Patterns

CellPress
OPEN ACCESS

Perspective

Methods for identifying emergent concepts in deep neural networks

Tim Räz[1,*]
[1]University of Bern, Institute of Philosophy, Länggassstrasse 49a, 3012 Bern, Switzerland
*Correspondence: tim.raez@posteo.de
https://doi.org/10.1016/j.patter.2023.100761

THE BIGGER PICTURE Deep neural networks (DNNs) are a key technology in many scientific domains. However, their inner workings are still insufficiently understood. The hope is that if we understand the inner working of DNNs, we can make sure that they do what we want, which is not necessarily the case today. One path to understanding is to associate the inner structure of models with concepts that humans can understand. Computer scientists have proposed methods to do exactly this. The present perspective reviews and discusses these methods from a philosophical perspective. For example, it has been argued that DNNs learn concepts that are relevant to their task from scratch: a model can learn something akin to the concept of "airplane" to recognize images of hangars. However, saying that the model learned the concept "airplane" requires human interpretation. In sum, while existing methods are promising first steps, we are still quite far from understanding the inner workings of DNNs.

Concept: Basic principles of a new data science output observed and reported

SUMMARY

The present perspective discusses methods to detect concepts in internal representations (hidden layers) of deep neural networks (DNNs), such as network dissection, feature visualization, and testing with concept activation vectors (TCAV). I argue that these methods provide evidence that DNNs are able to learn non-trivial relations between concepts. However, the methods also require users to specify or detect concepts via (sets of) instances. This underdetermines the meaning of concepts, making the methods unreliable. The problem could be overcome, to some extent, by systematically combining the methods and by using synthetic datasets. The perspective also discusses how conceptual spaces—sets of concepts in internal representations—are shaped by a trade-off between predictive accuracy and compression. I argue that conceptual spaces are useful, or even necessary, to understand how concepts are formed in DNNs but that there is a lack of method for studying conceptual spaces.

INTRODUCTION

There is a well-known story of how deep neural networks (DNNs) predict classes in an image classification task[1,2]: In the hidden layers of DNNs, progressively abstract concepts are represented. Take a model that classifies animals such as cats, dogs, and cows. According to the story, the model detects low-level concepts such as colors and textures in the first layers. In intermediate layers, the model detects higher-level concepts, such as body parts (eyes, ears) or complex textures (fur), by composing low-level concepts. In the final layer, the model detects animals by composing higher-level concepts. Importantly, these concepts are emergent, which means that they are not hardwired into the models and do not correspond to the labeled classes but are acquired through the learning process.

To verify this story, computer scientists examine the internal representations (hidden layers) of DNNs and try to detect concepts supposedly represented there. The main goal of the present perspective is to review and discuss methods to do this, in view of philosophical work on concepts. The perspective focuses on two issues. First, what do these methods tell us about concepts that are supposedly represented in a DNN, and how reliable are they? I argue that while these methods provide evidence that DNNs are able to learn non-trivial relations between concepts, the reliability with which concepts are detected runs into well-known philosophical problems. Second, how are conceptual spaces—sets of concepts in internal representations—shaped by the classes to be predicted and by the representational capacities of DNNs? I argue that while answering this question is important, as of now, there are few existing methods that tackle this important question.

BACKGROUND

Concepts

Before discussing methods to detect concepts in DNNs, it should be specified under what conditions DNNs possess concepts. There are various philosophical theories of concepts.[3] Here, I use an undemanding theory, or explication, of concept possession, which does not assume that concept possession requires mental states or consciousness or that concepts are abstract objects.[4] Rather, concepts are taken to be associated with abilities. An important distinction is between the extension and the meaning (intension) of a concept—the distinction goes back to Frege.[5] The extension of a concept is the collection of entities falling under it. DNNs show possession of concept extensions through activation patterns in their hidden or output layers. DNNs are trained on (partial) extensions, viz. labeled instances. When humans label instances, they do this on the basis of prior knowledge about the instances—the meaning of a concept. In DNNs, the possession of the meaning of a concept encompasses the representation of some inferential relations, e.g., a cat is an animal, has four legs, a head, and fur (usually), and so on. Both the extension and the meaning of a concept are relevant.

