Generalizing with overly complex representations

Anonymous Author(s)
Affiliation
Address
email

Abstract

Representations enable cognitive systems to generalize from known experiences to the new ones. Simplicity of a representation has been linked to its generalization ability. Conventionally, simple representations are associated with a capacity to capture the structure in the data and rule out the noise. Representations with more flexibility than required to accommodate the structure of the target phenomenon, on the contrary, risk to catastrophically overfit the observed samples and fail to generalize to new observations. Here, I computationally test this idea by using a simple task of learning a representation to predict unseen features based on the observed ones. I simulate the process of learning a representation that has a lower, matching, or higher dimensionality than the phenomenon it accounts for. The results suggest that the representations of the highest dimensionality consistently generate the best out-of-sample predictions despite perfectly memorizing the training observations. These findings are in line with the recently described "double descent" of generalization error – an observation that many learning systems generalize best when overparameterized (when their representational capacity far exceeds the task requirements).

1 Introduction

Simplicity is often viewed as a core cognitive principle. Humans and scientists alike prefer simple explanations [2, 19, 15, 16]. Simple representations and algorithms are claimed to govern robust behavior and successful perception, learning, and decision making [11, 6, 9]. Simple representations are viewed as more interpretable, memorable, resource-efficient, communicable, among of their other values. Perhaps one of the core virtues of simple representations is their superior generalization. Representations that "compress" the multidimensional world to the right extent can capture the essential structure in the agent's experiences and eliminate the noise, while more complex representations are notorious for overfitting the observations and amplifying the noise in them, thereby leading to arbitrarily poor generalization. Thus, to find patterns among noise, a successful cognitive system must compress information coming from the complex environment by constructing simple representations.

This link between simplicity and generalization has been studied formally [14]. The typical scenario of finding a representation of optimal complexity for a given problem juxtaposes representation's ability to account for the observed data/encountered situations and its ability to generalize to unseen data/situations resulting in the 'bias-variance' tradeoff [10]. Even though increasing representation complexity (e.g. adding more flexible parameters to it) allows it to account for the training data better, it is generally undesirable since more flexible parameters open up a space of solutions which catastrophically overfit the training data and miss the structure in them. This bias-variance tradeoff has been widely studied in statistical learning, resulting in the criteria that explicitly punish models for a number of their flexible parameters (such as BIC and AIC), aiming to find the perfect balance when a model is not too simple to misrepresent the data and not too complex to catastrophically

Submitted to Information-Theoretic Principles in Cognitive Systems Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

overfit the noise. The bias-variance tradeoff has been used to justify both the simplicity bias among scientists as well as simplicity as a principle that governs successful cognitive systems [14, 5, 18, 9].

Here, I study the influence of representation complexity on generalization in a simple prediction task. I test both the representations that compress the data to different extents and the representations that possess way more resources than the task requires. I find that the overly complex representations are the ones that best generalize to the unseen data.

2 Experiment: learning predictive representations

I consider a general representation learning scenario: an agent learns to predict the unseen properties of the data generated by the D-dimensional world with some structure.

2.1 Representation

Representation to be learned is formalized as a simple neural autoencoder with one hidden layer (activation functions: ReLU-ReLU). This autoencoder learns to predict masked (unknown) dimensions of an observation based on the recorded (known) ones [13]. I vary the number of hidden units [1,2,3,4,6,8,10,16,32,100,250,500,1000] as a measure of simplicity/complexity bias of the representation. All representations are trained with stochastic gradient descent with no regularization until convergence.

2.2 Representation complexity

The definitions of representation “complexity" vary and are often vague. Criteria for representation’s complexity can include the number of assumptions it makes about the world, the amount of flexible parts/parameters in it, human judgements of its’ simplicity (e.g. 5 is simpler than 4.9999), as well as informational and algorithmic complexity on the formal side. Here, I vary the complexity of a representation as formalized by a number of units in the hidden layer of the autoencoder (Fig.1E). This manipulation correlates with other indicators of complexity, such as number of flexible parameters and representation’s ability to perfectly fit the training data. Note that the “necessary" representation complexity for a given world cannot simply be judged based on the number of free parameters/hidden units as compared to the world dimensionality. Instead, I will use both “the ability to perfectly fit the training data" and the hidden layer dimensionality as markers of representation complexity relative to the world’s complexity.

