Using self-supervision and augmentations to build insights into neural coding

Mehdi Azabou* Georgia Tech
Max Dabagia* Georgia Tech
Ran Liu* Georgia Tech
Chi-Heng Lin Georgia Tech
Keith B. Hengen WashU-St Louis
Eva L. Dyer Georgia Tech

Abstract

Self-supervised learning (SSL) provides a powerful mechanism for building representations of complex data without the need for labels. In this perspective piece, we highlight recent progress in the application of self-supervised learning (SSL) to data analysis in neuroscience, discuss the implications of these results, and suggest ways in which SSL might be applied to reveal interesting properties of neural computation.

1 Introduction

Self-supervised learning (SSL) provides a powerful mechanism for building meaningful representations without the need for labels (1; 2; 3; 4). Instead of using a supervised objective to guide learning, state-of-the-art methods for self-supervision engineer invariance to task-irrelevant information through the use of *augmentations*: examples are transformed to create augmentations, and the representation is trained to maximize the similarity of transformed/augmented views. Thus, the choice of augmentations controls which invariances are built into the representation. Accordingly, designing good augmentations generally requires a significant level of domain knowledge.

In the application of these methods to biological datasets, there is an intriguing duality: Useful augmentations allow better representations to be learned, while if an augmentation yields good performance on a downstream task, we can infer that the encoding of information in the dataset is invariant to the augmentation. Whether an interesting transformation leaves an example's semantics intact or destroys meaningful information can be a source of great insight into how the biological system may encode task-relevant information.

Here, we highlight recent progress in the application of SSL to brain activity across different spatiotemporal scales (5; 6; 7; 8; 9; 10), and discuss ways in which SSL and the study of augmentations may lead to interesting discoveries in neuroscience. The focus of this perspective piece is on what SSL-based mechanisms, when paired with specific augmentations, might be able to tell us about population coding in neural circuits. We conclude by proposing a few directions along which we expect progress may continue.

2 Self-supervised learning and its applications to neuroscience

2.1 Information maximization as a guiding principle for representation learning

Many of the best performing methods for SSL are grounded in principles of *information maximization* and the use of *augmentations*. The idea behind these methods is to augment or transform a data

* These authors contributed equally. Contact: Eva L. Dyer - evadyer@gatech.edu.

Workshop on Self-Supervised Learning - Theory and Practice, 35th Conference on Neural Information Processing Systems

sample with a transformation sampled randomly from a set, and then learn a representation that brings these augmented samples closer to one another (or maximizes their mutual information) in the representation space of the network. When carefully designed, augmentations can be used to build rich invariances into representations that can be used for complex downstream tasks.

Self-supervised learning methods can generally be distinguished by whether they make use of "negative" examples – examples from the dataset which are assumed to be dissimilar to a target example, in contrast to augmentations of the target example which are assumed to be similar. These methods are known as "contrastive", and include `SimCLR` (1) and `MoCo` (2). Contrastive methods typically rely on optimizing different variants of the InfoNCE loss (11), which is a tractable lower bound on mutual information (12). More recently, non-contrastive methods like `BYOL` (3), `PBL` (13) and `SimSiam` (14) have shown that useful representations can be trained by maximizing the similarity between representations of similar examples, where each uses a clever strategy to avoid collapse in the representation. `BarlowTwins` (4) applies a redundancy reduction principle to this general framework, which aims to learn a representation that conserves the maximum amount of mutual information while being invariant to distortion.

2.2 Recent work in self-supervised representation learning in neural datasets

In neuroscience, we are often faced with challenges in collecting (and labeling) training data. Even with ample labeled data, it is rarely clear that labels – typically behavioral or environmental variables – are entirely reflective of the underlying brain state of an individual. Thus, the appeal of self-supervised learning is two-fold: it has potential to construct robust representations of brain activity without requiring labels, and moreover representations unbiased towards predicting a (rather arbitrary) set of external variables.

To build representations across distributed sets of neurons, `MYOW` recently (8) introduced a set of augmentations for neural population activity, and proposed to acquire additional similar views by adaptively selecting examples from across the dataset. The augmentations include temporal jitter (pairing off examples which are close in time) and dropout (randomly masking a subset of the input channels). `Swap-VAE` (9) combined augmentation-based self-supervised information maximization with a generative modeling framework to disentangle different sources of information in the latent representations of multi-unit neural recordings from non-human primates.

