probes that predict the latitude and longitude of renowned places and the active years of historical figures. The linear regression probes are expressed by Equation (1) [Alain and Bengio 16][Belinkov 22]. These probes produce predictions based on an activation dataset A and a target dataset Y, which includes spatial data reflecting geographical information and temporal data encompassing historical and current event information.

$$\hat{W} = \arg\min_{W} ||Y - AW||_2^2 + \lambda ||W||_2^2 = (A^T A + \lambda I)^{-1} A^T Y \quad (1)$$

The research by Gurnee et al. suggests that spatial and temporal features can be recovered through the linear probes, indicating that such information is encoded as linear representations. Moreover, the quality of these learned representations is noted to be significantly influenced by factors such as model scale and the volume of training data. However, as the study primarily focuses on LLMs, it does not delve deeply into the specific connections between LLMs and those world models that implicitly model the dynamic transition of states.

2.2 World Models

World models provide a broad framework for capturing implicit representations of external environments [Ding 24]. The notion of World Models [Ha 18] can be traced back to research showing that internal neural network structures can learn to represent environmental dynamics. Building on this foundation, the Dreamer series introduced powerful model-based reinforcement learning techniques, leveraging actor-critic methods and "latent imagination" to refine the learning process. In Dreamer, the agent learns a latent dynamics model—such as a Recurrent State Space Model (RSSM)—capable of predicting future states from compact representations of high-dimensional inputs. By generating "imaginary" trajectories within latent space, Dreamer trains a policy more efficiently than purely model-free methods. This approach ef-

Contact: Asato Imadate,
National Institute of Technology (KOSEN), Suzuka College, Shiroko-cho, Suzuka, Mie, 510-0294, Japan, h31i05@ed.cc.suzuka-ct.ac.jp

fectively combines actor-critic methods with a learned world model, enabling the agent to plan and update its policy based on these imagined rollouts rather than direct interactions with the environment. Each iteration of the Dreamer series introduced key improvements: DreamerV1 [Hafner 20] pioneered the RSSM, DreamerV2 [Hafner 21] incorporated discrete latent variables, and subsequent versions added stability measures and training optimizations. Collectively, the Dreamer series stands as a prime example of model-based reinforcement learning built atop actor-critic principles.

3. Proposed Method

3.1 Learning the world model

In this research, we use DreamerV3 [Hafner 23], the latest version of the well-established Dreamer model, as the world model for analyzing spatiotemporal representation learning. For the training environment, we adopt Crafter [Hafner 22], based on the following three reasons: (1) it allows exploration of relatively large maps, thereby providing sufficient spatial information; (2) the episode length, representing the agent's survival time, can be effectively utilized as temporal information; and (3) the computational cost for training is relatively low.

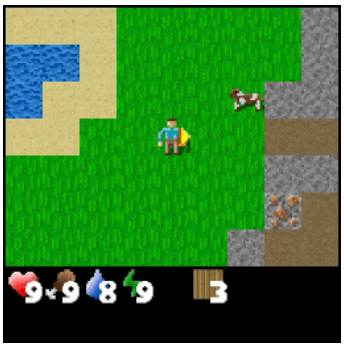

Figure 1: Example of crafter's playing screen [Hafner 23]

3.2 Dataset for probing

Following the methodology of Gurnee et al., in our experiments, we collected the internal activations of the deterministic states of Recurrent State Space Model (RSSM) and stochastic states in DreamerV3 by passing randomly sampled image vectors from Crafter through the model [Hafner 19]. The activation data used for training the probes (A in Equation (1)) is described in Table 1. Similarly, the target data used for training (Y in Equation (1)) is detailed in Table 2. The player's current position (*pos*) is represented as a two-dimensional xy-coordinate, while the elapsed time (*episode*) is represented as a scalar.

3.3 Probing

Following the methodology of Gurnee et al., this research also employs the linear probes and nonlinear probes described in Section 2.1 to analyze the representation of spatiotemporal information. Technically, we trained RidgeCV and multilayer perceptron (MLP) probes using the dataset explained in Section 3.2, with the training data, and then used the test data to evaluate the predictions made by the trained probes. For details on the training settings, refer to Section 3.4, and for the evaluation methods of the obtained predictions, refer to Section 3.5.

Table 1: Number of data for each activation

layer	vector	count
encoder	embedding observation	
rssm	deterministic state	9712
	stochastic state	9712
	logits state	
	policy	
decoder	embedding observation	

Table 2: Number of data for each target

target	count
pos	9712
episode	9712

3.4 Hyperparameters

We used the default hyperparameters of the DreamerV3 model to increase the reproducibility of our evaluation. The training loop ended with a total step of 16 million. During training, we collected 64×64 observation images, xy positions of the agent in the environment, and episode count for every 100 steps for the probing dataset. For training the linear regression model, we set the regularization parameter α by dividing the range $10^{0.8}$ to $10^{4.1}$ into 12 equal intervals. This configuration aligns with the smallest model size, Llama2-7B, used in the study by Gurnee et al. For the nonlinear model, we utilized a two-layer MLP with a hidden layer dimension of 256. For the weight decay of the optimizer, we tested four values—0.01, 0.03, 0.1, and 0.3—and selected the one that yielded the best performance. In both models, the training and testing datasets were randomly split at a ratio of 8:2.

