Discovering physical concepts with neural networks

Raban Iten*
Institute for Theoretical Physics
ETH Zürich
8093 Zürich, Switzerland
itenr@itp.phys.ethz.ch

Tony Metger*
Institute for Theoretical Physics
ETH Zürich
8093 Zürich, Switzerland
tmetger@ethz.ch

Henrik Wilming
Institute for Theoretical Physics
ETH Zürich
8093 Zürich, Switzerland
henrikw@phys.ethz.ch

Lídia del Rio
Institute for Theoretical Physics
ETH Zürich
8093 Zürich, Switzerland
lidia@phys.ethz.ch

Renato Renner
Institute for Theoretical Physics
ETH Zürich
8093 Zürich, Switzerland
renner@phys.ethz.ch

Abstract

We introduce a neural network architecture that models the human physical reasoning process: given an observation of a physical system, an encoder compresses it into a simple latent representation; a decoder is asked questions about the physical system that must be answered based only on the latent representation. This is analogous to the human process of describing a physical system by a few characteristic properties that suffice to predict the future behaviour. For a variety of simple physical systems, the network finds, in a fully unsupervised way, the physically relevant parameters, exploits conservation laws to make predictions, and can be used to gain conceptual insights — for example, the network allows us to recover the heliocentric model of the solar system only from observations made from Earth. On a theoretical level, we formalize the idea of a "simple" physical representation and we analyze it using methods from differential geometry. Our work provides a first step towards understanding the representations of physical data used by deep neural networks.

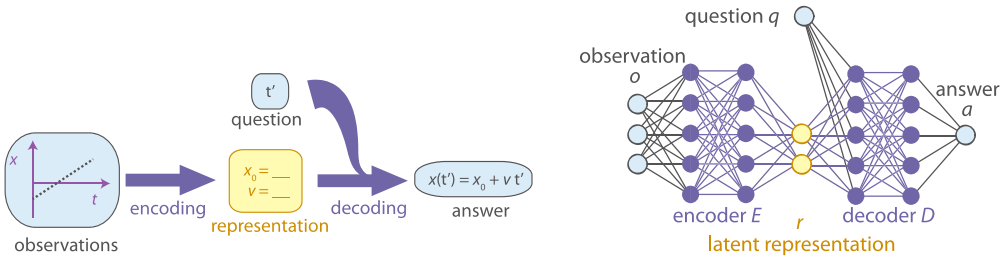
In an overview of challenges for artificial intelligence in the near future [1], Lake *et al.* wrote:

> "For deep networks trained on physics-related data, it remains to be seen whether higher layers will encode objects, general physical properties, forces and approximately Newtonian dynamics."

In this work, we provide a first step towards solving this question. We show that, in the case of simple systems, neural networks can be used to discover the physical properties and physical concepts we are used to from physics textbooks from experimental data without providing any prior knowledge about mathematics or physics. This is in contrast to most of the previous work that applies neural networks as black box predictors or uses some mathematical or physical prior-knowledge (for a detailed comparison and references, see our paper at arXiv:1807.10300).

*These authors contributed equally to this work.

For the full technical version, see arXiv:1807.10300. The source code is available at `https://github.com/eth-nn-physics/nn_physical_concepts`.



(a) Human learning.

(b) Neural network structure for *SciNet*.

Figure 1: **Learning physical representations. (a)** A physicist compresses experimental observations into a simple representation. When later asked any question about the physical setting, the physicist should be able to produce a correct answer using only the representation and not the original data. For example, the observations may be the first few seconds of the trajectory of a particle moving with constant speed; the representation could be the parameters "speed v" and "initial position x_0" and the question could be "where will the particle be at a later time t'?" **(b)** In our neural network, observations are encoded as real parameters fed to an encoder (a feed-forward neural network), which compresses the data into a latent representation. The question is again encoded in a number of real parameters, which, together with the representation, are fed to the decoder network to produce an answer. Note that the number of layers and neurons depicted is just an example, and not representative.

