An overview of functional alignment in artificial and biological neural networks: Current recommendations and open questions

Elizabeth DuPre (elizabeth.dupre@mail.mcgill.ca)
Montreal Neurological Institute, McGill University
Montreal, QC, Canada

Jean-Baptiste Poline (jean-baptiste.poline@mcgill.ca)
Montreal Neurological Institute, McGill University
Montreal, QC, Canada

Abstract

Functional alignment is a method for finding similarity in functional representations of both biological and artificial neural networks. Although it is actively developed in cognitive neuroscience and deep learning, each field prefers its own terminology for and variants of this method. There is, therefore, relatively little cross talk between the two spaces. In this brief review, we highlight three functional alignment methods successfully used in both fields: canonical correlation analysis, Procrustes analysis, and shared response modelling. We consider the relative strengths of each method and highlight situations in which each may be most appropriate. We conclude with open questions in functional alignment that may serve as collaborative opportunities for cognitive neuroscience and deep learning.

Keywords: functional alignment, canonical correlation, network similarity

Introduction

One of the fundamental challenges for cognitive neuroscience is to find similarity across neural diversity (Churchland, 1998); that is, to find shared or similar neural processes supporting the diversity of individual cognitive experience. This goal is not unique to cognitive neuroscience, however, and is in fact shared across biological and artificial neural networks. Indeed, it can be considered more generally as a problem of aligning functional representations. For the purposes of this work, we can define functional representations broadly as the parameterization of internal states of a neural system that carry informational content and thereby play a functional role (Bechtel, 1998). Practically, we can treat them as activation vectors within a high-dimensional space defined by e.g., the neurons or voxels of the network (Churchland, 1998). In deep learning, multiple random instantiations of the same neural network architecture on the same data set will yield different layerwise functional representations (Li, Yosinski, Clune, Lipson, & Hopcroft, 2015). In neuroscience, anatomical variability and poor structure-function correspondence across association cortex (Rodriguez-Vazquez et al., 2019; Paquola et al., 2019) yields misaligned functional representations across subjects for an identical stimulus, even following state-of-the-art anatomical normalization.

Despite the immediate potential of functional alignment methods, these tools are underutilized and often misunderstood within each field. Here, we review three methods used in functionally aligning both artificial and biological neural networks: Canonical Correlation Analysis (CCA), Procrustes analysis (also known in the neuroscience literature as hyperalignment), and Shared Response Modelling (SRM). Expanding on Barrett, Morcos, and Macke (2019), we argue that functional alignment is a promising direction for collaboration between deep learning and cognitive neuroscience. We note open questions in current formulations of functional alignment and suggest future research directions that may benefit both fields.

Canonical correlation analysis

As proposed by Hotelling (1936), Canonical Correlation Analysis (CCA) was originally designed to deal with multi-view samples where we have two views on the same data; for example, audio and visual recordings of the same speaker.

For input matrices $X \in \mathbb{R}^{n \times p_1}$ and $Y \in \mathbb{R}^{n \times p_2}$, where n is the number of samples (e.g., time points in fMRI), and p_1, p_2 are the number of units (e.g., neurons or voxels) for each network. Interestingly, the dimensionality of these matrices varies dramatically across fields, with neuroscience applications often considering n and p to be in the range of 100-1000, while deep learning applications consider n and p values in the range of 10,000-100,000.

When $p_1 \leq p_2$, CCA derives a vector of canonical correlation coefficients $\rho = \langle \rho_1, \rho_2, ... \rho_{p_1} \rangle$. We can assume that the matrices have been pre-processed to center their columns. For a given index i, then, ρ_i can be defined as

$$\begin{aligned} \rho_i = \max_{r_X^i, r_Y^i} & \text{corr}(\mathrm{Xr_x^i}, \mathrm{Yr_Y^i}) \\ \text{subject to } & \forall_{j<i} \; Xr_X^i \perp Xr_X^j \\ & \forall_{j<i} \; Yr_Y^i \perp Yr_Y^j \end{aligned} \tag{1}$$

We can also consider this maximizing correlation as minimizing distance (Xu, Lorbert, Ramadge, Guntupalli, & Haxby, 2012), in which case we can re-write CCA as

$$\begin{aligned} \min_{R_X, R_Y} & \; \|XR_X - YR_Y\|_F^2 \\ \text{subject to } & \; R_X^T X^T X R_X = I \\ & \; R_Y^T Y^T Y R_Y = I \end{aligned} \tag{2}$$

In functional alignment of biological or artificial neural networks, where X and Y are sub-sampled from two subjects (in the case of biological networks) or two initializations (in the case of artificial networks), some concerns emerge in using generic CCA. In particular, since CCA maximizes these canonical correlation coefficients, if the two data sources share correlated noise we will learn a joint representation driven by noise rather than signal. This is especially a concern for functional magnetic resonance imaging (fMRI) data with its low signal-to-noise ratio. Variants such as projection-weighted CCA (Morcos, Raghu, & Bengio, 2018) and L2-regularized CCA (Bilenko & Gallant, 2016) are thus designed to reduce the influence of noise, though they adopt different strategies in doing so.

