Behavioral and Brain Sciences

Decompressing the kinds of compression

--Manuscript Draft--

Manuscript Number:	
Full Title:	Decompressing the kinds of compression
Short Title:	Decompressing the kinds of compression
Article Type:	Open Peer Commentary
Corresponding Author:	Aaron M Bornstein University of California Irvine Irvine, CA UNITED STATES OF AMERICA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of California Irvine
Corresponding Author's Secondary Institution:	
First Author:	Dale Zhou
First Author Secondary Information:	
Order of Authors:	Dale Zhou
	Aaron M Bornstein
Order of Authors Secondary Information:	
Abstract:	Information compression is a part of many theories in cognitive science and neuroscience. How-ever, implementations and algorithms that perform compression vary by input modality, cognitive process, and situation. We provide an overview of the algorithms and architectures that compress information under resource constraints. Considering more kinds of compression can expand under-standing of how learning systems discover generalizable patterns from experience.

Decompressing the kinds of compression

Dale Zhou[1,2] and Aaron M. Bornstein[1,2,3]

[1]University of California, Irvine, Neurobiology and Behavior, 519 Biological Sciences Quad, Irvine, 92697, United States
[2]University of California, Irvine, Center for the Neurobiology of Learning and Memory, Qureshey Research Laboratory, Irvine, 92697, United States
[3]University of California, Irvine, Department of Cognitive Sciences, 2318 Social & Behavioral Sciences Gateway, Irvine, 92697, United States

Email addresses:
dale.zhou@uci.edu
aaron.bornstein@uci.edu

Websites:
https://dalezhou.com
https://aaron.bornstein.org/

Authors' note: Dale Zhou and Aaron M. Bornstein are co-corresponding authors.

Abstract: Information compression is a part of many theories in cognitive science and neuroscience. However, implementations and algorithms that perform compression vary by input modality, cognitive process, and situation. We provide an overview of the algorithms and architectures that compress information under resource constraints. Considering more kinds of compression can expand under-standing of how learning systems discover generalizable patterns from experience.

Dubova & Sloman ask how minds allocate more representational resources than necessary to encode past experiences in an excess-capacity system. In contrast, a constrained-capacity system compresses experiences, where compression is defined as constructing an internal representation that is simpler than the input experiences by discarding detail deemed irrelevant. A model that learns signal and discards noise can *generalize* beyond its past experiences. Excess capacity regimes do not need to discard any information, because they can memorize arbitrary input-output mappings. Yet, they still retain the ability to learn generalities, an ability commonly thought to depend on compression.

The excess capacity framework productively explains how cognitive and neural systems need not be overly constrained to learn generalities. It generates novel research directions about how strongly to apply constraints of resource efficiency when developing models of neural and cognitive processes. Both resource constraints and abundance are important for network efficiency, robustness, control, and inference (Avena-Koenigsberger, Misic, & Sporns, 2018; Srivastava et al., 2020; Sterling & Laughlin, 2015; Noh, Zhou, et al., 2026; Noh, Cooper, et al., 2026).

The excess-capacity framework generates novel directions for future research. How can an excess capacity learning system allocate sufficient representational resources? How can a single learning system dynamically transition between constrained (compressed) and excess-capacity regimes in response to task demands? These are questions addressable by more full-featured theories of compression than considered by Dubova & Sloman.

In the target article, compression is largely considered to fall under the constrained capacity systems and under lossy compression. For example, bottlenecking the representational architecture to constrain information flow produces compressed representations in the autoencoder family of neural networks (Hinton & Salakhutdinov, 2006). In contrast, convolutional neural networks trained on visual classification tasks are considered to fall under the excess capacity regime. These networks can memorize arbitrary input-output mappings, yet still generalize (Zhang, Bengio, Hardt, Recht, & Vinyals, 2017; Zhang, Bengio, Hardt, Mozer, & Singer, 2020).