There is evidence that DNNs learn non-trivial inferential relations between concepts that are not the predicted classes, but which emerge as a function of learning to predict the classes. Also, DNNs apparently learn concepts that are relevant to predict several classes. To use the example of classifying animals, in order to classify cats and dogs, a DNN may learn concepts such as fur, head, paw, eye, and so on that are shared by cats and dogs. This issue has been explored for some time. DNNs apparently exploit that many classes share low-level features in order to improve generalization.[1] If we can confirm these findings, it means that DNNs are in fact able to learn non-trivial inferential relations between concepts, which implies that DNNs learn information about meaning rather than extensions. However, one of the main challenges with existing methods is that they require users to specify or recognize concepts with the help of (small) sets of instances, or extensions, and this may make it hard for users to determine which concepts are actually represented.

One could argue that the exercise of examining concepts in internal representations is superfluous because the predictive successes of DNNs show that they are able to automatically identify predictively salient concepts. If this were not the case, DNNs would not be able to generalize as well as they do. However, DNNs are not always successful; there are known failure modes such as adversarial examples.[6] Also, predictively successful models are not necessarily models that represent their target system adequately; this is true for scientific models as well as for DNNs, as philosophers know.[7,8] Thus, investigating whether and how concepts are represented in DNNs is a worthwhile enterprise.

Conceptual spaces

DNNs may be able to learn concepts that are relevant to predict more than one class. Other factors may shape how concepts are represented in DNNs as well. First, the predicted classes may not share certain concepts and may be mutually exclusive (to some extent). In the example of animal classification, in order to classify cows, a useful concept to be learned by a DNN may be horns. This concept does not contribute positively to the classification of cats and dogs because cats and dogs are not horned. Second, the concepts populating the internal representation take up some space in the internal representation: they are in competition for a finite amount of representational space. This competition may lead to compression and thus shape the internal representation of all concepts. Third, individual concepts may be compressed as well: if the representation of a concept contains predictively irrelevant details, this will lead to overfitting.

All these factors contribute to the formation of a conceptual space, the set of concepts in an internal representation of a DNN. Conceptual spaces are formed as a function of both the set of predicted classes and the representational capacity of the DNN. There have been some studies of how conceptual spaces are formed in DNNs, but less is known about this than about the emergence of individual concepts. Below, we will see some evidence for compression due to competing concepts, and I will argue that understanding conceptual spaces may be necessary to understand how individual concepts are represented in DNNs.

Limitations

Not all aspects of the internal representation of a DNN have an interpretation in terms of concepts. An internal representation may not relate to any concept in that (1) there may be a failure to represent, as in adversarial examples,[6] or (2) what is represented may not be accessible or comprehensible to humans and therefore not correspond to a concept.[9] Note that adversarial examples may also constitute predictively useful patterns, or artifacts.[10] Cases (1) and (2) are examples of non-conceptual content of an internal representation. Understanding the scope and limits of non-conceptual content is important, but the following discussion will focus on the modes of representation of concepts that can be grasped by humans. Also, we will focus on post hoc methods to extract concepts from trained models, excluding methods like concept whitening[11] or concept bottleneck[12] that modify the architecture of DNNs to enhance interpretability.

DETECTING EMERGENT CONCEPTS

In this section, I review empirical work on the emergence of concepts in the internal representation of DNNs. I focus on concepts that are relevant to several of the predicted classes because such concepts indicate non-trivial inferential relations.

Network dissection

Network dissection by Bau et al.[13] proposes to detect concepts associated with individual neurons in convolutional neural networks (CNNs). Specifically, the emergence of object detectors in scene classifiers is examined. For example, according to Bau et al., the CNN learned the concept "airplane" in the process of classifying "airfield" and "hangar." Concepts are detected automatically by matching the region of the input that maximizes activation of a neuron with a region associated with a concept given by an image segmentation method. The authors

found that many concepts are important for the classification of multiple scenes.