In contrast to multi-cellular neural recordings described above that capture the firing and activity of individual neurons, modalities like EEG and ECoG measure aggregate activity across many brain regions. In this case, the net electrical activity of large collections of neurons are captured through many channels that are spatially organized across the scalp (EEG) or surface of the brain (ECoG). To build representations of these macroscale brain data, Cheng, et al. (7) proposed to use subject-specific augmentations and adversarial training techniques to study diverse physiology datasets including electroencephalography (EEG). Banville et al. (6) explored a wide range of temporal pretext tasks applied to EEG for sleep decoding and pathology screening across patients. Recently, Peterson et al. proposed a cross-modal deep clustering approach (10) that builds representations across EEG, electrocorticography (ECoG), and behavior in a self-supervised manner. Transformer-based models like BENDR (15) utilize self-supervised sequence modeling techniques to learn a latent representation from EEG signals.

2.3 Contrastive methods and SSL as potential mechanisms for learning in neural systems

Contrastive losses like those used by `CPC` (11) and `SimCLR` (1) have been proposed as a mechanism for learning in biological neural networks (16; 17; 18). This idea has been extended further to build a model of both the dorsal and ventral stream after using `CPC` to train the model (18). These results suggest that self-supervised losses may be more predictive of neural activity than supervised loss functions that emphasize classification. One can easily imagine that various forms of self-supervision, especially temporal or contrastive prediction, could be performed in neural circuits.

3 What can augmentations and SSL tell us about the brain?

The approaches described in the previous section can learn representations that selectively preserve shared information across different transformations of an example. In what follows, we will describe how different augmentations and pairing of views may be able to provide insights into neural computations.

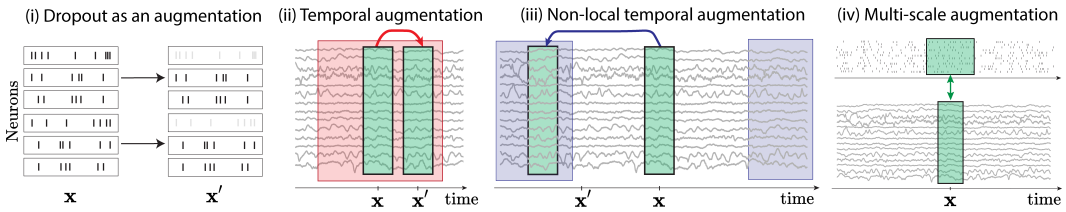

Figure 1: *Overview of augmentations used for neural activity.* Each block illustrates a strategy employed to harness mutual information shared across spatial, temporal or multi-modal samples.

3.1 Temporal augmentations

In applications of SSL to both recordings from populations of neurons (8; 9) and macroscale electrophysiology (5; 6), temporal augmentations are amongst the most widely used. The implicit assumption underlying these approaches is that neural activity is temporally smooth – that is, the content a circuit is encoding changes slowly relative to the sampling rate. In this case, it is possible to treat samples from nearby points in time (i.e. in a window around the current sample) as positive examples (6). Relative positioning (RP) is based on predicting whether a pair of examples are temporally close (19); temporal shuffling (TS), which is a variation on this strategy, instead predicts whether a trio of examples are in the correct order or shuffled (20); `CPC` further leverages temporal structure by training an autoregressive model to predict future time steps given a sequence of past ones (11).

Implications: The length of time over which examples can be paired together and remain similar is a natural measure of the smoothness of the underlying temporal sequence. When there are many change points in the sequence (or when the window of assumed smoothness is too large), the temporal smoothness assumption breaks down and dissimilar brain states may be matched as positive views. As a result, the learned representation may obscure changes over time, with many distinct states collapsed into one.

3.2 Neuron dropout

Dropout is one of the most widely used forms of regularization in neural networks (21), where units throughout the network are randomly set to zero, typically independently. Less commonly used (but still used, especially in vision) is dropout of the inputs (22; 23; 24), where input features are masked before feeding them into the network, in which case it is naturally interpreted as an augmentation. A representation which is invariant to dropout is sensitive only to structured, population-wide changes in activity.

