Geometry of Shared Representations

Gregory Henselman-Petrusek[1] (gh10@princeton.edu)

Simon Segert[1] (ssegert@princeton.edu)

Bryn Keller[2] (bryn.keller@intel.com)

Mariano Tepper[2] (mariano.tepper@intel.com)

Jonathan D. Cohen[1] (jdc@princeton.edu)

[1] Princeton Neuroscience Institute, Washington Road, Princeton, New Jersey, 08540

[2] Intel Labs, NE 25th Avenue, Hillsboro, Oregon, 97124

Abstract

Advances in the use of neural networks in both cognitive neuroscience and machine learning have generated new challenges: while they have proven powerful at learning complex tasks, *what* they learn and *how* they come to perform those tasks often remains a mystery. Here, we examine a novel approach to these challenges, inspired by recent spatial and algebraic analyses of abstraction and generalization in network architectures. We evaluate it, and compare it to other measures, by using it to test theoretical predictions regarding the influence that training has on the development of shared vs. separated representations, and their impact on network performance. We find that the proposed measure outperforms all others in identifying a theoretically predicted, low dimensional set of linear spatial relationships that, in turn, best predict network performance.

Keywords: cognitive representation

Introduction

The application of deep learning methods to artificial neural networks has led to remarkable progress in solving difficult problems in machine learning (e.g., computational vision and natural language processing). However, the solutions that these systems arrive at often remain difficult to interpret in lower dimensional forms than the networks themselves (Li, Farkhoor, Liu, & Yosinski, 2018; Bengio, Courville, & Vincent, 2013; Bau, Zhou, Khosla, Oliva, & Torralba, 2017).A similar problem arises when using deep learning as models of cognitive and/or brain function: even when a network can be trained to perform a given task or set of tasks, it is not always clear what this reveals about *how* the task is performed, and/or whether it is being done in the same way as the human brain. This presents an impediment in both engineering (e.g., in generalization) and for understanding natural systems (e.g., whether the solutions correspond to those used by the brain (Yamins et al., 2014; Kriegeskorte, Mur, & Bandettini, 2008; Han et al., 2019; Horikawa & Kamitani, 2017; Seeliger, Güçlü, Ambrogioni, Güçlütürk, & Van Gerven, 2018; Wen et al., 2017)). These challenges have motivated the search for analysis methods that can determine whether a high dimensional network has learned more compressed, lower dimensional representations and, if so, can explicitly identify them.

The most common current approaches use simple statistical procedures, such as cross-correlations of activity patterns or connection weights, and/or principal components analysis (PCA) based on those. While often useful, these methods generally make strong assumptions about the underlying distributions that are generally violated by neural networks. Motivated by these considerations, we consider a novel geometric approach, based on work recently reported by Bernardi et al. (2018). To evaluate the success of this method, we use a simple three-layered network, and a theoretically-motivated training paradigm. The theory predicts well-specified forms of structure that should emerge in the hidden layer of a network under particular training conditions, that can be empirically validated in network performance. Experiments indicate that the new approach successfully captures this predicted structure, and provides a rich body of new geometric data.

Background

Previous theoretical work has identified a fundamental tension between shared representations that facilitate learning and generalization, and separated representations that facilitate simultaneous execution of multiple processes (Feng, Schwemmer, Gershman, & Cohen, 2014; Musslick, Dey, & Musslick, 2017; Alon et al., 2017). The former is exploited by standard multi-task training strategies in deep learning (Caruana, 1997; Bengio et al., 2013), while the latter is exploited by traditional multiprocessor architectures ("embarrassing parallelism") (Jin et al., 2011). This same tension is thought to explain the distinction between controlled and automatic processing in humans, in which shared representations enable rapid, flexible generalization to novel task domains, at the expense of serial, control-dependent processing; while separated, task-dedicated representations support efficient, parallel execution, but take longer to learn (Sagiv, Musslick, Niv, & Cohen, 2018).Previous work has explored this tension indomains that share a set of input and output dimensions, and each task involves a mapping of information from a particular input dimension to a particular output dimension. This work has shown that training on tasks individually generally favors the formation of shared representations in the hidden layers of a network, which allows tasks that share a given in-

put dimension (but map that to different output dimensions) to profit from all training trials involving that dimension (Musslick et al., 2017). However, such shared representations can produce cross-talk if the network is required to perform more than one task at a time. As a consequence, when networks are explicitly trained to perform two or more tasks simultaneously (that is, on *concurrent multitasking*, as distinct from *multi-task learning*), they develop separated representations dedicated to each task. Measurements of performance are consistent with these predictions: networks trained on tasks sequentially learn more quickly than ones trained explicitly to multitask, but suffer when tested for multitasking performance; conversely, training on simultaneous performance takes longer, but leads to efficient multitasking capability. Despite these observations, it has been difficult to directly confirm the predicted forms of representation presumed to be responsible for these effects. This invites new methods of network analysis.