Network dissection has several advantages: it is automatic and allows for a quantitative evaluation of similarity and for visual inspection of image regions. A drawback is that the image segmentation method can only detect a fixed, limited set of concepts. If a concept is not included in this set, it cannot be detected. Therefore, network dissection falls prey to a version of the "bad lot" argument by van Fraassen[14,15]: if we explain scientific evidence (here: region with high activation by a neuron) using the best hypothesis from a limited set (here: concepts from image segmentation), it is not clear that the best hypothesis is also true; the concept detected by the CNN may simply not be in the scope of the image segmentation method. This problem can be overcome by letting users inspect regions with high activation in order to identify the concept. However, a concept we associate with a certain region need not be the concept used by the CNN, because the same region of an image is usually associated with various meanings. A CNN may see a shiny tube with horizontal bars where we see a plane. This is a version of the so-called indeterminacy of reference described by Quine.[16,17]

Feature visualization

Feature visualization by Olah et al.[18] proposes to detect concepts by constructing input instances that maximize the activation of neurons (or other parts) of CNNs. The method generates synthetic images that maximize activation of a neuron. Olah et al. note that direct, unregularized optimization can lead to degeneracies (akin to adversarial examples) and that different kinds of regularization have to be used to obtain natural-looking images. Feature visualization can be used to show that low-level concepts are combined and form higher-level concepts, e.g., a car detector is assembled from features like windows, car body, and wheels.[19]

Feature visualization has the advantage that one does not need to infer the meaning of a concept from a set of instances. Rather, it provides a single visualization (or a few). However, a user still needs to determine meaning from the visualization. As Olah et al. acknowledge, while many visualizations have a rather clear semantic interpretation, some visualizations appear to have a mixed meaning (so-called polysemantic neurons, more on these below), and some visualizations have no discernible meaning at all. Thus, the indeterminacy of the reference is an issue here as well. Furthermore, the visualizations depend on the choices made in optimization, the regularizations in particular, which may introduce artifacts. The use of optimization raises further concerns. For example, the method could get stuck in a local optimum. Optimization in DNNs is not very well understood from a theoretical point of view,[20,21] and the possibility of local optima makes the method susceptible to a bad lot-type argument.

Testing with concept activation vectors (TCAV)

TCAV by Kim et al.[22] is a method to examine how strongly a user-defined concept is associated with a predicted class in a particular layer. Concepts are defined extensionally by the user through a set of input examples of that concept and a set of random counterexamples. The concept activation vector of a layer is the vector normal to the hyperplane that best separates the activations of examples and counterexamples. One can test how strong the association of this concept with a predicted class is by measuring how well its vector aligns with the vector of that class. Kim et al. claim that DNNs learn emerging concepts with considerable accuracy. Classifiers of low-level concepts (colors, shapes) achieve high accuracy in early layers, while more complex concepts (race, gender) achieve higher accuracy in later layers. Note that other researchers have explored the activation of layers with linear classifiers.[23]

The main advantage of TCAV is that it allows users to choose the concepts to be detected through customized sets of examples. TCAV thereby overcomes, to some extent, the problem of indeterminacy of reference: in principle, there is no limit on the number and variety of instances to define a concept extensionally. However, the extensional definition of concepts nevertheless limits the control on the meaning of the concept being defined. Also, there are practical limitations on the instances used. A further drawback of the method is its limitation to testing for linear information in the layers.

Non-local representation of concepts

The above methods differ in how they propose to detect concepts, but they also vary in where they take concepts to be represented (in single neurons, entire layers, spread over several layers). It is known that concepts are not (only) represented by individual neurons but have distributed representations. There is evidence that the representation of concepts is not limited to single layers. Yosinski et al.[24] examined concept representations from the perspective of transfer learning. They found that feature representation in intermediate layers is distributed over consecutive layers: freezing only a portion of consecutive intermediate layers led to a worse performance than freezing all intermediate layers in question. This is indirect evidence that the relevant concepts are distributed over these layers.

CONCEPTUAL SPACES

In this section, I discuss how conceptual spaces arise as a function of the predicted classes and of compression. The discussion is more speculative than in the last section because there is less empirical work on the global perspective of conceptual spaces.