3.5 Evaluation

To investigate the spatiotemporal representations in the pre-trained DreamerV3 model, we trained our Ridge regression and MLP probes. Our spatial probes take internal activations as input and predict xy-coordinates, while temporal probes predict episode counts from the activations. With reference to the methodology of Gurnee et al, this investigation will also evaluate the performance of our investigation by reporting the decision index R^2 and standard regression indices for the test dataset and discussing the results

Table 3: Standard regression indices for test data showing the results of probing for spatial information in the linear regression model at each step

steps	**activation**	R^2 of X	R^2 of y	R^2	MAE of X	MAE of y	MAE	RMSE
1.1	deterministic state	-6.5e-4	-5.6e-5	-3.5e-4	8.4	8.6	8.5	11.6
	stochastic state	0.21	0.23	0.22	7.3	7.8	7.5	10.3
1.6	deterministic state	-5.9e-3	-9.4e-4	-3.4e-3	9.5	9.7	9.6	12.3
	stochastic state	0.24	0.23	0.23	8.0	8.2	8.1	10.8

Table 4: Standard regression indices for test data showing the results of probing for spatial information of MLP at each step

steps	**activation**	R^2 of X	R^2 of y	R^2	MAE	RMSE
1.1	deterministic state	-1.9e-3	-1.1e-4	-1.0e-3	8.5	11.7
	stochastic state	0.22	0.25	0.23	7.4	10.2
1.6	deterministic state	-2.4e-3	-7.4e-3	-4.9e-3	9.6	12.3
	stochastic state	0.25	0.24	0.24	8.1	10.7

Table 5: Standard regression indices for test data showing the results of probing for time information of MLP at each step

steps	**activation**	R^2	MAE	RMSE
1.1	deterministic state	-4.8e-4	65.9	75.6
	stochastic state	3.3e-4	65.8	75.6
1.6	deterministic state	-2.3e-3	63.5	73.7
	stochastic state	-8.9e-4	63.5	73.7

Table 6: Standard regression indices for test data showing the results of probing for time information of MLP at each step

steps	**activation**	R^2	MAE	RMSE
1.1	deterministic state	-3.4e-4	65.9	75.6
	stochastic state	5.6e-3	65.4	75.4
1.6	deterministic state	-4.5e-3	63.6	73.8
	stochastic state	-2.6e-2	64.2	74.6

of this investigation.

4. Experimental Results

Our experimental results are shown in Tables 3 to 6. Overall, our conclusion is that world models fail to store useful representations that can be consistently reconstructed by linear or non-linear probes. However, the result clearly shows that stochastic states succeed in constructing the linearly consistent spatiotemporal representation better than deterministic states do. The R^2 scores of stochastic state excel that of deterministic state, indicating that stochastic states have less variance in the prediction and successfully contain linear representation for spatiotemporal understanding compared to the deterministic states. Non-linear probes also demonstrate a similar results to the linear probing, nonetheless, MLP succeeded in improving the R^2 scores of deterministic states, suggesting that the deterministic states contain spatiotemporal representations in a non-linear form.

5. Conclusion

In our work, we investigated the nature of spatiotemporal representations in world models and used probes to reconstruct spatiotemporal information from the activations obtained through the model to help understand the linearity and non-linearity of such representations. Our research suggested that while deterministic states lack rich linearity, stochastic states contain more consistent representations that can be decoded linearly by Ridge regressor models. On the other hand, the R^2 scores of deterministic states increased in the non-linear probing settings, indicating the complexity of those representations. Although our work could not identify meaningful representations stored in the pre-trained DreamerV3, future research should be done to investigate how the amount of training impacts the understanding of spatiotemporal representations and the linearity and non-linearity of spatiotemporal representations in other pre-trained world models.

Acknowledgements

This research was conducted as part of the ‘ World Models 2024 ’ course, endowed by the University of Tokyo's World Models and Simulators. We would like to express our deepest gratitude to Professor Yutaka Matsuo and the lecturers for providing us with this opportunity.

References

[Ha 18] Ha, D. and Schmidhuber, J.: World Models (2018)

[Ding 24] Ding, J., Zhang, Y. et al.: Understanding World or Predicting Future? A Comprehensive Survey of World Models (2024)

[Gurnee 23] Gurnee, W. and Tegmark, M.: Language Models Represent Space and Time (2023)

[Touvron 23] Touvron, H., Martin, L. et al.: Llama 2: Open Foundation and Fine-Tuned Chat Models (2023)

[Alain and Bengio 16] Alain, G. and Bengio, Y.: Understanding Intermediate Layers Using Linear Classifier Probes (2016)

[Belinkov 22] Belinkov,Y. : Probing Classifiers: Promises, Shortcomings, and Advances (2022)

[Hafner 23] Hafner, D., Pasukonis, J., Ba, J. and Lillicrap, T.: Mastering Diverse Domains through World Models (2023)

[Hafner 22] Hafner, D.: Benchmarking the Spectrum of Agent Capabilities (2022)

[Hafner 19] Hafner D., Lillicrap T. et al.: Learning Latent Dynamics for Planning from Pixels (2019)

[Hafner 20] Hafner, D., Ba, J., Norouzi, M. and Fischer, I.: Dream to Control: Learning Behaviors by Latent Imagination (2020)

[Hafner 21] Hafner, D., et al.: Mastering Atari with Discrete World Models. In International Conference on Learning Representations (2021)