2.3 World

For each simulation, I randomly seed a mixture of N [1,10,100] multivariate gaussian distributions that span across D [4,8,100] dimensions for the agent to learn about (Fig.1A and Fig.1B). These parameters of the world determine its complexity.

2.4 Learning

For learning, I sample 300 training observations from the world. Each observation has R [$\frac{1}{4}$D,$\frac{1}{2}$D] recorded dimensions and D-R masked dimensions. The agent learns to predict values along the unknown dimensions with supervision (see Fig.1C and Fig.1D). Training performance is evaluated with respect to the average absolute error of the predicted dimensions. Note that increasing the number of world’s dimensions (D) makes the task less noisy, as the agent gets more information to base the predictions upon.

2.5 Evaluation

For evaluation, I sample 10000 observations from the world and look at the mean absolute error of the predicted dimensions.

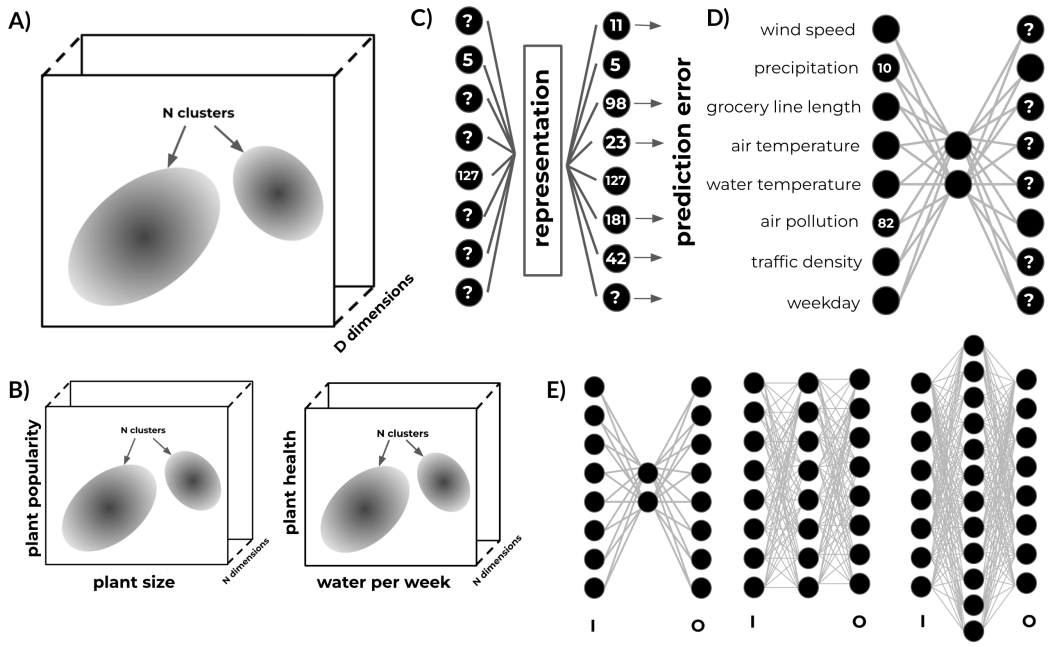

Figure 1: Model. A) The world, B) Example interpretations of the world: as a structure in the features (left) or a structure in the outcomes of possible causal interventions, C) The task: learning a representation for predicting masked (unknown) dimensions of an observation based on the known ones, D) Example interpretation of the prediction task, E) three types of representations: bottleneck (left), matching dimensionality of the world (center), and extending world's dimensionality (right).

3 Results

I conducted 12 simulation runs per condition, resulting in 2808 observations in total.