Polysemantic neurons

Feature visualization provides indirect, local evidence for competing concepts (see above). If concepts are disjunctive and in competition for representational space, say, in a layer of a DNN, then one observable consequence may be that some concepts have imperfect representations and become mixed. This phenomenon has been observed by Goh et al.[25] They find that while many neurons maximize activation for a single, identifiable concept, so-called polysemantic neurons are composites of different, seemingly unrelated concepts, e.g., a neuron representing a mix of cats and cars. Goh et al. point out that one possible explanation of this sort of disjunctive neuron is that they could make "concept packing more efficient."[25] This idea is discussed in more detail by Olah et al.[19] as the superposition hypothesis.

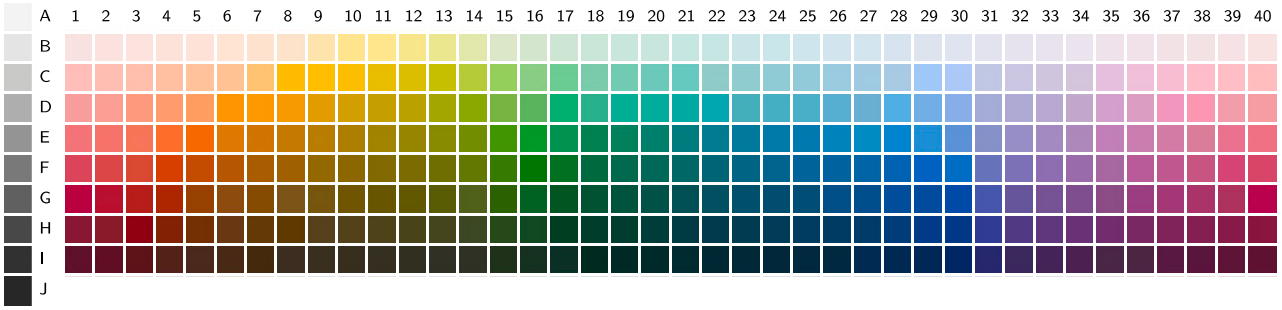

Figure 1. The WCS palette
The WCS palette, reproduced from Zaslavsky et al., arXiv.1808.03353, 2018.[34,35]

Completeness-aware concept-based explanations

Completeness-aware concept-based explanation (CCE) proposed by Yeh et al.[26] is a method geared toward discovering sets of concepts that are not only positively relevant to the predicted classes but also complete. A complete set is akin to a sufficient statistic, a function of the input that retains all information relevant to prediction.[27] CCE identifies concepts by partitioning linear directions in the activation space of a hidden layer such that similar concepts are as close as possible and dissimilar ones as distant as possible. The meaning of concepts is determined by inspecting input instances. This approach distinguishes itself by identifying complete sets of concepts as opposed to single concepts. However, it affords little control on whether the discovered concepts are meaningful.

Minimal sufficient statistics and the information bottleneck

DNNs may learn compressed, efficient representations of concepts because the space to represent concepts is limited. From a statistical point of view, the layers of a DNN form a Markov chain, which means that in deeper layers, information is lost, and internal representations become more abstract.[23,28] But what are the rules that guide how concepts are compressed? To answer this, a global perspective on the representation of concepts is necessary. The CCE approach provides a global perspective in the form of a complete set of concepts (a sufficient statistic). However, in order to account for the idea of an efficient representation of concepts, minimal sufficient statistics (MSSs) are needed.[27] MSSs are sufficient statistics that are as coarse as possible and thus provide the most efficient representation without losing predictive power.

Some have argued that DNNs cannot learn MSSs.[29] MSSs only yield a useful degree of compression for a very particular kind of data distribution,[27] which is not given for most empirical datasets processed by DNNs. A helpful framework that generalizes MSSs and can be applied to DNNs is the so-called information bottleneck (IB) method.[29–31] Tishby and collaborators propose that the IB explains how internal representations of DNNs arise as a trade-off between predictive accuracy (sufficiency) and compression (minimality). Formally, the IB trade-off is a constrained optimization problem, which yields a predictively optimal representation for a given level of compression (information loss). Tishby et al. argue that layers of DNNs in fact approximate the optimum given by the IB trade-off. Note that while the IB framework has been applied extensively, it has been contested whether it is an adequate account of how internal representations arise in DNNs.[32,33]

Visualizing conceptual spaces: Color naming

The IB framework provides a theoretical picture of how entire conceptual spaces emerge in DNNs. Unfortunately, it is unclear what we can learn about actual conceptual spaces in given DNNs, that is, what is learned, even if the theoretical picture of how learning works as provided by the IB framework is correct.[9] In order to illustrate what could be learned if conceptual spaces in DNNs were accessible, we will now consider an application of the IB framework to concepts in an empirical context.