The present focus on convolutional neural networks may understate how many learning systems are likely to operate in the excess capacity regime. While there is some empirical support suggesting the representations learned by convolutional neural networks are similar to neural representations (Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007; Lindsay, 2021), there are open questions as to what a biologically plausible architecture and algorithm to form such representations would be (Kim, Ricci, & Serre, 2018; Bowers et al., 2023). Moreover, as the performance of these networks improves, the learned representations diverge from humans (Linsley et al., 2023). This opens the excess capacity framework to similar criticisms that such networks are simply not good models of human learning systems, because they can do things like memorize arbitrary mappings when most humans cannot.

In any case, convolutional neural networks may perform some level of compression via fixed transforms as inputs are processed through deeper layers the neural networks. This complicates the understanding of compression as being characteristic of constrained capacity and distinct from excess-capacity regimes. While these networks are not explicitly designed to perform compression and do not have theoretical guarantees on the bounds of information retained or lost, they do lose information to form less detailed representations which are useful for downstream tasks such as classification or decision-making (Zhou et al., 2025). There is an overall dimensionality reduction of representations, suggesting compression (Recanatesi et al., 2019). It remains an open question how to establish what neural and cognitive processes explicitly compress information (Humphries, 2020). A brain region does not necessarily compress information because an experimenter can find compressible neural activity. The eye does not compress away sound just because it “filters” that kind of information. Learning systems that are not capable of representing some forms of information in the first place do not actively discard such information. This makes it challenging to infer compression from descriptive measures of dimensionality and complexity, because it hinges on unresolved questions about how a brain region or cognitive module represents information. To go beyond the focus on convolutional neural networks and whether they perform compression, one could expand what kinds of compression fall under the excess capacity system.

Akin to memorization without loss of information, excess capacity systems could include systems capable of lossless compression. In the computer science of compression, deep neural networks can be trained to perform neural compression (Yang, Mandt, & Theis, 2023). Like the excess capacity regime, these large neural networks use the flexibility gained from a large number of parameters in order to optimize for a better probabilistic, predictive model of the data. In the brain, probabilistic generative models of data are thought to be learned by cortico-hippocampal systems (Spens & Burgess, 2024). Moreover, predictive representations like the successor and predecessor representations recode sensory information in terms of its predictions about past and future states (Yoo, Chrastil, & Bornstein, 2024). These compressed representations support generalization and can exist alongside more granular state representations. Excess capacity systems may accomplish generalization by lossless compression or representation learning.

A natural question is how valid these systems and compression schemes are as models of human learning systems. There are several biologically plausible options. Dubova & Sloman describe a complementary learning systems model, C-HORSE, a biologically realistic model of hippocampal and cortical function (Schapiro, Turk-Browne, Botvinick, & Norman, 2017). They note parts of the model falling under the constrained or excess capacity regimes, depending on the learning environment. Beyond task settings, these neural networks use Hebbian learning which performs compression via dimensionality reduction (R. C. O'Reilly, Munakata, Frank, Hazy, & Contributors, 2024). Hebbian plasticity adjusts weights to represent the largest principal components of experience, the dimensions that explains the most variance in the input data (Sanger, 1989). In the artificial neural network, the amount of dimensionality reduction is determined by different neural architectures and physiological operating ranges (R. O'Reilly, Munakata, & Mcclelland, 2000). Similarly, in the brain network, patterns of dimensionality expansion and reduction across cortical and hippocampal systems are shaped by distinct white matter architectures and the distribution of physiological resources (Zhou, Lynn, et al., 2022; Zhou, Kim, et al., 2022). Hence, a single learning system may transition between constrained and excess capacity by differentially activating different subsystems with specialized connectivity and physiological operating ranges implementing a hierarchy of compression.

The varieties of compression point to the importance for many theories and frameworks that depend on compression as a computation to consider its myriad algorithmic and implementational instantiations. Together, this overview and expansion of the excess-capacity framework indicate that both compression and generalization could be multiply realizable.

Competing interests: none.

Funding statement:
D.Z. acknowledges funding from the George E. Hewitt Foundation for Medical Research. A.M.B. acknowledges funding from NIA R01AG088306.

References

Avena-Koenigsberger, A., Misic, B., & Sporns, O. (2018). Communication dynamics in complex brain networks. *Nature reviews neuroscience*, *19* (1), 17–33.

