Mechanising Conceptual Spaces using Variational Autoencoders

Max Peeperkorn[1], Rob Saunders[2], Oliver Bown[3], Anna Jordanous[1]
[1] School of Computing, University of Kent, Canterbury, United Kingdom
[2] Leiden Institute for Advanced Computer Science, Leiden University, Leiden, Netherlands
[3] Faculty of Art and Design, University of New South Wales, Sydney, Australia
mp770@kent.ac.uk

Abstract

In this pilot study, we explore the Variational Autoencoder as a computational model for conceptual spaces in a social interaction context. Conceptually, the Variational Autoencoder is a natural fit for this purpose. We apply this idea in an agent-based social creativity simulation to explore and understand the effects of social interactions on adapting conceptual spaces. We demonstrate a simple simulation setup and run experiments with a focus on establishing a baseline. While ongoing work needs to identify if adaption was appropriate, the results so far suggest that the Variational Autoencoder appears to adapt to new artefacts and has potential for modelling conceptual spaces.

Introduction

In society, humans share their ideas and exchange artefacts. We draw inspiration from these interactions, and this sparks our imagination to produce new ones (Vygotsky 2004). Every individual has a unique perspective, a style of thought, embedded in a conceptual space (Boden 2004). While ideas and artefacts can be attributed to individuals, they are shaped by others, leading to a distributed emergence of creativity.

In this paper, we explore the use of the Variational Autoencoder (VAE) (Kingma and Welling 2014) as a computational model for the conceptual space in an artificial social context. This is an initial study investigating how to embed and maintain VAEs in an agent-based Computational Social Creativity (Saunders and Bown 2015) simulation.

Background

Conceptual Spaces... There are two views on conceptual spaces: a creativity view (Boden 2004) and a general cognitive view (Gärdenfors 2004). Gärdenfors proposed conceptual spaces as a geometric mental structure to organise thought, with the aim to bridge the symbolic and the subsymbolic. It allows finding similarities between symbols that cannot be derived from the symbolic level alone. According to this theory, concepts are convex regions in the conceptual space, and the axes represent properties. Boden's view of the conceptual space is well-known and a central part of her creativity framework concerning the three modes of creativity. This definition is abstract and less defined, simply the set of artefacts that follow the rules of a given domain. While useful to reason about creativity, Boden's abstract definition is unsuited for computational purposes. However, in this paper, we are less concerned with the formal definition and use both views to inform our choices in the simulation. We use Boden's view to examine the creative act and use Gärdenfors' view to inform traversing the conceptual space.

...and Variational Autoencoders Due to its probabilistic nature, and compression and generative capabilities, we explore the idea that VAE is conceptually a natural fit for approximating conceptual spaces. The VAE is a deep generative model that learns fuzzy relations in the data and maps this onto smooth latent spaces—which is reminiscent of Gärdenfors' geometric conceptual space. The latent space can be queried to find similar artefacts and sampled to generate new artefacts. This makes it particularly interesting to use as a way for agents to perceive, interpret, and produce artefacts. Based on its characteristics, we assume that the VAE is a reasonable abstraction for the formation of concepts and properties.

Simulation

Like other simulations of social creativity (Saunders 2012), the DIFI model (Feldman, Csikszentmihalyi, and Gardner 1994) provides the conceptual model for the simulation presented here. To explore how to embed and maintain the VAEs in a simulation, we use them in two ways: as the conceptual space for each agent and as a recommender system for the whole domain. Next, we discuss the data representation, VAE architecture, each component of the DIFI model, and further discuss the details of the utility of VAE in the simulation.

Data Representation

For use in the simulation, the VAEs require pre-training that can be likened to providing basic education for each agent. Initially, we used a generated dataset in a simplified musical domain of short melodies of 16 timesteps of 12 pitches (chromatic scale) (Peeperkorn, Bown, and Saunders 2020). Further work proved this dataset to be problematic and led to heavy overfitting when pre-training the VAEs. To mitigate this, we generated a dataset using Hidden Markov Mod-

