Novelty-seeking Guides Formation of Disentangled Representations

Pingsheng Li (pingsheng.li@epfl.ch)
EPFL, Neuro-X Institute, Switzerland

Abstract

Disentangled representations are common in the brain, where many neurons are tuned to single factors of task variation, such as place cells or object-vector cells. Previous work has shown that neural networks trained with biological constraints can also learn disentangled representations if trained on disentangled data, i.e., data generated by independent factors. However, in real-world, open-ended environments, such neatly disentangled data may not always be available. This raises a fundamental question: how can agents collect experiences that help them form disentangled representations? Intrinsic motivations, such as novelty, efficiently guide humans and artificial agents during exploration of unfamiliar environments but it is unclear whether they also support disentangled representation learning. Using a novel method to extract representation-specific novelty signals, we compute novelty signals from the latent representations of autoencoders (AEs) and discrete variational autoencoders (D-VAEs) and use them as intrinsic exploration rewards for an artificial agent performing unsupervised learning. We show that these novelty signals favor exploration of disentangled over entangled data, and help the agent learn disentangled representations.

Keywords: disentangled representation; unsupervised learning; intrinsic motivations; exploration; novelty-seeking

Introduction

Many neurons in the brain are tuned to individual factors underlying task variations (O'Keefe, 1976; Høydal et al., 2019), forming representations that are referred to as 'disentangled' (Bengio, Courville, & Vincent, 2014). Neural networks trained with biological constraints, such as non-negativity and energy efficiency, can also learn such disentangled representations from data characterized by independent factors **if trained on disentangled data**. (Whittington et al., 2023) (see also (Plumbley, 2003; Hyvarinen, Khemakhem, & Morioka, 2023) for related methods). However, in real-world environments, data may not always be neatly disentangled. Intrinsic motivations, such as novelty-seeking (Bellemare et al., 2016; Xu et al., 2021), (Barto, Mirolli, & Baldassarre, 2013; Becker, Modirshanechi, & Gerstner, 2024) guide agents exploring unfamiliar environments and help them learn about sparse rewards by sampling the environment efficiently. However, it is unclear whether seeking intrinsic motivations also helps agents sample disentangled data that they need for disentangled representations learning. In this work, we explore the role of novelty-seeking behaviors in shaping disentangled representations. Using a novel method proposed in Becker et al. (2024), we can extract novelty signals from existing network representations in autoencoders (AEs) (Hinton & Salakhutdinov, 2006) and discrete latent variational autoencoders (D-VAEs) (Friede et al., 2023) and use them as intrinsic rewards to guide agents toward data generated by disentangled factors. In a simple multi-armed bandit setup, we show that novelty seeking enhances disentangled representation learning. This suggests that novelty-seeking could promote learning of disentangled representations in an open-ended environment.

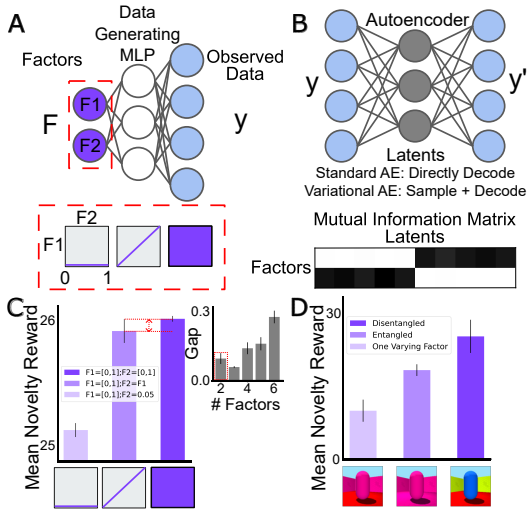

Figure 1: (A) (Top) A data-generating MLP maps latent factors F to observed data y. (Bottom) Factor-sampling strategies: single varying, entangled, and disentangled factors. Purple indicates the domain of factors for each strategy. (B) (Top) An autoencoder learns latent representations from the generated data via reconstruction. (Bottom) Example mutual information (MI) matrix quantifies how much information each latent encodes about each factor. White indicates high MI. (C) Mean novelty reward in trained AEs over a sequence of 1000 i.i.d. data points sampled from each strategy. Inset: the gap of mean reward between disentangled and entangled data increases as the number of generative factors increases. (D) Similar to (C), but using D-VAEs trained on the 3D-Shapes dataset (Burgess & Kim, 2018).