1 Network structure and training

We introduce a neural network architecture, which we call *SciNet* for brevity, which mimics a physicist's modelling process (Figure 1a), and apply it to study various physical scenarios. For a purely input-output (black box) analysis, the modelling process of *SciNet* can be seen as a map $F : \mathcal{O} \times \mathcal{Q} \to \mathcal{A}$ from the sets of possible observations $\mathcal{O}$ and questions $\mathcal{Q}$ to the set of possible answers $\mathcal{A}$. We can split this map into an encoder $E : \mathcal{O} \to \mathcal{R}$ mapping the original observation to a compressed latent representation $\mathcal{R}$ followed by a decoder $D : \mathcal{R} \times \mathcal{Q} \to \mathcal{A}$ that takes the representation and the question to produce an answer.[2] The corresponding network structure is shown in Figure 1b. A similar network architecture was recently applied for scene representation and rendering [3]. We use fully connected feed-forward neural networks to implement the encoder and the decoder of *SciNet*. It is this decomposition, $F(o, q) = D(E(o), q)$, that will allow us to interpret the network's learned representation, by analyzing how it changes as we tweak known parameters of the setting.

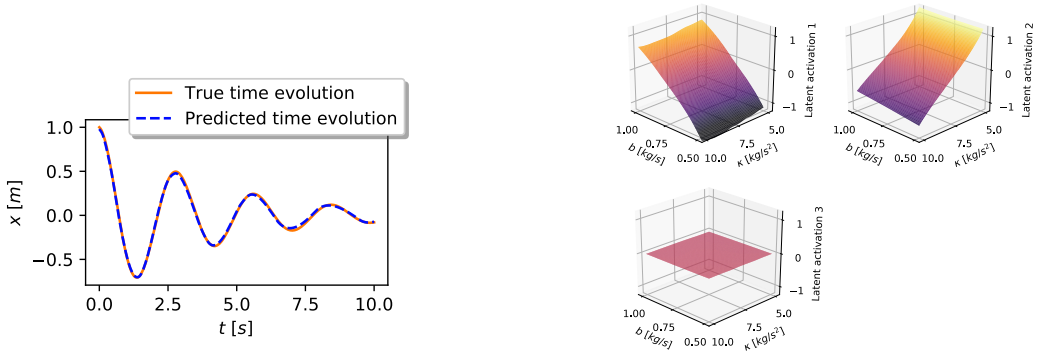
We train *SciNet* with data samples of the form $(o, q, a_{\text{cor}}(o, q))$, where the observation o and question q are chosen from the sets $\mathcal{O}$ and $\mathcal{Q}$ of all possible observations and questions, respectively, and where $a_{\text{cor}}(o, q)$ denotes the correct reply to question q given observation o. The structure of the training data and of the network does not fall into the standard categorisation of "unsupervised" versus "supervised". However, *SciNet* can be regarded as a generalisation of the idea of autoencoders for all examples presented here, since the answers to the questions correspond to subsets of collected measurement data and do not require human labelling. In this sense, the training is unsupervised.

2 Minimal uncorrelated representation

We use methods of disentangling variational autoencoders [2, 4–6] to encourage our network architecture to learn "simple" latent representations, which we formalize as follows. An *uncorrelated (sufficient) representation* for the input data set $\mathcal{O}$ with respect to a set of questions $\mathcal{Q}$ and described by a random variable R (whose distribution is determined trough the distribution of the input data and the encoder mapping $E : \mathcal{O} \to \mathcal{R}$) is defined by the following properties:

1. **Sufficient (with smooth decoder)**: There exists a smooth map $D : \mathcal{R} \times \mathcal{Q} \mapsto \mathcal{A}$ such that $D(E(o), q) = a_{\text{cor}}(o, q)$ for all possible observations $o \in \mathcal{O}$ and questions $q \in \mathcal{Q}$.

[2] For our implementation, we use stochastic mappings as it is common for variational autoencoders [2].



(a) Trajectory prediction of *SciNet*.

(b) Representation learned by *SciNet*.

Figure 2: **Damped pendulum**. *SciNet* is fed a time series of the trajectory of a damped pendulum. It learns to store the two relevant physical parameters, frequency and damping, in the representation, and makes correct predictions about the pendulum's future position. **(a)** Here, the spring constant is $\kappa = 5\text{kg/s}^2$ and the damping factor is $b = 0.5\text{kg/s}$. *SciNet's* prediction is in excellent agreement with the true time evolution. **(b)** The plots show the activations of the three latent neurons of *SciNet* as a function of the spring constant κ and the damping factor b. The first two neurons store the damping factor and spring constant, respectively. The activation of the third neuron is close to zero, suggesting that only two physical variables are required. On an abstract level, learning that one activation can be set to a constant is encouraged by searching for uncorrelated latent variables, i.e., by minimizing the common information of the latent neurons during training.

2. **Uncorrelated**: The latent variables, described by random variables $R_1, R_2, \ldots, R_{|R|}$, are mutually independent.