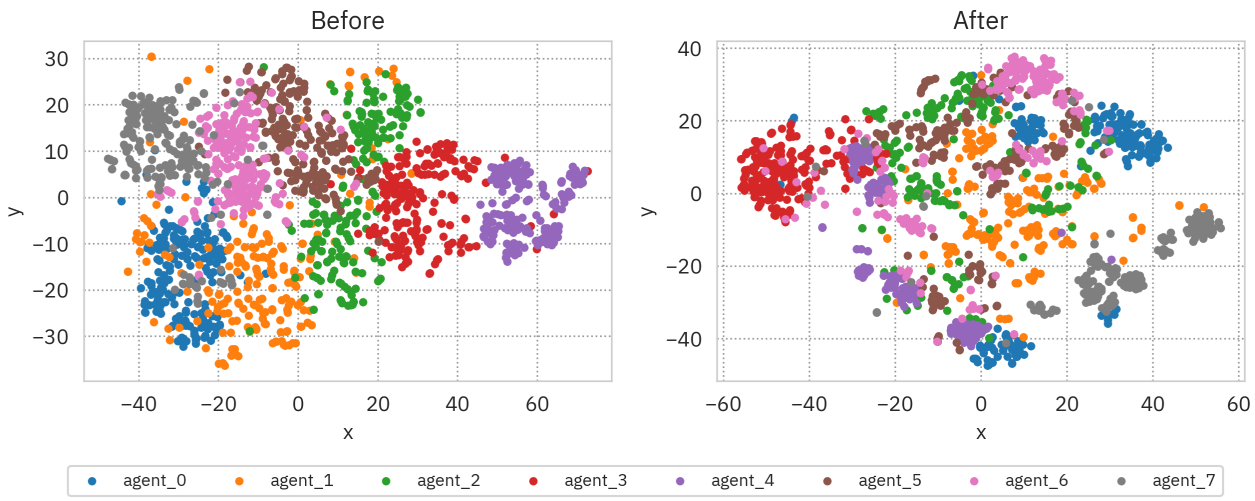

Figure 1: Sampling from Agent VAEs before and after the simulation and compared in the Domain VAE projected using T-SNE.

els fitted to real music data.[1] Subsequently, we generated a combined dataset of 400k samples of 16 timesteps and 88 pitches. We considered other datasets, such as images of typefaces, but the benefit of using categorical data is that it allows for exact reconstructions.

Recurrent VAE architecture

We used a simple recurrent VAE (Fabius and Van Amersfoort 2014) using Long Short-Term Memory (LSTM) layers. A big issue with recurrent VAEs is posterior collapse which occurs when the network learns to ignore the latent space. The Kullback-Leibler (KL) term is annealed in the early stages of the training (Bowman et al. 2015) to mitigate this issue allowing the VAE to extract informative features before the full penalty smooths the latent encodings. The final VAE network has a 32-dimensional latent space. The encoder and decoder consist of two hidden LSTM layers with 128 nodes. For initial training, we used a batch size of 512 and KL-annealing over the first 200 epochs.

DIFI model Setup

Domain The domain is explained as a cultural repository of knowledge (Csikszentmihalyi 2014). In this work, there is no single repository for agents to access. Instead, the domain is distributed amongst the agents' conceptual spaces, each with a personal subset of embedded knowledge. This does not allow artefact comparison on the individual level, and therefore, we introduce a static and pre-trained Domain VAE. It operates as an archimedean point that enables the analysis of the distributed domain. Additionally, the Domain VAE is used to split the dataset into different slices for each agent using a 2D PCA projection of the latent encodings.

Individual Each agent in the simulation has a personal VAE, each trained on a different slice. In contrast to the Domain VAE, the individual agent uses the VAE to learn from and generate new artefacts. Generating is done by randomly sampling from a gaussian distribution, and decoding the latent vector to produce the artefact. We assume that the standard deviation can be used as a proxy for novelty preference. A narrow distribution produces less varied artefacts, and conversely, a wide distribution produces high variation.

Field The field acts as a gatekeeper for what artworks are selected for circulation, according to the ideology of society (Csikszentmihalyi 2014). Different ideologies use different selection criteria, and subsequently, influence the social interactions taking place in the domain. The field acts according to an ideology, a social policy, for selecting artefacts for the next round in the simulation. In the current setup, we use a neutral policy, i.e. that every artefact has an equal chance of being "put on display" in the field. The Recommender System (Domain VAE) informs the field of its choices. As such, the field fulfils two roles in the model: the matchmaker and the gatekeeper. The matchmaker takes the newly produced artefacts and determines the agent's position to find neighbours who share their artefacts. Subsequently, each agent has a different pool from which the gatekeeper will select for the next round.