In the sections that follow, we describe a novel approach to analyzing the geometry of hidden layer representations, apply it to networks trained as outlined above, and compare it with other standard approaches to network analysis in their ability both to identify theoretically predicted representational structure, and its association with corresponding patterns of performance.

Methods

Network architectures

For simplicity, we focus on feed-forward neural networks with the following architecture (see Figure 1). Each network has two input layers, *stimulus* $\mathbf{x}$ and *task* $\mathbf{u}$, a single *output* layer $\mathbf{z}$, and one or more associative (*hidden*) layers. Units in the stimulus and output layers are subdivided into subgroups termed *dimensions*. For each pair (i, j) of input and output dimensions, we associate (i) a *task mapping* t_{ij} from activity patterns in i to activity patterns in j, and (ii) a unique *task unit* u_{ij} in the task input layer, taking values in $\{0, 1\}$. Given training data, composed of a set of pairs $(\mathbf{x}_i, \mathbf{z}_j)$ of activity patterns in i and j such that $t_{ij} : \mathbf{x}_i \mapsto \mathbf{z}_j$, the learning objective is to regress $\mathbf{x}_i$ into $\mathbf{z}_j$ whenever unit u_{ij} is active. Multiple units may be simultaneously active.

To perform a set of tasks $\{t_{i_0 j_0}, \ldots, t_{i_k j_k}\}$ means to realize the relevant mappings from input to output dimensions, while clamping all output dimensions other than $j_0, \ldots, j_k$ to zero.

Representations

Formal treatments of network representation have been developed in a number of disciplines, e.g. (Chung, Lee, & Sompolinsky, 2018). Here, we propose a variant of a novel geometric measure proposed by Bernardi et al. (2018), and evaluate its ability to identify forms of representation predicted to arise in response to training on single task vs. multitasking performance, as discussed above. Specifically, the theory predicts that single task training will lead to sharing of representations for input dimensions that are common to a set of tasks, whereas training on multitasking will lead to separated representations, one for each task.

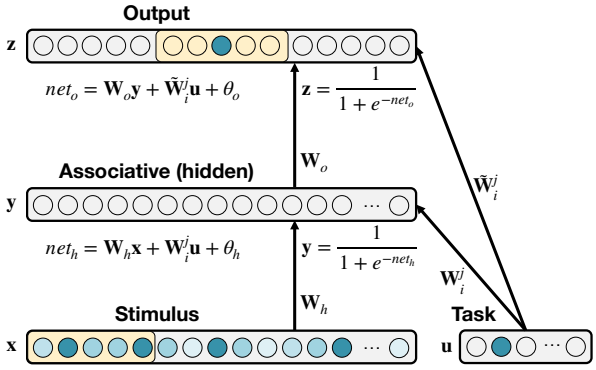

Figure 1: Network architecture from (Musslick et al., 2017). At initialization, we set additive biases $\theta_h = \theta_o = -2$. We update these biases and the weight matrices $\mathbf{W}_i^j$, $\tilde{\mathbf{W}}_i^j$, $\mathbf{W}_h$, and $\mathbf{W}_o$ via stochastic gradient descent and backpropagation.

For example, consider a task space compromised of two input dimensions (e.g., colors and words) and two output dimensions (e.g., verbal and manual responses). The following four tasks can be defined: color naming, word reading, color pointing (e.g., point left for red and right for green) and, analogously, word pointing. There are two extreme forms of hidden layer representation a network could learn for performing these four tasks. It could learn a single set of representations for colors that are mapped to both spoken and manual responses, and similarly for words. Task units would then select the appropriate input representations (color or word) and output dimension for a given task. This efficiently exploits a single set of representations for each input dimension, but precludes performing more than one task at a time. For example, simultaneously color naming and word pointing elicits cross talk from color pointing and word reading. This can be solved the other extreme of representation: dedicating a set of input representations to each task (i.e., combination of input and output dimensions); for example, different representations of colors for spoken and manual responses. While less efficient and longer to learn, this permits simultaneous performance of color naming and word pointing (and conversely, color pointing and word reading).

Table 1 shows these schemes, where each set of representations is respresented as a "node." Each entry in the tuples indicates activity of the associated nodes, 1 for high activity and 0 for low. These tuples can be visualized as points in a vector space. The points from Network 1 (shared representations) lie on a 2d plane, and, in particular, on the vertices a parallelogram (Figure 2, left). Approximate arrangement of points For Network 2 (separated representations) is shown in Figure 2, right, though an exact representation in 3 dimensions is not possible.