We define a *minimal uncorrelated representation* R as an uncorrelated (sufficient) representation with a minimal number of latent parameters $|R|$, i.e., there does not exist a minimal uncorrelated representation with less than $|R|$ latent neurons. This formalizes what we consider to be a "simple" representation of (physical) data.

Without the assumption that the decoder is smooth, it would, in principle, always be sufficient to have a single latent variable, since a real number can store an infinite amount of information. Hence, methods from standard information theory, like the information bottleneck [7], are not the right tool to give the number of variables a formal meaning. In the full technical version, we use methods from differential geometry to show that the number of variables $|R|$ in a minimal (sufficient) representation corresponds to the number of relevant degrees of freedom in the observation data required to answer all possible questions.

3 Results

We train *SciNet* with raw (simulated) experimental data from a variety of simple systems in classical and quantum mechanics and extract conceptual information from the learned representation:

1. The representation stores the physically relevant parameters, like the frequency of a pendulum, which it recovers from a time series of its position (see Figure 2).
2. *SciNet* finds and exploits conservation laws: it stores the total angular momentum to predict the motion of two colliding particles.
3. Given measurement data of a simple quantum mechanical system, *SciNet* finds a minimal representation of it, correctly recognizing the underlying degrees of freedom.
4. Given a time series of the positions of the Sun and Mars as observed from Earth, *SciNet* discovers the heliocentric model of the solar system — that is, it encodes the data into the angles of the two planets as seen from the Sun.

4 Related work

Independently of our work, physical variables were extracted in an unsupervised way from time series data of dynamical systems [8]. The network structure used in [8] is built on interaction networks [9–11] and it is well adapted to physical systems consisting of several objects interacting in a pair-wise manner. The prior knowledge included in the network structure allows the network to generalise to situations that differ substantially from those seen during training. Where the focus in [8] is on good generalization properties, our aim was to discover physical properties and concepts without putting prior knowledge about the physical system into the network structure.

5 Conclusion

The main aim of this work is to show that neural networks can be used to discover physical concepts without any prior knowledge. To achieve this goal, we introduced a neural network architecture that generalizes autoencoders. The examples illustrate that this architecture allows us to extract physically relevant data from experiments, without imposing further knowledge about physics or mathematics.

Moreover, we formalized the notion of a "simple representation" as a minimal uncorrelated representation that is sufficient with respect to a fixed set of questions and investigated its properties with methods from information theory and differential geometry. This generalizes the idea of learning a full representation of some given data and we expect it to be applicable to different tasks in representation learning.

References

[1] Lake, B.M. et al. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017. ISSN 0140-525X, 1469-1825. doi: 10.1017/S0140525X16001837. URL https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/building-machines-that-learn-and-think-like-people/A9535B1D745A0377E16C590E14B94993.

[2] Kingma, D.P. and Welling, M. Auto-encoding variational bayes. 2013. doi: 10.1051/0004-6361/201527329. URL https://arxiv.org/abs/1312.6114.

[3] Eslami, S.M.A. et al. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, June 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aar6170. URL http://science.sciencemag.org/content/360/6394/1204.

[4] Higgins, I. et al. beta-vae: learning basic visual concepts with a constrained variational framework. *ICLR*, 2017. URL https://openreview.net/references/pdf?id=Sy2fzU9gl.

[5] Kim, H. and Mnih, A. Disentangling by factorising. 2018. URL https://arxiv.org/abs/1802.05983.

[6] Burgess, C.P. et al. Understanding disentangling in beta-vae. *NIPS*, 2018. URL https://openreview.net/pdf?id=Sy2fzU9gl.

[7] Tishby, N., Pereira, F.C. and Bialek, W. The information bottleneck method. April 2000. URL http://arxiv.org/abs/physics/0004057.

[8] Zheng, D. et al. Unsupervised learning of latent physical properties using perception-prediction networks. 2018. URL http://arxiv.org/abs/1807.09244.

[9] Battaglia, P. et al. Interaction networks for learning about objects, relations and physics. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4509–4517, 2016. URL http://dl.acm.org/citation.cfm?id=3157382.3157601.

[10] Chang, M.B. et al. A compositional object-based approach to learning physical dynamics. December 2016. URL http://arxiv.org/abs/1612.00341.

[11] Raposo, D. et al. Discovering objects and their relations from entangled scene representations. February 2017. URL http://arxiv.org/abs/1702.05068.