These and related CCA variants have been successfully used in functionally aligning both fMRI data (Bilenko & Gallant, 2016) as well as deep neural networks (Raghu, Gilmer, Yosinski, & Sohl-Dickstein, 2017; Morcos et al., 2018). Recent work, however, has begun to ask whether CCA's invariance to invertible linear transformations is a desirable property in assessing similarity (Kornblith, Norouzi, Lee, & Hinton, 2019). In particular, Kornblith et al. (2019) show that for $p_2 \geq n$ data sets (e.g., wide convolutional network layers with more neurons than examples in the training data set) similarity indices that are invariant to invertible linear transforms gives the same result. This is a common situation in neuroimaging, where the number of voxels is often much greater than the number of available examples.

Procrustes analysis

Named for Procrustes, the ancient Greek innkeeper who stretched or cut off traveller's limbs so they would fit his bed, Procrustes analysis seeks to conform datasets through a series of rigid-body transformations (Schönemann, 1966). In the case where $p_1 = p_2$, we can define an orthogonal rotation matrix $R_X \in \mathbb{R}^{p \times p}$ such that we can

$$\begin{aligned} \min_{R_X} \ & \|XR_X - Y\|_F^2 \\ \text{subject to } & R_X^T R = I \end{aligned} \tag{3}$$

Although this is only defined in the case of exactly two data sets, Procrustes analysis has been extended to a generalized framework (Gower, 1975) wherein two or more data sets of the same dimensionality can be compared by first aligning to a reference subject and then iterating on this alignment. It was this Generalized Procrustes Analysis which was introduced to the neuroscience literature as *hyperalignment* in Haxby et al. (2011).

Procrustes analysis has been used successfully used both for aligning biological neural networks constructed from fMRI data (Haxby et al., 2011; Guntupalli et al., 2016) as well as artificial neural networks (Smith, Turban, Hamblin, & Hammerla, 2017). Two constraints emerge in applying Procrustes analysis to these data types, however. The first is that data sets must be of equivalent dimensionality. Thus, for example, convolutional neural network (CNN) hidden layers must have the same width to be aligned using Procrustes transformations. The second constraint is that each minimal unit (i.e., voxels in fMRI data or neurons in CNN hidden layers) is considered in the analysis, meaning that very large data sets often suffer from estimation problems. In particular, we need $\geq p$ samples for the estimation to be well-posed; this is rarely the case in fMRI studies, where our sampled time points $n \ll p$. To date, investigators have circumvented this issue by performing functional alignment only in anatomically- or functionally-defined regions of interest.

Shared response modelling

A more recently proposed method is Shared Response Modelling (SRM; P.-H. Chen et al., 2015). The intuition is that rather than aligning networks individually, we now want to develop a common basis set or coordinate system into which we can project additional networks.

Thus for m subjects, we want to learn an individual transformation basis $W \in \mathbb{R}^{p \times k}$ and a common or shared time series $S \in \mathbb{R}^{k \times n}$, where k is an experimenter-selected parameter to control the dimensionality of the model. As before, p is the number of units (e.g., neurons or voxels) in the network and n is the number of samples (e.g., time points in an fMRI analysis). Because all subjects are considered simultaneously in learning the shared response, the data matrix X now contains sub-matrices for each subject i such that $X_i \in \mathbb{R}^{p \times n}$. Note that since all subjects are included in X, there is thus no longer a need for the Y matrix. For subject i, then, we want to learn

$$\begin{aligned} \min_{W_i, S} \ & \Sigma_i \|X_i - W_i S\|_F^2 \\ \text{subject to } & W_i^T W_i = I_k \end{aligned} \tag{4}$$

For a fixed S, this formulation resembles (3) but with the transformation matrix–R_X in (3), W_i in (4)–now applied to the second term rather than the first. These two formulations are in fact equivalent when the experimenter-selected dimensionality k is equal to the number of minimal units (i.e., voxels or neurons) p_1. However, in the case where $k < p_1$, applying the transformation matrix directly to the subject data X_i leads to an uninformative shared response S (P.-H. Chen et al., 2015).