Consider a network that begins like Network 1, and gradually transitions to that of Network 2 (here we identify the length-4 tuples of Network 1 with length-6 tuples by appending

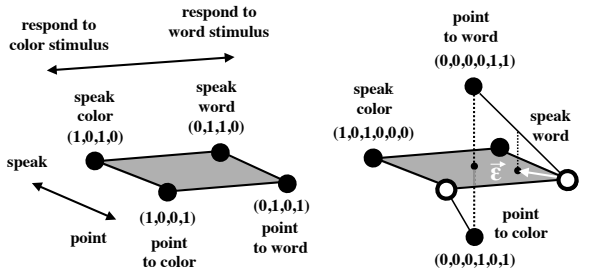

Figure 2: Shared and distributed coding schemes. The white displacement vector $\vec{\varepsilon}$ represents a projection onto the plane an an intermediate point on the path between two representations for word pointing.

zeros to the end). Suppose that the tuples for speaking stay fixed, while the those for pointing move along straight paths to their destinations, as indicated in Figure 2. Early on, the network will be vulnerable to cross-talk if it attempts to perform multiple tasks simultaneously, since, for example, the nonzero elements of word pointing and color pointing overlap. Later this effect will diminish, as the entries in the first three slots of the word pointing pattern descend to 0. In general, the degree of cross-talk should scale proportionally with the overlap in these three entries.

This idea represents the fundamental geometric intuition for the present work. In practice, the notions of tuple and entry become unwieldy. To compensate, we project not onto the first three coordinate axes, as in the idealized example, but instead onto the 2d plane spanned by the parallelogram. This method proves powerful.

Table 1: Task specifications for two neural networks. Letters C, W, S, and P denote dedicated nodes for color, word, speech, and pointing in Network 1. Symbol C_S denotes the node for color in Network 2 that is specifically dedicating to speaking. Symbols W_S, C_P, and W_P are defined similarly.

	Network 1	Network 2
Task	(C,W,S,P)	$(C_S, W_S, S, C_P, W_P, P)$
speak color	$(1,0,1,0)$	$(1,0,1,0,0,0)$
speak word	$(0,1,1,0)$	$(0,1,1,0,0,0)$
point color	$(1,0,0,1)$	$(0,0,0,1,0,1)$
point word	$(0,1,0,1)$	$(0,0,0,0,1,1)$

Experiments

Cross validation We first validated the measure against network performance in simultaneous multitasking, as defined by mean square error (MSE). Ten variants of the network were generated by varying parameters (e.g., number of hidden units and in/out dimensions, L2 weight regularization, and training corpus). For each variant, 10 networks with random initial weights were trained to perform tasks in sequence, and another 10 to perform them simultaneously. Degree of representational separation was measured using eight variants of the proposed method, and compared with standard similarity measures (e.g., Euclidean distance and correlation) using two criteria: (i) ability of a standard linear classifier to predict training condition (single vs. multi) from the measure, and (ii) its Pearson correlation with multitasking MSE. Among these, only the proposed measure and PCA-driven dimension estimates (e.g. inverse participation ratio) maintained classification accuracy above 95% and Pearson correlation above 0.75 across all benchmarks.

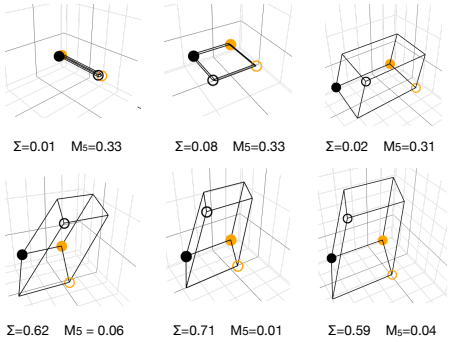

Figure 3: Variation in representation geometry across benchmark data sets. Each plot displays an isometric embedding of 4 mean task activation patterns. The input dimension of each task is indicated by color, output by solid/open circles. Symbols M_5 and Σ denote 5-way multitasking MSE and the proposed sharing measure, respectively.

Dynamic tracking We also tested precision in tracking evolution of network structure across time under two distinct training regimes. In the first experiment, the proposed measure accurately tracked MSE (mean Pearson correlation with MSE over training: 0.97 for 5-way multitasking; 0.99 for 2-way multitasking). Both curves were essentially monotonic, however, raising a concern that correlation could reflect merely motion in a consistent direction over time. To test we applied a more complex regime characterized by alternating periods of strictly-single and strictly-multitask training. As hoped, this regime induced non-monotonic trajectories due to periodic "catastrophic interference." Pearson correlation with MSE loss curve and geometric sharing remained high (0.96 for both 2- and 5-way multitasking) over 10 networks. More significantly, the sharing measure captured salient monotone features of the loss curve, see Figure 4.