Methods

Similarity-based Novelty. We define the novelty $N^{(t)}$ of a stimulus s at time t as a nonlinearly decreasing function of its empirical frequency $p^{(t)}(s)$ (Barto et al., 2013; Xu et al., 2021), i.e. $N^{(t)}(s) = -\log p^{(t)}(s)$. Using a novel method called similarity-based novelty Becker et al. (2024), we compute the empirical frequency directly from an existing representation, using a mixture model: $p^{(t)}(s) = \sum_{j=1}^{N} w_j^{(t)} k_j(s)$,

where $k_1, ..., k_N$ are probability density functions ('components') each of which covers a subset of the stimulus space. The novelty weights $w^{(t)}$, constrained by $\sum_j^N w_j^{(t)} = 1$, determine the contribution of each component to $p^{(t)}(s)$ at time t. The weights are updated to maximize the likelihood of the observed sequence, yielding the iterative update rule: $\Delta w_j^{(t)} = \alpha^{(t)} \left(\frac{k_j(s)}{p^{(t-1)}(s)} - 1 \right) w_j^{(t-1)}$, where $\alpha^{(t)}$ is the learning rate.

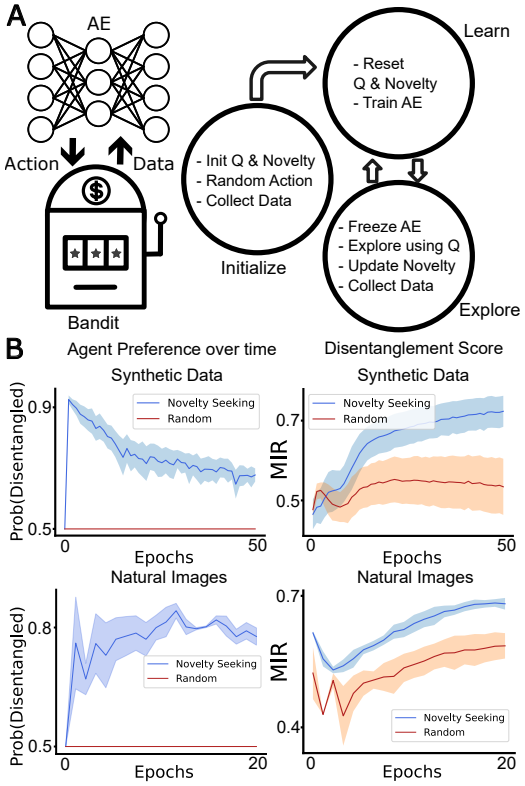

Figure 2: (A) Problem Setup & Training Procedure. Left: A multi-arm bandit setting where an AE (or D-VAE) selects between two actions to receive entangled or disentangled data points, without explicit rewards. Right: The training procedure: (1) Initialization: random data collection without any updates, (2) Learning: training AEs/D-VAEs on collected data, and (3) Exploration: while freezing AE/D-VAE, the agent explores using Q-values, updates novelty weights, and stores data in an experience buffer. (B) Novelty-Seeking vs. Random Exploration. Left column: novelty-seeking agents preferentially collect data generated by disentangled factors. Right column: novelty-driven exploration enables the network to learn more disentangled representation, measured by MIR score proposed in Whittington et al. (2023). Standard errors shown in shaded area.

Novelty in AEs & D-VAEs. For autoencoders (AEs, Fig.1B), we interpret the vector of latent activations in response to an input s as the vector of component activations $k(s) = (k_1(s), ..., k_N(s))$. This allows us to read-out similarity-based novelty directly from the network's activation. In D-VAEs, the encoder outputs $p(z|s)$ can serve as $k(s)$, where z is a categorical latent variable with N classes. Furthermore, if a stimulus s can be factorized into F independent factors, $s = (s^1, \ldots, s^F)$, then $N^{(t)}(s) = -\log p^{(t)}(s) = -\sum_{f=1}^F \log p^{(t)}(s_f) = -\sum_{f=1}^F N^{(t)}(s_f)$. Therefore, in disentangled AEs/D-VAEs, we can compute novelty separately for groups of neurons, each corresponding primarily to one factor. For AEs, latent neurons are partitioned into groups and trained with auxiliary losses that encourage intra-group correlation and discourage inter-group correlation. D-VAEs, with their categorical latents, naturally capture this separation.

Multi-armed Bandit Framework. We consider a modified multi-armed bandit where pulling an arm, instead of yielding an explicit reward, returns a data point s with an implicit novelty reward $r^{(t)} = N^{(t)}(s)$ (Fig.2A). The agent selects arms using Boltzmann exploration and updates the estimated value $Q(a_i)$ for arm i as a running average, where no future discounted reward is considered: $Q^{(t+1)}(a_i) = \frac{1}{C_i} r^{(t)} + \frac{C_i - 1}{C_i} Q^{(t)}(a_i)$, where C_i is the count of how many times arm i has been chosen. Each arm generates either disentangled data or entangled data (Fig.2). For 3D-shape data, the entangled arm produces samples where wall, object, and floor share the same hue, while the disentangled arm makes them independent.