Interaction After initialising the VAEs, the simulation iterates through three stages. The first stage is associated with the individual, where the newly observed artefacts are used to fine-tune the agents' latent space for a given learning budget to extract new features and then produce several new artefacts sampled according to the novelty preference. The second stage is where the field receives the position of each agent, queried from the Recommender System using the mean of the newly produced artefacts. In the third stage, the positions are used to determine the agents' nearest neighbours. The neighbour shares their artefacts, which form a pool of artefacts. Subsequently, the field selects artefacts from this pool for the next round according to its ideology.

[1]The data is gathered from the Humdrum database (https://kern.humdrum.org), selecting the 8 genres with the most samples.

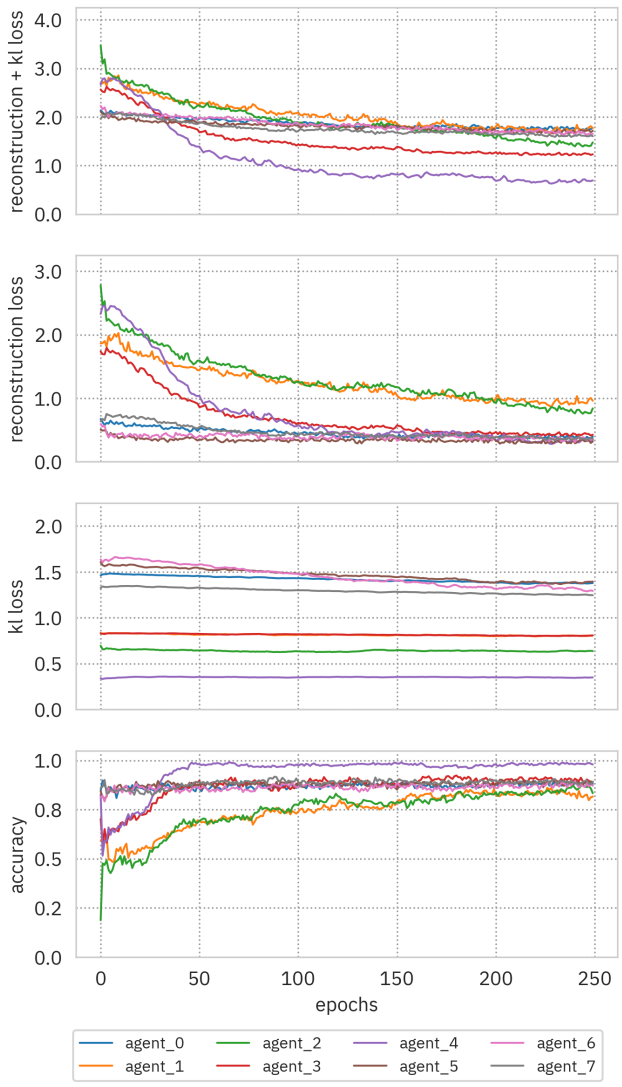

Figure 2: Agent VAE performances evaluating artefacts over a sliding window of 25 epochs.

Results

The simulation experiments use the following settings: 250 epochs with 8 agents, the neutral ideology, and novelty preference set to 0.25. Each round, the field selects 128 artefacts, individuals produce 4 new artefacts, and 1 neighbour shares their artefacts. Each agent has a 5-epoch budget for fine-tuning using a learning rate of 10^{-4}.

The VAEs are trained on the respective datasets using a 70/30 train/validation split. Table 1 shows that Domain VAE performs very well. The agents show clear clusters after the initialisation (Fig. 1). However, the agent VAE pre-training show very mixed results and some perform well (>80% accuracy), while others do not (<30% accuracy).

Post-simulation sampling of the agent VAEs suggests that they mingled as expected (Fig. 1). However, there are a few very dense clusters, which could signify that the latent space is collapsing.

Table 1: Pre-training results show loss and accuracy after 2000 epochs. The Agents VAE shows the mean results for 8 agents.