Across the world complexity conditions, increasing the representation dimensionality predictably led to the better fit for the training observations (Fig.2A). Surprisingly, there was a nuanced relationship between the representation complexity and its out-of-sample (generalization) performance (Fig.2B). For smaller worlds, the generalization performance almost monotonically decreased with the increase of representation capacity. For the world with the highest dimensionality (D=100), the generalization error followed the traditional U-curve up until the dimensionality that allowed the representation to perfectly fit the training data. Increasing the representation dimensionality from this point led to decrease in generalization error. Increasing the dimensionality of the representation beyond the capacity necessary to overfit the training data also improved the reliability of generalization performance, measured as a standard deviation of generalization performance across different runs. Thus, more complex representations were more predictable in their generalization performance, whereas simpler representations ore representations that matched the dimensionality of the world produced a wider range of generalization errors. Over time, the generalization performance of more complex representations steadily decreased with more training observations, whereas simpler representations' generalization performance was sometimes harmed by more observations (Fig.2C). Overall, the most complex representation (H=1000) exhibited both the best training performance and the best generalization performance.

4 Discussion

Simplicity bias governs both the theory-building process in cognitive science and cognitive scientists' assumptions about successful cognitive systems. We assume that to capture patterns in the multi-dimensional and noisy world one must learn relatively simple representations. The present results speak against this intuition: to learn a predictive representation that generalizes best one should

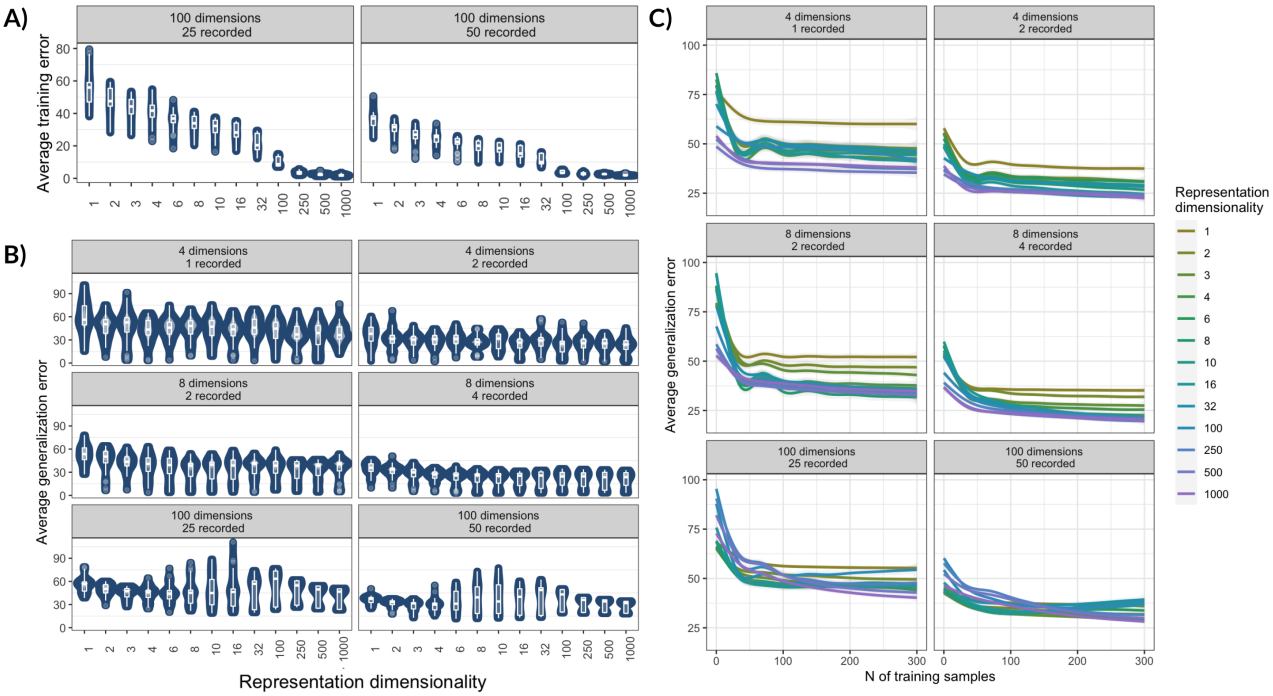

Figure 2: Results of learning representations of different complexity. A) Average training error in the 100-dimensional world condition, B) Average generalization error, C) Average generalization error over the number of training samples.

learn a representation that is even more complex than the problem itself; both for simpler and more complex problems. Moreover, the results suggest that the ability of the representation to perfectly accommodate the observed examples (overfitting) is not necessarily harmful for generalization [12].