Zaslavsky et al.[34] use the IB framework to explain how color-naming systems arise as a result of efficiency. Different natural languages use different systems to name colors. Based on a standard representation of colors (the WCS stimulus palette; see Figure 1), one can determine how speakers of different languages name the color chips on this palette.

This yields different (soft) partitions of the color space, corresponding to the color systems of these languages (cf. Figure 2, top row).

Zaslavsky et al. propose a theoretical explanation of the origin of these color-naming systems of different languages. They argue that the different empirical partitions (Figure 2, top row) match closely with partitions that are derived from the IB framework (Figure 2, bottom row). The main difference between languages is the number of color concepts they use. A language with more color concepts yields a more fine-grained partition and a language with less colors a more coarse-grained partition. The partitions derived from the IB framework are determined to a large degree by the trade-off between accuracy and compression, controlled by the parameter β_I, which yields different numbers of concepts (the theoretical predictions also depend on the so-called least informative prior). The close fit between theoretical and empirical partitions suggests that color-naming systems in different languages have evolved to communicate accurately about colors at a given level of compression and that the level of compression is due to the different communicative needs of the societies using the languages.

How is this related to conceptual spaces in DNNs? To spell out the analogy, the naming systems correspond to internal

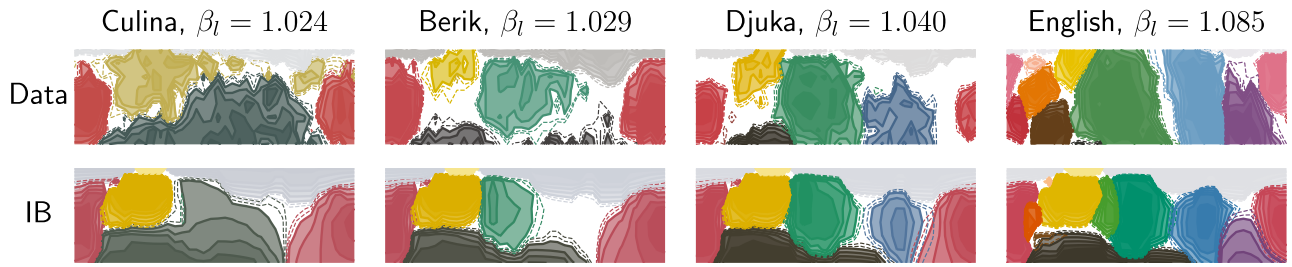

Figure 2. Color naming of four different languages, determined empirically (top row) and theoretically from the IB (bottom row)
The parameter β_l controls the degree of compression in the IB trade-off.[34,35]

representations, e.g., partitions of activation patterns in a hidden layer. The cells of the partition (colors) correspond to clusters of activation patterns with a meaning (concepts). The degree of compression is measured by the number of concepts, which is determined by communicative need in the case of color-naming systems and the predicted classes and representational capacity in the case of DNNs. The analogy is substantive to the extent that both color spaces and representations in DNNs are driven by the IB objective.

The analogy allows us to get a sense of how sets of concepts may emerge holistically in DNNs, that is, as a function of predicting classes while having a limited representational capacity. If we compare the different partitions in Figure 2, we can see that as the number of colors changes, the entire partition changes and therefore the representations of all concepts. This illustrates how the representation of single concepts depends on the representational capacity of a DNN and on all other concepts learned at the same time.

