Patterns

CellPress
OPEN ACCESS

Preview

# The overfitted brain hypothesis

Luke Y. Prince[3] and Blake A. Richards[1,2,3,4,*]
[1]Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada
[2]School of Computer Science, McGill University, Montreal, Quebec, Canada
[3]Mila, Montreal, Quebec, Canada
[4]Learning in Machines and Brains Program, CIFAR, Toronto, Ontario, Canada
*Correspondence: blake.richards@mila.quebec
https://doi.org/10.1016/j.patter.2021.100268

**What is the purpose of dreaming? Many scientists have postulated a role for dreaming in learning, often with the aim of improving generative models. In this issue of *Patterns*, Erik Hoel proposes a novel hypothesis, namely, that dreaming provides a means to reduce overfitting. This hypothesis is interesting both for neuroscience and for the development of new machine-learning systems.**

The reasons for how and why we dream are poorly understood; however, the suppression of sleep stages most closely associated with dreaming have been known to impair learning in mammals for some time.[1] Given the limits of experimental techniques in cognitive and systems neuroscience, theories and experiments that address the role of sleep in learning often do not consider the impact that dreaming *specifically* may have. In a new perspective in this issue of *Patterns,* titled "The overfitted brain: Dreams evolved to assist generalization", Erik Hoel explores this question using concepts from machine learning.[2]

Interestingly, within machine learning, there is actually a long history of algorithmic techniques that use dream-like processes for learning.[3–6] Many of these were motivated by the challenges of training probabilistic generative models. Specifically, machine learning researchers often wrestle with the dilemma of needing to find a model that maximizes the marginal likelihood or *evidence* of observed data, despite the fact that they are unable to evaluate this intractable quantity. A large amount of machine learning research is devoted to circumventing this problem by instead maximizing a suitable lower bound on this quantity, referred to as the evidence lower bound (ELBO) and sometimes as negative variational free energy.[7]

A canonical example is the expectation-maximization (EM) algorithm,[8] the iterative two-stage procedure used in fitting Gaussian mixture models and other hierarchical probabilistic models without closed-form marginal likelihoods. This involves an *expectation,* E-step, in which we fix model parameters and compute the expected posterior mixture assignments, and a *maximization,* M-step, in which we update model parameters to maximize the ELBO while keeping the posterior mixture assignments fixed. One can also think of EM as performing two steps of maximization of the ELBO, because during the E-step one is finding a posterior distribution that maximizes the negative variational free energy.[9]

Iterative two-step training procedures such as these are ubiquitous in approaches to solving the problem of fitting hierarchical probabilistic models. Possibly the most relevant in relation to Hoel's theory is the wake-sleep algorithm for training Helmholtz machines.[3,5] The two phases of the wake-sleep algorithm[5] correspond to an input-driven "wake" phase, akin to the M-step of the EM algorithm, and an internal representation-driven "sleep" phase, akin to the E-step of the EM algorithm. During the wake phase, layers in a deep neural network are activated by fixed feedforward connections, while feedback connections are updated to maximize the probability of reconstructing the input (the M-step). The sleep phase plays out in reverse: layers are activated by fixed feedback connections, while feedforward connections are updated to find a posterior that better matches the generative distribution (the E-step). In this case, hidden layer activations during the sleep phase are interpreted as "dreams" because they are internally generated data constructed during a time when the network is cut off from sensory inputs. According to this theory, dreams are a way of aligning our recognition pathways with our generative pathways, and so, over time, dreams should come more and more to resemble that which is experienced during waking. Within neuroscience and psychology, recent theories inspired by these machine-learning algorithms have attempted to explain learning and the role that sleep and dreaming might contribute in a similar light.[10]

However, Hoel proposes a different role for dreams, also inspired by machine learning. Specifically, Hoel's hypothesis is that dreams help to prevent overfitting. Specifically, he proposes that the purpose of dreaming is to aid generalization and robustness of learned neural representations obtained through interactive waking experience. Dreams, Hoel theorizes, are augmented samples of waking experiences that guide neural representations away from overfitting waking experiences. Hoel argues that the properties of these augmentations explain why dreams are both less detailed and more fantastic than waking experience but retain the sequential ordering that we are familiar with.