These counterintuitive results corroborate a series of recent findings showing the benefits of overparameterization for generalization in a wide range of models and tasks [8, 17, 3, 4, 20]. Increasing the model complexity beyond the point when it has enough resources to perfectly fit the training data often leads to improvements in generalization. The classic bias-variance tradeoff characterizes generalization dynamics well only for learning systems in the underparameterized regime. Increasing the model complexity beyond the point that allows the representation to perfectly fit the training data leads to the second decrease in generalization error, often resulting in better generalization performance of a sufficiently overparameterized model compared to the model of perfect complexity in the underparameterized regime (bottom of the U-curve).

Importantly, the results presented here do not indicate that humans or other cognitive systems necessarily operate with complex representations. Instead, the current work challenges only one of the commonly assumed virtues of simple representations – their superior generalization ability. Cognitive systems might still prefer simpler representation for their other virtues, such as interpretability and compactness despite their inferior generalizability. Moreover, the modeling approach presented here has a number of limitations. First of all, the representation and the world complexity need further articulation. Second, the current approach does not allow testing representation's performance in the extrapolation settings (e.g. generalizing outside of the learned distribution). Finally, the potential goals of cognitive systems extend well beyond the prediction task used here: for example, we often need to explain or control the world [1] – representations that would best support these aims do not necessarily align with the representations which enable best prediction [7].

Uncovering the non-trivial relationship of the representation complexity and its generalization ability is essential for understanding cognitive systems. As cognitive scientists, we might need to abandon the simplicity bias if we want to build more generalizable theories of cognitive systems: successful cognitive systems and theoretical accounts of such systems might be constructed with much more complex representations than we thought.

References

[1] Anonymous author(s). How predictive minds explain and control dynamical systems. 2022.

[2] Alan Baker. Simplicity. 2004.

[3] Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

[4] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[5] Henry Brighton and Gerd Gigerenzer. How heuristics handle uncertainty. In *Ecological rationality: Intelligence in the world*, pages 33–60. Oxford University Press, 2012.

[6] Nick Chater and Paul Vitányi. Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences*, 7(1):19–22, Jan 2003.

[7] Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204, 2013.

[8] Yehuda Dar, Vidya Muthukumar, and Richard G. Baraniuk. A farewell to the bias-variance tradeoff? an overview of the theory of overparameterized machine learning. *arXiv preprint arXiv:2109.02355*, 2021.

[9] Jacob Feldman. The simplicity principle in perception and cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(5):330–340, 2016.

[10] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

[11] Gerd Gigerenzer and Peter M Todd. *Simple heuristics that make us smart*. Oxford University Press, USA, 1999.

[12] Uri Hasson, Samuel A Nastase, and Ariel Goldstein. Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron*, 105(3):416–434, 2020.

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 16000–16009, 2022.

[14] Kevin T. Kelly. *Simplicity, truth, and probability*, page 983–1024. Elsevier, 2011.

[15] Daniel RB Little and Richard Shiffrin. Simplicity bias in the estimation of causal functions. In *Proceedings of the annual meeting of the cognitive science society*, volume 31, 2009.

[16] Tania Lombrozo. Simplicity and probability in causal explanation. *Cognitive Psychology*, 55(3):232–257, Nov 2007.

[17] Neil Mallinar, James B. Simon, Amirhesam Abedsoltan, Parthe Pandit, Mikhail Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: A taxonomy of overfitting. *arXiv preprint arXiv:2207.06569*, 2022.

[18] Hansjörg Neth and Gerd Gigerenzer. *Heuristics: Tools for an uncertain world*, page 1–18. Wiley Online Library, 2015.

[19] Elliott Sober. *Ockham's razors*. Cambridge University Press, 2015.

[20] Peter Sollich and Anders Krogh. Learning with ensembles: How overfitting can be useful. *Advances in neural information processing systems*, 8, 1995.