Conclusion

We present a novel extension of a recent geometric approach to study spatial and functional relationships between distinct cognitive representations. The results prove effective in tracking system-level phenomena such as 5-way multitasking performance and catastrophic interference, and offer rich new

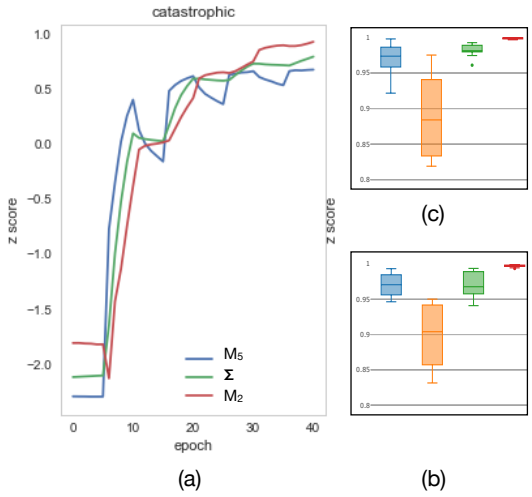

Figure 4: (a) Learning trajectory under a training regime designed to induce catastrophic interference. Traces are Z-scored values of 5-way multitasking MSE (M_5), geometric representation sharing (Σ), and 2-way multitasking MSE (M_2). MSE scores were first multiplied by -1 for ease of visual alignment. Curves are averaged over 10 networks. (b) Variation between networks under the training regime associated to (a). In each network, two-way multitasking MSE was correlated with PCA-based dimension estimates - a common proxy for sharing. These estimates included inverse participation ratio across all tasks (blue), and averaged over 4-element task subsets (orange), the proposed sharing measure (red), and the norm of the displacement vector from which the proposed measure is taken (as the horizontal component). (c) Similar to (b), for the 10 networks trained under the original, non-catestrophic training regime.

shape statistics to describe network geometry. While mutually validating, these statistics complement and expand current methods of compression and network analysis.

Acknowledgements

Supported in part by the Swartz Center for Theoretical Neuroscience at Princeton University.

References

Alon, N., Reichman, D., Shinkar, I., Wagner, T., Musslick, S., Cohen, J. D., ... others (2017). A graph-theoretic approach to multitasking. In *NIPS Proceedings* (pp. 2097–2106).

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 6541–6549).

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828.

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, D. (2018, September). *The geometry of abstraction in hippocampus and prefrontal cortex* (Preprint). Neuroscience. doi: 10.1101/408633

Caruana, R. (1997). Multitask learning. *Machine learning*, *28*(1), 41–75.

Chung, S., Lee, D. D., & Sompolinsky, H. (2018, Jul). Classification and geometry of general perceptual manifolds. *Phys. Rev. X*, *8*, 031003. doi: 10.1103/PhysRevX.8.031003

Feng, S. F., Schwemmer, M., Gershman, S. J., & Cohen, J. D. (2014, March). Multitasking versus multiplexing: Toward a normative account of limitations in the simultaneous execution of control-demanding behaviors. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(1), 129-146. doi: 10.3758/s13415-013-0236-9

Han, K., Wen, H., Shi, J., Lu, K.-H., Zhang, Y., Fu, D., & Liu, Z. (2019). Variational autoencoder: An unsupervised model for encoding and decoding fmri activity in visual cortex. *NeuroImage*.

Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, *8*, 15037.

Jin, H., Jespersen, D., Mehrotra, P., Biswas, R., Huang, L., & Chapman, B. (2011). High performance computing using mpi and openmp on multi-core parallel systems. *Parallel Computing*, *37*(9), 562–575.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 4.

Li, C., Farkhoor, H., Liu, R., & Yosinski, J. (2018). Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*.

Musslick, S., Dey, B., & Musslick, S. (2017). A Formal Approach to Modeling the Cost of Cognitive Control A Formal Approach to Modeling the Cost of Cognitive Control. (July).

Sagiv, Y., Musslick, S., Niv, Y., & Cohen, J. D. (2018). Efficiency of learning vs. processing: Towards a normative theory of multitasking. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (pp. 1004–1009).

Seeliger, K., Güçlü, U., Ambrogioni, L., Güçlütürk, Y., & Van Gerven, M. (2018). Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, *181*, 775–785.

Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., & Liu, Z. (2017). Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, *28*(12), 4136–4160.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.