SRM has been successfully used in aligning both fMRI data (J. Chen et al., 2017) as well as deep neural networks (Lu et al., 2018). Like CCA, SRM has the advantage that the layer width or number of voxels considered does not need to be equivalent across networks. Similarly to SVCCA, a CCA variant developed by Raghu et al. (2017), learning the hyperparameter k also provides researchers an understanding of how many directions meaningfully contribute to the alignment. Nonetheless, the problem of hyperparameter selection requires cross-validation to assess its impact on the learned shared response, potentially requiring more data than available in standard analyses.

Current recommendations

Although each of the considered methods have been used in functionally aligning both artificial and biological neural net-

Table 1: Summary of functional alignment methods used to date.

Method	Quantify dimensionality	Tunable hyperparameters
Canonical Correlation Analysis	In SVCCA	In rCCA
Procrustes Analysis	✗	✗
Shared Response Modelling	✓	✓

works, their relative use differs dramatically across fields. CCA shows higher popularity in deep learning than in neuroscience, while Procrustes analysis (under the name *hyperalignment*) and SRM are more consistently used in neuroscience research. Although this disparity is due in part to non-overlapping terminology between the two fields, there are also field-specific constraints which in part guide these decisions. For example, one of the advantages of CCA is that data set sizes do not need to match exactly. This is more likely to appeal to deep learning researchers as it enables comparison of layers with different widths. Neuroscience researchers, however, are more likely to work with regions-of-interest from functional or anatomical parcellations that are standardized to the same number of voxels.

We hope, however, that this brief review will introduce researchers to the range of functional alignment methods available, enabling them to use those methods that best match their data set and research question. To this end, we have summarized some of the key features for each method in Table 1. Although these follow our understanding of functional alignment methods as they exist today, there are still several open questions which we draw attention to here.

Open questions and discussion

In considering current methods for functional alignment, at least two immediate questions arise. The first is what kind of similarity we should be assessing and what are the transformations to which these scores should be invariant; for example, whether we should allow for isotropic scaling of representations during alignment as in CCA, and therefore how to choose a similarity measure for a given use case. The second question is how to interpret calculated similarity. Deriving a "similarity score" could be useful for diagnosing network architecture and performance or for comparing experimental conditions; however, its interpretation after hyperparameter optimization is unclear. We review each of these questions in turn.

What kind of similarity metric should we use?

The question of what kind of similarity we should be examining is a fundamental one, with connections to many other mathematical fields such as clustering (Estivill-Castro, 2002). In their recent work, Kornblith et al. (2019) argue that similarity should not be invariant to invertible linear transformation. Besides the practical problem of data set size outlined above, choosing similarity metrics that are invariant to invertible linear transformation implies that the scale of activation space is irrelevant. That is, that representations that are only similar on small eigenvalues should have the same similarity index as representations that are only similar on large eigenvalues. The success of deep learning methods such as style transfer suggest that these distances are meaningful, however (Dumoulin, Shlens, & Kudlur, 2016). Neuroscience has only begun to quantify the dimensionality supporting similar representations (Ahlheim & Love, 2018), but we argue that a similar case is likely to hold for this field as well.

Should we define or improve similarity?

Hyperparameter selection in SRM or regularized CCA (Bilenko & Gallant, 2016) significantly improves our ability to transfer functional representations between networks. Unfortunately, it also obscures the definition of similarity. For example, many deep learning researchers use functional alignment in order to gain insight into the development of functional representations across training. In this case, a summary statistic of similarity can be meaningfully used to learn how different training regimes such as freeze training impact learned representations. If similarity is not only calculated between two networks, however, but optimized as in SRM then the interpretation of such a metric and its use across data sets is unclear.

Although a future alignment method may develop which preserves interpretability while maximizing similarity, we argue that such an extension is unlikely. Instead, we suggest that researchers carefully consider what they hope to learn from functionally aligning their networks and to choose a method which best meets their research goals with a clear understanding of the methods specificity and differences.

Conclusions

Functional alignment methods are being actively developed in both cognitive neuroscience and deep learning, though to date these research programs have been pursued largely in parallel. We argue, however, that there is substantial overlap and opportunities for collaboration in exploring the alignment of biological and artificial neural networks. Indeed, future investigations directly aligning these two kinds of networks seem close at hand. We hope that developing a common language for and implementations of these methods will inspire scientists to bridge this gap.

Acknowledgements

This work was supported in part by funding provided by Brain Canada, in partnership with Health Canada, for the Canadian Open Neuroscience Platform initiative. J.-B.P. was partially funded by NIH-NIBIB P41 EB019936 (ReproNim) NIH-NIMH R01 MH083320 (CANDIShare) and NIH 5U24 DA039832 (NIF), as well as the Canada First Research Excellence Fund, awarded to McGill University for the Healthy Brains for Healthy Lives initiative.