Implication: Neural activity is typically modeled by population dynamics (25), where single-neuron responses fail to fully capture the underlying behavior. Under this assumption, the neural code should be nearly invariant to the activity of single neurons. In other words, if the task-relevant information is encoded by the activity of only a single neuron in the recording, then dropout would be a poor strategy for augmenting neural activity, as views which mask out this neuron's activity would destroy this information. However, if neural circuits indeed encode information in a redundant, distributed manner as hypothesized, then dropout and masking augmentations should produce views that preserve the underlying information, allowing the network to develop a consistent representation. The success of methods like `MYOW` (8) and `Swap-VAE` (9) which use dropout as an augmentation suggests that this is indeed the case.

3.3 Nonlocal yet similar views

A major challenge in neuroscience is learning a mapping between the brain and behavior. The standard approach used to estimate this mapping is to drive repeated behaviors or sensations/perceptions in minimally variable trial-based designs. However, this makes it difficult to address the problem of how behaviors generalize. In reality, robust and reliable behavior plays out across complex environments, brain states, environmental conditions, and motivations. It is unclear whether the insights gleaned from repetitive, trial-based experiments apply to natural behavior. Ideally, we could find repeated patterns in complex dynamics by taking advantage of a maximally diverse set of internal and external conditions.

The combination of long-term recordings of free behavior (26) with emerging methods provides a promising solution to this problem. In extended recordings, we observe repeated instances of behaviors, such as eating, drinking, and sleeping, across a variety of internal and external conditions, such as fatigue, social interaction, and light/dark. Typically, this type of variability would be avoided by design. However, a recent method called Mine Your Own vieW (`MYOW`) (8) shows how similar views can be found adaptively by "mining" nearby samples in representation space. Different brain states are linked through a predictive mechanism, making local associations in time within one projector and then making more global associations between different points in time in a second cascaded projector. These methods operate on the assumption that when an animal repeats a behavior in a variety of conditions, the similarity of the associated brain states is closely related to the repeated behavior.

Implication: Consistent decoding over long time scales is complicated by the myriad processes that modulate brain activity despite similar context, such as attentional states (27; 28; 29), neuron death (30; 31), and electrode movement (32; 33; 34), among others. With labels, different instances of the same behavior (e.g., the same movement across varying levels of attention) can be used as positive views to build invariances into the representation. Techniques like `MYOW` can do something similar without labels; by finding two different instances in time that lead to similar representations, invariances across time and trials can be learned to improve generalization on downstream brain decoding tasks.

3.4 Views from different modalities

Self-supervised approaches which train representations simultaneously for multiple modalities have recently shown tremendous potential, for example in audio and video (35). In neuroscience, many studies involve recording from different data streams, ranging from several different measures of neural activity at various spatiotemporal scales, to physiological signals like muscle movement.

The results of (10) demonstrated the effectiveness of this approach: cross-modal self-supervised learning combining EEG or ECoG data streams with a simultaneously measurement of arm position in a reaching task led to a state-of-the-art neural decoder. Their model built on the deep clustering-based approach of (35) and extended it to an arbitrary number of modalities, exhibiting a three modality decoder that exceeded the performance of a supervised decoder.

Implication: Existing approaches have focused on comparing a measure of neural activity with an external measurement of behavior or stimuli. However, one could imagine pairing different streams of neural activity data like ECoG and EEG to provide insights into the shared information of neural recording modalities across different temporal or spatial scales. For example, the approach in `Swap-VAE` can be viewed as decomposing the representation into common and private spaces (9), and could be used to find a common space of information shared across modalities and private spaces which capture modality-specific variability.

3.5 Adversarial augmentations

Finally, we consider adversaries that try to generate views that obscure our ability to decode downstream information. The idea of adversarial augmentations has been studied extensively in vision (36; 37; 38) and more recently in graphs (39). In domains like neuroscience where it is less clear how information is encoded, an adversary that impairs the network's performance on a downstream decoding task could tell us something about how information relevant to that task is encoded.

Implication: The design of adversarial augmentations could provide insights into which neurons are the conduits of information in a neural circuit and whether these messenger neurons remain consistent or switch across time. Much like the questions posed in our discussion of dropout (Section 3.2), we can also use adversarial augmentations to gain insights into how distributed or localized the neural representation of different behaviors can be in different conditions.