Synthetic Data Generation for AEs Following Whittington et al. (2023), illustrated in Fig. 1A, we sample M independent variables $F \in [0,1]^M$ to generate data $y \in \mathbb{R}^{50}$ via a one-hidden-layer MLP with ReLU units, with all weights drawn i.i.d. from a normal distribution.

Results

Novelty Rewards Incentivize Disentangled Data Data generated by disentangled factors yield higher intrinsic novelty rewards in trained networks, for both synthetic data (AE, Fig.1C) and more realistic data from the 3D-Shapes data set (D-VAE, Fig.1D). We also note that this result still holds even if we increase the number of data-generating factors (AE, Fig.1C inset) since it allows more possible combinations of factors in the generated data.

Novelty-Seeking Shapes Disentangled Representations In the multi-armed bandit, agents seeking similarity-based novelty prefer the arm yielding data generated from disentangled factors, i.e. data that is most useful for disentangled representation learning. So novelty-seeking agents learn more disentangled representations (Fig.2B), in contrast to agents who explore randomly.

Discussion & Future Directions

We use similarity-based novelty Becker et al. (2024) to guide agents via representation-specific novelty signals, encouraging exploration of data with independent factors and supporting disentangled representation learning. Our results suggest intrinsic motivations may shape representation learning by influencing the sampling of the environment.

Acknowledgments

The author thanks Sophia Becker for useful discussion, and acknowledges the participation of Wulfram Gerstner.

References

Barto, A., Mirolli, M., & Baldassarre, G. (2013). Novelty or surprise? *Frontiers in Psychology*, *4*. Retrieved from https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2013.00907 doi: 10.3389/fpsyg.2013.00907

Becker, S., Modirshanechi, A., & Gerstner, W. (2024). Representational similarity modulates neural and behavioral signatures of novelty. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/early/2024/12/10/2024.05.01.592002.full.pdf doi: 10.1101/2024.05.01.592002

Bellemare, M. G., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., & Munos, R. (2016). *Unifying count-based exploration and intrinsic motivation.* Retrieved from https://arxiv.org/abs/1606.01868

Bengio, Y., Courville, A., & Vincent, P. (2014). *Representation learning: A review and new perspectives.* Retrieved from https://arxiv.org/abs/1206.5538

Burgess, C., & Kim, H. (2018). *3d shapes dataset.* https://github.com/deepmind/3dshapes-dataset/.

Friede, D., Reimers, C., Stuckenschmidt, H., & Niepert, M. (2023). Learning disentangled discrete representations. In *Machine learning and knowledge discovery in databases: Research track: European conference, ecml pkdd 2023, turin, italy, september 18–22, 2023, proceedings, part iv* (p. 593–609). Berlin, Heidelberg: Springer-Verlag. Retrieved from https://doi.org/10.1007/978-3-031-43421-1_35 doi: 10.1007/978-3-031-43421-1_35

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504-507. Retrieved from https://www.science.org/doi/abs/10.1126/science.1127647 doi: 10.1126/science.1127647

Hyvarinen, A., Khemakhem, I., & Morioka, H. (2023). *Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning.* Retrieved from https://arxiv.org/abs/2303.16535

Høydal, O. A., Skytøen, E. R., Andersson, S. O., Moser, M.-B., & Moser, E. I. (2019, April). Object-vector coding in the medial entorhinal cortex. *Nature*, *568*(7752), 400–404. Retrieved from https://doi.org/10.1038/s41586-019-1077-7 doi: 10.1038/s41586-019-1077-7

O'Keefe, J. (1976). Place units in the hippocampus of the freely moving rat. *Experimental Neurology*, *51*(1), 78-109. Retrieved from https://www.sciencedirect.com/science/article/pii/0014488676900558 doi: https://doi.org/10.1016/0014-4886(76)90055-8

Plumbley, M. (2003). Algorithms for nonnegative independent component analysis. *IEEE Transactions on Neural Networks*, *14*(3), 534-543. doi: 10.1109/TNN.2003.810616

Whittington, J. C. R., Dorrell, W., Ganguli, S., & Behrens, T. (2023). Disentanglement with biological constraints: A theory of functional cell types. In *The eleventh international conference on learning representations.* Retrieved from https://openreview.net/forum?id=9Z_GfhZnGH

Xu, H. A., Modirshanechi, A., Lehmann, M. P., Gerstner, W., & Herzog, M. H. (2021, 06). Novelty is not surprise: Human exploratory and adaptive behavior in sequential decision-making. *PLOS Computational Biology*, *17*(6), 1-32. Retrieved from https://doi.org/10.1371/journal.pcbi.1009070 doi: 10.1371/journal.pcbi.1009070