References

Ahlheim, C., & Love, B. C. (2018, October). Estimating the functional dimensionality of neural representations. *Neuroimage*, *179*, 51–62.

Barrett, D. G. T., Morcos, A. S., & Macke, J. H. (2019). Analyzing biological and artificial neural networks: challenges with opportunities for synergy? *Current Opinion in Neurobiology*, *55*, 55-64.

Bechtel, W. (1998, July). Representations and cognitive explanations: Assessing the dynamicist's challenge in cognitive science. *Cogn. Sci.*, *22*(3), 295–317.

Bilenko, N. Y., & Gallant, J. L. (2016, November). Pyrcca: Regularized kernel canonical correlation analysis in python and its applications to neuroimaging. *Front. Neuroinform.*, *10*, 49.

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017, January). Shared memories reveal shared structure in neural activity across individuals. *Nat. Neurosci.*, *20*(1), 115–125.

Chen, P.-H., Chen, J., Yeshurun, Y., Hasson, U., Haxby, J., & Ramadge, P. J. (2015). A Reduced-Dimension fMRI shared response model. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems 28* (pp. 460–468). Curran Associates, Inc.

Churchland, P. M. (1998). Conceptual similarity across sensory and neural diversity: the Fodor/Lepore challenge answered. *J. Philos.*, *95*(1), 5–32.

Dumoulin, V., Shlens, J., & Kudlur, M. (2016, October). A learned representation for artistic style.

Estivill-Castro, V. (2002, June). Why so many clustering algorithms: A position paper. *SIGKDD Explor. Newsl.*, *4*(1), 65–75. Retrieved from http://doi.acm.org/10.1145/568574.568575 doi: 10.1145/568574.568575

Gower, J. C. (1975, March). Generalized procrustes analysis. *Psychometrika*, *40*(1), 33–51.

Guntupalli, J. S., Hanke, M., Halchenko, Y. O., Connolly, A. C., Ramadge, P. J., & Haxby, J. V. (2016, June). A model of representational spaces in human cortex. *Cereb. Cortex*, *26*(6), 2919–2934.

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., ... Ramadge, P. J. (2011, October). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, *72*(2), 404–416.

Hotelling, H. (1936, December). Relations between two sets of variates. *Biometrika*, *28*(3-4), 321–377.

Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019, May). Similarity of neural network representations revisited.

Li, Y., Yosinski, J., Clune, J., Lipson, H., & Hopcroft, J. (2015). Convergent learning: Do different neural networks learn the same representations? In D. Storcheus, A. Rostamizadeh, & S. Kumar (Eds.), *Proceedings of the 1st international workshop on feature extraction: Modern questions and challenges at NIPS 2015* (Vol. 44, pp. 196–212). Montreal, Canada: PMLR.

Lu, Q., Chen, P.-H., Pillow, J. W., Ramadge, P. J., Norman, K. A., & Hasson, U. (2018, November). Shared representational geometry across neural networks.

Morcos, A., Raghu, M., & Bengio, S. (2018). Insights on representational similarity in neural networks with canonical correlation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems 31* (pp. 5727–5736). Curran Associates, Inc.

Paquola, C., De Wael, R. V., Wagstyl, K., Bethlehem, R. A., Hong, S.-J., Seidlitz, J., ... others (2019). Microstructural and functional gradients are increasingly dissociated in transmodal cortices. *PLOS Biology*, *17*(5), e3000284.

Raghu, M., Gilmer, J., Yosinski, J., & Sohl-Dickstein, J. (2017). SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In I. Guyon et al. (Eds.), *Advances in neural information processing systems 30* (pp. 6076–6085). Curran Associates, Inc.

Rodriguez-Vazquez, B., Suarez, L. E., Shafiei, G., Markello, R., Paquola, C., Hagmann, P., ... Misic, B. (2019). Gradients of structure-function tethering across neocortex. *BioRxiv*, 561985.

Schönemann, P. H. (1966, March). A generalized solution of the orthogonal procrustes problem. *Psychometrika*, *31*(1), 1–10.

Smith, S. L., Turban, D. H. P., Hamblin, S., & Hammerla, N. Y. (2017, February). Offline bilingual word vectors, orthogonal transformations and the inverted softmax.

Xu, H., Lorbert, A., Ramadge, P. J., Guntupalli, J. S., & Haxby, J. V. (2012, August). Regularized hyperalignment of multi-set fMRI data. In *2012 IEEE statistical signal processing workshop (SSP)* (pp. 229–232).