This proposal is different from EM-style proposals because the dream phase is not used to improve the match between a generative model and a recognition model, but rather to regularize a single model. This essentially proposes that brains use a secondary training phase in order to engage in some regularization process courtesy of using corrupted versions of the original data. This is akin, to some degree, to the use of augmented data for training

machine-learning models to improve robustness and generalization. Hoel's proposal is interesting from a neuroscience perspective, as it provides a normative theory of dreaming that, unlike the EM-style proposals, can explain why dreams do not become more realistic over time. But it is also interesting from a pure machine-learning perspective. If Hoel is correct, there should be ways to incorporate the phenomenology of dreaming to algorithm design for training and regularizing artificial neural networks (ANNs).

Given that a fundamental flaw of current deep neural networks is their inability to learn representations from data that generalize to out-of-distribution samples, this is an interesting proposal. Cognitive scientists and machine-learning researchers alike have argued that this ability is key to learning causal world models,[11] and it relies on being able to decompose sensory experiences into disjoint hierarchies of discrete objects with locally continuous features. For example, the ability to recognize a coworker in an unfamiliar location relies on the ability to separate the coworker from the workplace and attire one usually perceives them in. In this case the disjoint hierarchical objects are the coworker, attire, and environment, which themselves can be decomposed into further sub-objects (facial features, items of clothing, workplace furniture, etc.).

While solutions to this problem in deep-learning research are still in their infancy, the most common paradigm is to include a mechanism that guides learned representations away from the training data, either explicitly through regularization in the objective function or implicitly by injecting simple independent noise to the training data. More complex strategies try to generate synthetic training examples that exploit counterfactuals to observations, thereby supporting learning of causal relations between discrete objects. Hoel's proposals for the role of dreaming in human and animal learning can be interpreted as exploiting each of these strategies.

Finally, Hoel speculates on the use of dream substitutions—dream-like stimuli generated to aid learning during wakefulness or to ameliorate the effects of sleep deprivation. This offers clear empirical predictions for human studies and implications for the use of augmented reality technologies. Forming our own speculations, other waking experiences such as the use of psychedelics may also offer alternative dream substitutions. Recent work shows how the use of psychedelics alters variability in neural activity.[12] Should this variability impact learning in the way Hoel's theory suggests, we might also predict that psychoactive substances or other consciousness-altering experiences might promote more robust learning.

## REFERENCES

1. Walker, M.P., and Stickgold, R. (2004). Sleep-dependent learning and memory consolidation. Neuron *44*, 121–133.
2. Hoel, E. (2021). The Overfitted Brain: Dreams evolved to assist generalization. Patterns *2*. https://doi.org/10.1016/j.patter.2021.100244.
3. Dayan, P., Hinton, G.E., Neal, R.M., and Zemel, R.S. (1995). The Helmholtz machine. Neural Comput. *7*, 889–904.
4. Hinton, G.E., and Sejnowski, T.J. (1986). Learning and relearning in Boltzmann machines. In Parallel distributed processing: Explorations in the microstructure of cognition, J. McLellan, ed. (MIT Press), p. 2.
5. Hinton, G.E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. Neural Comput. *18*, 1527–1554.
6. Hinton, G.E., Dayan, P., Frey, B.J., and Neal, R.M. (1995). The "wake-sleep" algorithm for unsupervised neural networks. Science *268*, 1158–1161.
7. Kingma, D.P., and Welling, M. (2013). Auto-encoding variational bayes. arXiv. arXiv:1312.6114v10.
8. Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. J. R. Stat. Soc. B *39*, 1–38.
9. Neal, R.M., and Hinton, G.E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In Learning in graphical models, M.I. Jordan, ed. (Springer), pp. 355–368.
10. Friston, K. (2010). The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. *11*, 127–138.
11. Schölkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Towards Causal Representation Learning. arXiv. arXiv:2102.11107v1.
12. Schartner, M.M., Carhart-Harris, R.L., Barrett, A.B., Seth, A.K., and Muthukumaraswamy, S.D. (2017). Increased spontaneous MEG signal diversity for psychoactive doses of ketamine, LSD and psilocybin. Sci. Rep. *7*, 46421.