Functional and causal inductive biases for deep generative model representations

Michel Besserve, Technical University Braunschweig

Abstract:

Learning useful internal representations of the external world is an important goal for both artificial and biological intelligent systems, but how to do it “right” remains a largely open question. We investigate this question in Deep Generative Models (DGM), which can be conceived as Latent Variable Models (LVM) that learn to map a latent representation to the data using a feedforward deep neural network.

While DGM have become remarkably efficient at learning to synthesize samples of complex data (e.g. images) that “look” realistic, it is often less clear to which extent they can learn meaningful representations that can be useful beyond imitating the training data. One mathematically precise formulation of this problem is the question of identifiability of LVMs. Broadly construed, it addresses whether one can unambiguously learn the properties of a ground truth LVM, based only on the distribution of the data it generates.

It turns out that DGMs are not identifiable without additional assumptions about the data generative process, i.e. identification requires inductive biases. I will go over sets of assumptions that guaranty such identifiability by leveraging (1) constraints on the LVM’s function space; (2) constraints on the causal structure of the generative process, which decomposes it into mechanisms that can be selectively intervened on. I will moreover illustrate how these inductive biases are implicitly leveraged in two popular approaches, variational autoencoders and self-supervised learning, therefore providing justifications for their empirical success.

While these results provide insights into the conditions for learning faithful representations of the external world, they do not explicitly enforce a simplification of reality, key to human understanding. I will introduce ongoing work on causal model reduction that address this issue by mapping a high-dimensional model to a low-dimensional one with an interpretable causal structure. Finally, I will elaborate on the potential of these identifiable and simplified representations to assist humans in solving complex real-world problems.

References:

• Luigi Gresele*, Julius von Kügelgen*, Vincent Stimper and Bernhard Schölkopf, Michel Besserve_._Independent mechanism analysis, a new concept? NeurIPS 2021.

• Patrik Reizinger*, Luigi Gresele*, Jack Brady*, Julius von Kügelgen, Dominik Zietlow, Bernhard Schölkopf, Georg Martius, Wieland Brendel and Michel Besserve. Embrace the Gap: VAEs Perform Independent Mechanism Analysis. NeurIPS 2022.

- Simon Buchholtz, Michel Besserve and Bernhard Schölkopf. Function Classes for Identifiable Nonlinear Independent Component Analysis. NeurIPS 2022.
- Armin Kekić, Bernhard Schölkopf and Michel Besserve. Target Reduction of Causal Models. UAI 2024.
- Patrick Burauel, Frederick Eberhardt, Michel Besserve. Controlling for discrete unmeasured confounding in nonlinear causal models. CleaR 2025 (accepted).

Brief Bio:

Michel Besserve is Professor of Artificial Intelligence at the Computer Science department of TU Braunschweig since October 2024. He is particularly interested in how machine learning can help uncover the principles governing complex natural and artificial systems, notably through the lens of causality. Michel studied electrical engineering and applied mathematics at ENS Paris-Saclay. After investigating brain-computer interfaces during his doctorate at Paris-Saclay University, he joined as a postdoc the Cognitive Neurophysiology department at the Max Planck Institute for Biological Cybernetics of Tübingen in 2008, where he led modeling and data analysis efforts to understand distributed information processing in the brain till 2019, notably in relation to vision and episodic memory. From 2015, Michel also started investigating the principles of causal machine learning in the department of Empirical Inference at the Max Planck Institute for Intelligent Systems of Tübingen, where he led theoretical work on Machine Learning for Complex Systems, motivated by applications ranging from deep generative models to sustainability.