Of course, this is only an illustration; the analogy has its limits. For one, colors are special concepts because they are disjunctive, which need not be the case for other concepts. Also, the representation of colors is non-hierarchical, in contrast to complex representations in DNNs. Note that the conceptual spaces of other kinds of objects have been investigated, but they do not allow for similarly striking visualizations.[36] An important open question about compressed representations concerns the mechanism by which compression is achieved. It is unclear whether compression is due to limited representational space because many successful DNNs are overparametrized, as witnessed by the double-descent risk curve.[20,37] Compression could also be an effect of randomness induced by stochastic gradient descent.[29]

DISCUSSION

Robust detection of concepts

All methods we examined above require that concepts are specified or detected via partial extensions, which is problematic because partial extensions underdetermine the meaning of concepts. Feature visualization relies on optimization, which raises other issues. These problems can be seen as in-principle, philosophical obstacles to detecting concepts in DNNs. From a more pragmatic perspective, the individual weaknesses of these methods could be overcome to some extent by combining them and performing what is known as robustness analysis. Robustness analysis, first proposed in population biology, determines whether different, imperfect methods arrive at the same prediction to increase reliability, under the slogan "truth is at the intersection of independent lies."[38–42] In analogy, robust detection of concepts means using multiple methods like TCAV and feature visualization to detect the same concept. If different methods detect the same concept independently, this should raise our confidence that the methods are somewhat reliable. The required independence of methods seems to be given because, e.g., feature visualization depends on optimization, while TCAV does not. Robustness analysis is limited in that it will not yield an absolute confirmation of concepts[43]—it is only as good as the set of methods in combination—but it is better than using only one method. A combination of different methods contributing to interpretability has been proposed and explored.[22,44] If new methods of concept detection are proposed, it would be desirable that they use a different path than existing methods, in order to increase robust detection.

Testing methods with synthetic data

The methods for detecting concepts are limited to extracting local or linear information. It would be desirable to extend the scope of the methods to encompass concepts with distributed and non-linear representations. However, this will be hard to carry out by sticking to the extensional paradigm of concepts. Defining concepts with sets of instances, like TCAV, only allows for limited control on the meaning of concepts. One possibility to gain more control of meaning would be to create synthetic datasets in which not only the predicted classes (animals) are labeled but also intermediate concepts (body parts, textures, etc.), which may re-emerge in internal representations of DNNs—interpretable datasets, so to speak. This approach has been proposed in the context of interpretable architectures.[12] However, synthetic datasets could also be used to test methods for non-interpretable architectures, such as TCAV or feature visualization. In the context of physical modeling, the use of simulation data has led to some progress in developing DNN emulators for which emerging, high-level properties (e.g., energy conservation) can be checked.[45–47] One of the main challenges of this approach would be to come up with a principled system for labeling intermediate concepts.

The need for conceptual spaces

Above, I discussed both concepts and conceptual spaces. One could ask whether both are really necessary because once we have found all the concepts in an internal representation,

we have arguably also found the conceptual space. This argument presupposes that the detection of individual concepts in internal representations is reliable and leads to a neat partition of the representational space. However, this presupposition is not realized in practice. Existing methods are not (yet) reliable. Also, conceptual spaces may contain elements like polysemantic neurons, as well as artifacts, which do not have neat conceptual counterparts. Understanding how entire conceptual spaces are formed is an additional path to understanding how individual concepts are formed. Bottom-up methods, which allow the detection of individual concepts, and top-down methods, which examine entire conceptual spaces, should not be seen as competing but as complementary ways of triangulating concepts in internal representations, ultimately making the triangulation more reliable.

Conclusions

Reviewing methods for concept detection, we saw evidence that DNNs are able to represent non-trivial inferential relations between predicted classes and emergent concepts. This indicates that DNNs may be able to acquire information that is not purely extensional. However, detecting emergent concepts in the first place is unreliable because existing methods rely on partial extensions of concepts, which makes them susceptible to philosophical problems such as the indeterminacy of reference and the bad lot argument. These limitations should give us pause, given that we have used an undemanding theory of concepts. Finally, the problem of understanding how entire sets of concepts arise holistically in internal representations through trade-offs between predictive accuracy and compression is underexplored. Novel methods to detect concepts as well as conceptual spaces are needed.

ACKNOWLEDGMENTS

The author thanks audience members of the philosophy of science colloquium in Bern as well as reviewers of this journal for useful feedback on an earlier version of this perspective. This work is funded by the Swiss National Science Foundation through grant number 197504. Note that Figures 1 and 2 fall under a PNAS exclusive license to publish and are excluded from the [CC] license to publish.