Bowers, J. S., Malhotra, G., Dujmovi´c, M., Montero, M. L., Tsvetkov, C., Biscione, V., . . . others (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, *46* , e385.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, *313* (5786), 504–507.

Humphries, M. D. (2020). Strong and weak principles of neural dimension reduction. *arXiv preprint arXiv:2011.08088* .

Kim, J., Ricci, M., & Serre, T. (2018). Not-so-clevr: learning same–different relations strains feedforward neural networks. *Interface focus*, *8* (4), 20180011.

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of cognitive neuroscience*, *33* (10), 2017–2031.

Linsley, D., Rodriguez Rodriguez, I. F., Fel, T., Arcaro, M., Sharma, S., Livingstone, M., & Serre, T. (2023). Performance-optimized deep neural networks are evolving into worse models of inferotem-poral visual cortex. *Advances in Neural Information Processing Systems*, *36* , 28873–28891.

Noh, S. M., Cooper, K. W., Guo, S., Zhou, D., Stark, C. E., & Bornstein, A. M. (2026). Multistep inference across the human lifespan can be improved with individualized memory interventions. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, *81* (5), gbag038.

Noh, S. M., Zhou, D., Cooper, K. W., Guo, S., Dinh, E. T., & Bornstein, A. M. (2026). Sparsity and memory constraints interact with training sequence to bias learning of associative maps. *Neuropsy-chologia*, 109396.

O'Reilly, R., Munakata, Y., & Mcclelland, J. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. MIT Press. Retrieved from https://books.google.com/books?id=gL5NEAAAQBAJ

O'Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E., & Contributors. (2024). *Computational cognitive neuroscience*. Online Book, 5th Edition, URL: https://book.compcogneuro.org. Retrieved from https://book.compcogneuro.org

Recanatesi, S., Farrell, M., Advani, M., Moore, T., Lajoie, G., & Shea-Brown, E. (2019). Dimensionality compression and expansion in deep neural networks. *arXiv preprint arXiv:1906.00443* .

Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural networks*, *2* (6), 459–473.

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372* (1711), 20160049.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE transactions on pattern analysis and machine intelligence*, *29* (3), 411–426.

Spens, E., & Burgess, N. (2024). A generative model of memory construction and consolidation. *Nature human behaviour* , *8* (3), 526–543.

Srivastava, P., Nozari, E., Kim, J. Z., Ju, H., Zhou, D., Becker, C., . . . Bassett, D. S. (2020). Models of communication and control for brain networks: distinctions, convergence, and future outlook. *Network Neuroscience*, *4* (4), 1122–1159.

Sterling, P., & Laughlin, S. (2015). *Principles of neural design*. MIT press.

Yang, Y., Mandt, S., & Theis, L. (2023). An introduction to neural data compression. *Foundations and Trends in Computer Graphics and Vision*, *15* (2), 113–200.

Yoo, J., Chrastil, E. R., & Bornstein, A. M. (2024). Cognitive graphs: Representational substrates for planning. *Decision*.

Zhang, C., Bengio, S., Hardt, M., Mozer, M. C., & Singer, Y. (2020). Identity crisis: Memorization and generalization under extreme overparameterization. In *International conference on learning representations (iclr).* Retrieved from https://arxiv.org/abs/1902.04698

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. In *International conference on learning representations (iclr).* Retrieved from https://arxiv.org/abs/1611.03530

Zhou, D., Kim, J. Z., Pines, A. R., Sydnor, V. J., Roalf, D. R., Detre, J. A., . . . Bassett, D. S. (2022). Compression supports low-dimensional representations of behavior across neural circuits. *arXiv preprint arXiv:2211.16599* .

Zhou, D., Lynn, C. W., Cui, Z., Ciric, R., Baum, G. L., Moore, T. M., . . . others (2022). Efficient coding in the economics of human brain connectomics. *Network Neuroscience*, *6* (1), 234–274.

Zhou, D., Noh, S. M., Harhen, N. C., Banavar, N. V., Kirwan, C. B., Yassa, M. A., & Bornstein, A. M. (2025). A compressed code for memory discrimination. *arXiv preprint arXiv:2510.10791* .