	Loss		**Accuracy**	
	Train	Val	Train	Val
Domain VAE	2.028	2.034	.937	.934
Agents VAEs	2.593	2.894	.559	.497

The results in Fig. 2 on the other hand, appear to indicate that agents adapt well, within a 25-epoch sliding window, to the artefacts selected each round as accuracy goes up and reconstruction loss goes down. It is somewhat surprising given the agent initialisation results (Table 1). While this is desirable, it might also indicate overfitting. The KL loss is level, suggesting latent space stability, but an issue is that, for some agents, it is already very low after pre-training.

Discussion

The results suggest the conceptual spaces drift stably, which, in turn, suggests that the VAEs adapt. However, it does not inform to what extent they adapted and if it is appropriate according to the social dynamics and interactions. With the current setup, it is very difficult to observe exact agent behaviours. Crucial for future work is to further investigate VAE performance during the simulation and rule out the previously mentioned issues, such as overfitting or posterior collapse. Even though the VAEs appear operable, the performance still causes some concern. It could be due to the datasets, but it might also be that the domain requires a more sophisticated VAE architecture, such as the Hierarchical decoder (Roberts et al. 2018).

This paper focuses on getting the VAE to work and less on the social dynamics. It does provide opportunities for examining different novelty preferences or ideologies, for example, progressive (seeking novelty) and conservative (seeking familiarity). These research directions are interesting to explore, but they depend on the ability to look inside the simulation and inspect the VAE behaviour. The main challenge remains: to develop the tools leveraging latent traversals to increase understanding of how the VAE behaves throughout the simulation. This is necessary to see if social dynamics and interactions explain agent VAE divergences. But this work establishes an initial baseline for future work.

Conclusion

The work presented here is an initial study into mechanising conceptual spaces using VAEs. The results suggest the potential for the VAE as a computational model for conceptual spaces. We stress that additional sophisticated analysis is necessary to further examine the VAE behaviours. However, it shows the potential of VAEs for modelling ill-defined domains without predetermined rules, which is so often the case with creative domains.

Author Contributions

MP designed the simulation and experiments with RS. MP implemented the simulation and experiments. MP and AJ contributed to improving the pre-training dataset setup. RS, OB, and AJ contributed valuable ideas, insights and supervisory feedback. MP wrote the paper. MP and AJ edited the paper.

Acknowledgements

We would like to thank Marek Grześ, Piotr Sawicki, and Dan Brown for their insightful suggestions and comments. This work is a continuation of the MP's master thesis supervised by RS and OB for the Media Technology MSc programme at Leiden University, NL. MP is supported by the University of Kent GTA Studentship Award, Prins Bernhard Cultuurfonds, Hendrik Mullerfonds, and Vreedefonds.

References

Boden, M. 2004. *The Creative Mind: Myths and Mechanisms*. London: Routledge.

Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A. M.; Józefowicz, R.; and Bengio, S. 2015. Generating sentences from a continuous space. *CoRR* abs/1511.06349.

Csikszentmihalyi, M. 2014. *Society, Culture, and Person: A Systems View of Creativity*. Dordrecht: Springer Netherlands. 47–61.

Fabius, O., and Van Amersfoort, J. R. 2014. Variational recurrent auto-encoders. *arXiv preprint arXiv:1412.6581*.

Feldman, D. H.; Csikszentmihalyi, M.; and Gardner, H. 1994. *Changing the world: A framework for the study of creativity*. Westport: Praeger Publishers.

Gärdenfors, P. 2004. *Conceptual spaces: The geometry of thought*. MIT press.

Kingma, D. P., and Welling, M. 2014. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR) 2014*, volume abs/1312.6114.

Peeperkorn, M.; Bown, O.; and Saunders, R. 2020. The maintenance of conceptual spaces through social interactions. In *Proceedings of BNAIC/BENELEARN 2020*. Master Thesis Abstract.

Roberts, A.; Engel, J.; Raffel, C.; Hawthorne, C.; and Eck, D. 2018. A hierarchical latent vector model for learning long-term structure in music. In *International conference on machine learning*, 4364–4373. PMLR.

Saunders, R., and Bown, O. 2015. Computational social creativity. *Artificial Life* 21(3):366–378.

Saunders, R. 2012. Towards autonomous creative systems: A computational approach. *Cognitive Computation* 4(3):216–225.

Vygotsky, L. S. 2004. Imagination and creativity in childhood. *Journal of Russian & East European Psychology* 42(1):7–97.