4 Discussion

This article provides a starting point for thinking about how self-supervision may yield new insights in neuroscience. Self-supervised learning is still a nascent field, and we are only beginning to explore

its applications to neuroscience. Yet, considering the immense quantity of neural recordings currently being produced, it seems clear that the appeal of weakly or entirely unsupervised approaches – of which SSL is arguably the most promising – will only grow.

Much of this article focused on effective augmentation strategies, which are at the core of the success of SSL. From a certain perspective, we have put the cart before the horse; after all, if we knew exactly what neural encoding schema are invariant to, we would have a level of understanding of neural circuits which neuroscience has been aspiring to for a century. The augmentations we examined are based on a few simple ideas – that representations are distributed, temporally smooth, and preserved across similar behaviors – which are in keeping with everything we know about neural encoding, and it is difficult to imagine an encoding scheme which does not rely on them. Moving forward, we expect many approaches will follow `MYOW` and develop clever mechanisms to identify similar examples (across time or even individuals), perhaps guided by small quantities of labeled data.

We eagerly anticipate the insights to be gleaned from these models. One tantalizing prospect is to compare across states of consciousness, by comparing neural activity during wakeful behavior with replays of the same behavior during REM sleep, *and perhaps even to use these dream states as augmentations*. It was recently proposed that dreams could help the brain to generalize and achieve robustness (40), and it would be satisfying to use dreaming to engineer generalizable and robust brain decoders.

References

[1] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[2] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

[3] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," *arXiv preprint arXiv:2006.07733*, 2020.

[4] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," 2021.

[5] H. Banville, I. Albuquerque, A. Hyvärinen, G. Moffat, D.-A. Engemann, and A. Gramfort, "Self-supervised representation learning from electroencephalography signals," in *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2019.

[6] H. Banville, O. Chehab, A. Hyvarinen, D. Engemann, and A. Gramfort, "Uncovering the structure of clinical EEG signals with self-supervised learning," *Journal of Neural Engineering*, 2020.

[7] J. Y. Cheng, H. Goh, K. Dogrusoz, O. Tuzel, and E. Azemi, "Subject-aware contrastive learning for biosignals," *arXiv preprint arXiv:2007.04871*, 2020.

[8] M. Azabou, M. G. Azar, R. Liu, C.-H. Lin, E. C. Johnson, K. Bhaskaran-Nair, M. Dabagia, K. B. Hengen, W. Gray-Roncal, M. Valko, *et al.*, "Mine your own view: Self-supervised learning through across-sample prediction," *arXiv preprint arXiv:2102.10106*, 2021.

[9] R. Liu, M. Azabou, M. Dabagia, C.-H. Lin, M. G. Azar, K. B. Hengen, M. Valko, and E. L. Dyer, "Drop, swap, and generate: A self-supervised approach for generating neural activity," *bioRxiv*, 2021.

[10] S. M. Peterson, R. P. Rao, and B. W. Brunton, "Learning neural decoders without labels using multiple data streams," *bioRxiv*, 2021.

[11] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[12] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, "On variational bounds of mutual information," in *International Conference on Machine Learning*, pp. 5171–5180, PMLR, 2019.

[13] D. Guo, B. A. Pires, B. Piot, J.-b. Grill, F. Altché, R. Munos, and M. G. Azar, "Bootstrap latent-predictive representations for multitask reinforcement learning," *arXiv preprint arXiv:2004.14646*, 2020.

[14] X. Chen and K. He, "Exploring simple siamese representation learning," *arXiv preprint arXiv:2011.10566*, 2020.

[15] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "Bendr: using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data," *arXiv preprint arXiv:2101.12037*, 2021.

[16] Y. Zhu, Y. Xu, F. Yu, Q. Liu, S. Wu, and L. Wang, "Deep graph contrastive representation learning," *arXiv preprint arXiv:2006.04131*, 2020.

[17] A. Nayebi, N. C. Kong, C. Zhuang, J. L. Gardner, A. M. Norcia, and D. L. Yamins, "Unsupervised models of mouse visual cortex," *bioRxiv*, 2021.