REFERENCES

1. Bengio, Y. (2009). Learning Deep Architectures for AI (Now Publishers Inc).
2. LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444.
3. Margolis, E., and Laurence, S. (2021). Concepts. In The Stanford Encyclopedia of Philosophy, E.N. Zalta, ed. (Metaphysics Research Lab, Stanford University).
4. Buckner, C. (2018). Empiricism without magic: transformational abstraction in deep convolutional neural networks. Synthese *195*, 5339–5372.
5. Zalta, E.N. (2022). Gottlob Frege. In The Stanford Encyclopedia of Philosophy, E.N. Zalta, ed. (Metaphysics Research Lab, Stanford University).
6. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural nets. Preprint at arXiv. https://doi.org/10.48550/arXiv.1312.6199.
7. Jebeile, J., Lam, V., and Räz, T. (2021). Understanding climate change with statistical downscaling and machine learning. Synthese *199*, 1877–1897.
8. Räz, T., and Beisbart, C. (2022). The importance of understanding deep learning. Erkenn, 1–18. forthcoming.
9. Boge, F.J. (2021). Two dimensions of opacity and the deep learning predicament. Minds Mach. *32*, 43–75. https://doi.org/10.1007/s11023-021-09569-4.
10. Buckner, C. (2020). Understanding adversarial examples requires a theory of artefacts for deep learning. Nat. Mach. Intell. *2*, 731–736.
11. Chen, Z., Bei, Y., and Rudin, C. (2020). Concept whitening for interpretable image recognition. Nat. Mach. Intell. *2*, 772–782.
12. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. (2020). Concept bottleneck models. In International Conference on Machine Learning (PMLR), pp. 5338–5348.
13. Bau, D., Zhu, J.-Y., Strobelt, H., Lapedriza, A., Zhou, B., and Torralba, A. (2020). Understanding the role of individual units in a deep neural network. Proc. Natl. Acad. Sci. USA *117*, 30071–30078.
14. Douven, I. (2021). Abduction. In The Stanford Encyclopedia of Philosophy, E.N. Zalta, ed. (Metaphysics Research Lab, Stanford University).
15. Van Fraassen, B.C. (1989). Laws and Symmetry (Clarendon Press).
16. Michaelson, E., and Reimer, M. (2022). Reference. In The Stanford Encyclopedia of Philosophy, E.N. Zalta, ed. (Metaphysics Research Lab, Stanford University).
17. Quine, W.V.O. (2013). Word and Object (MIT press).
18. Olah, C., Mordvintsev, A., and Schubert, L. (2017). Feature visualization. Distill *2*, e7. https://distill.pub/2017/feature-visualization.
19. Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. (2020). Zoom in: An Introduction to Circuits (Distill). https://distill.pub/2020/circuits/zoom-in.
20. Berner, J., Grohs, P., Kutyniok, G., and Petersen, P. (2021). Theory of Deep Learning. In The Modern Mathematics of Deep Learning (Cambridge University Press). ch.
21. Vidal, R., Bruna, J., Giryes, R., and Soatto, S. (2017). Mathematics of deep learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.1712.04741.
22. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and Sayres, R. (2018). Interpretability beyond feature attribution: quantitative testing with concept activation vectors (tcav). In International conference on machine learning (PMLR), pp. 2668–2677.
23. Alain, G., and Bengio, Y. (2016). Understaning intermediate layers using linear classifier probes. Preprint at arXiv. https://doi.org/10.48550/arXiv.1610.01644.
24. Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In Advances in Neural Information Processing Systems, *27*.
25. Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., and Olah, C. (2021). Multimodal neurons in artificial neural networks. Distill *6*, e30. https://distill.pub/2021/multimodalneurons.
26. Yeh, C.-K., Kim, B., Arik, S., Li, C.-L., Pfister, T., and Ravikumar, P. (2020). On completeness-aware concept-based explanations in deep neural networks. In Advances in Neural Information Processing Systems, *33*, pp. 20554–20565.
27. Casella, G., and Berger, R.L. (2002). Statistical Inference, second ed..
28. Achille, A., and Soatto, S. (2018). Information dropout: learning optimal representations through noisy computation. In IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI).
29. Shwartz-Ziv, R., and Tishby, N. (2017). Opening the black box of deep neural networks via information. Preprint at arXiv. https://doi.org/10.48550/arXiv.1703.00810.
30. Räz, T. (2022). Understanding deep learning with statistical relevance. Philos. sci. *89*, 20–41.
31. Shamir, O., Sabato, S., and Tishby, N. (2010). Learning and generalization with the information bottleneck. Theor. Comput. Sci. *411*, 2696–2711.

32. Geiger, B.C., and Kubin, G. (2020). Information Bottleneck: Theory and Applications in Deep Learning. Entropy *22,* 12:1408.

33. Saxe, A.M., Bansal, Y., Dapello, J., Advani, M., Kolchinsky, A., Tracey, B.D., and Cox, D.D. (2018). On the information bottleneck theory of deep learning. In ICLR.

34. Zaslavsky, N., Kemp, C., Regier, T., and Tishby, N. (2018). Efficient compression in color naming and its evolution. Proc. Natl. Acad. Sci. USA *115*, 7937–7942.

35. Zaslavsky, N., Kemp, C., Regier, T., and Tishby, N. (2018). Efficient human-like semantic representations via the information bottleneck principle. Preprint at arXiv. https://doi.org/10.48550/arXiv.1808.03353.

36. Zaslavsky, N. (2019). Information-theoretic Principles in the Evolution of Semantic Systems (The Hebrew University of Jerusalem). PhD thesis, Ph. D. dissertation.

37. Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proc. Natl. Acad. Sci. USA *116*, 15849–15854.

38. Boge, F.J. (2021). Why trust a simulation? models, parameters, and robustness in simulation-infected experiments. Br. J. Philos. Sci. https://doi.org/10.1086/716542.

39. Knuuttila, T., and Loettgers, A. (2011). Causal isolation robustness analysis: the combinatorial strategy of circadian clock research. Biol. Philos. *26*, 773–791.

40. Levins, R. (1966). The strategy of model building in population biology. Am. Sci. *54*, 4.

41. Wimsatt, W.C. (2012). Robustness: material, and inferential, in the natural and human sciences. In Characterizing the Robustness of Science, L. Soler, M. Brewer and B. Collins, eds. (Springer), pp. 89–104. ch. 3.

42. Wimsatt, W.C. (2012). Robustness, reliability, and overdetermination. In Characterizing the Robustness of Science, L. Soler, M. Brewer and B. Collins, eds. (Springer), pp. 61–87. ch. 2.

43. Orzack, S.H., and Sober, E. (1993). A critical assessment of levins's the strategy of model building in population biology (1966). Q. Rev. Biol. *68*, 533–546.

44. Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., and Mordvintsev, A. (2018). The building blocks of interpretability. Distill *3*, e10. https://distill.pub/2018/building-blocks.

45. Beucler, T., Rasp, S., Pritchard, M., and Gentine, P. (2019). Achieving conservation of energy in neural network emulators for climate modeling. Preprint at arXiv. https://doi.org/10.48550/arXiv.1906.06622.

46. Couvreux, F., Hourdin, F., Williamson, D., Roehrig, R., Volodina, V., Villefranque, N., Rio, C., Audouin, O., Salter, J., Bazile, E., et al. (2021). Process-based climate model development harnessing machine learning: I. a calibration tool for parameterization improvement. J. Adv. Model. Earth Syst. *13*, e2020MS002217.

47. Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., and Yacalis, G. (2018). Could machine learning break the convection parameterization deadlock? Geophys. Res. Lett. *45*, 5742–5751.

About the author

Tim Räz is a postdoctoral researcher at the Institute of Philosophy, University of Bern, Switzerland. He has obtained degrees in philosophy (PhD 2013, University of Lausanne) and mathematics (MSc 2019, University of Bern). A philosopher of science by training, he works on conceptual issues in computer science, in particular group fairness and individual fairness, interpretability and understanding in machine learning, and the use of machine learning in climate modeling and criminal justice.