[18] S. Bakhtiari, P. Mineault, T. Lillicrap, C. C. Pack, and B. A. Richards, "The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning," *bioRxiv*, 2021.

[19] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1422–1430, 2015.

[20] I. Misra, C. L. Zitnick, and M. Hebert, "Shuffle and learn: unsupervised learning using temporal order verification," in *European Conference on Computer Vision*, pp. 527–544, Springer, 2016.

[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[22] X. Bouthillier, K. Konda, P. Vincent, and R. Memisevic, "Dropout as data augmentation," *arXiv preprint arXiv:1506.08700*, 2016.

[23] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[24] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6023–6032, 2019.

[25] S. Saxena and J. P. Cunningham, "Towards the neural population doctrine," *Current Opinion in Neurobiology*, vol. 55, pp. 103–111, 2019.

[26] Z. Ma, G. G. Turrigiano, R. Wessel, and K. B. Hengen, "Cortical circuit dynamics are homeostatically tuned to criticality in vivo," *Neuron*, vol. 104, no. 4, pp. 655–664, 2019.

[27] J. F. Mitchell, K. A. Sundberg, and J. H. Reynolds, "Spatial attention decorrelates intrinsic activity fluctuations in macaque area v4," *Neuron*, vol. 63, no. 6, pp. 879–888, 2009.

[28] M. R. Cohen and J. H. Maunsell, "Attention improves performance primarily by reducing interneuronal correlations," *Nature Neuroscience*, vol. 12, no. 12, p. 1594, 2009.

[29] C. J. McAdams and J. H. Maunsell, "Effects of attention on orientation-tuning functions of single neurons in macaque cortical area v4," *Journal of Neuroscience*, vol. 19, no. 1, pp. 431–441, 1999.

[30] D. McCreery, V. Pikov, and P. R. Troyk, "Neuronal loss due to prolonged controlled-current stimulation with chronically implanted microelectrodes in the cat cerebral cortex," *Journal of Neural Engineering*, vol. 7, no. 3, p. 036005, 2010.

[31] V. S. Polikov, P. A. Tresco, and W. M. Reichert, "Response of brain tissue to chronically implanted neural electrodes," *Journal of Neuroscience Methods*, vol. 148, no. 1, pp. 1–18, 2005.

[32] W. M. Grill, S. E. Norman, and R. V. Bellamkonda, "Implanted neural interfaces: biochallenges and engineered solutions," *Annual Review of Biomedical Engineering*, vol. 11, pp. 1–24, 2009.

[33] A. Prasad, Q.-S. Xue, V. Sankar, T. Nishida, G. Shaw, W. J. Streit, and J. C. Sanchez, "Comprehensive characterization and failure modes of tungsten microwire arrays in chronic neural implants," *Journal of Neural Engineering*, vol. 9, no. 5, p. 056015, 2012.

[34] A. Sridharan, S. D. Rajan, and J. Muthuswamy, "Long-term changes in the material properties of brain tissue at the implant–tissue interface," *Journal of Neural Engineering*, vol. 10, no. 6, p. 066001, 2013.

[35] H. Alwassel, D. Mahajan, B. Korbar, L. Torresani, B. Ghanem, and D. Tran, "Self-supervised learning by cross-modal audio-video clustering," *arXiv preprint arXiv:1911.12667*, 2019.

[36] X. Zhang, Q. Wang, J. Zhang, and Z. Zhong, "Adversarial autoaugment," *arXiv preprint arXiv:1912.11188*, 2019.

[37] C. Luo, H. Mobahi, and S. Bengio, "Data augmentation via structured adversarial perturbations," *arXiv preprint arXiv:2011.03010*, 2020.

[38] C. Gong, T. Ren, M. Ye, and Q. Liu, "Maxup: Lightweight adversarial training with data augmentation improves neural network training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2474–2483, 2021.

[39] K. Kong, G. Li, M. Ding, Z. Wu, C. Zhu, B. Ghanem, G. Taylor, and T. Goldstein, "Flag: Adversarial data augmentation for graph neural networks," *arXiv preprint arXiv:2010.09891*, 2020.

[40] E. Hoel, "The overfitted brain: Dreams evolved to assist generalization," *Patterns*, vol. 2, no. 5, p. 100244, 2021.