Modeling Brain Microarchitecture with Deep Representation Learning

Aishwarya H. Balwani [1] **Eva L. Dyer** [2 1]

Abstract

Models of neural architecture and organization are critical for many tasks in neuroscience. However, building these models in an automated, data-driven manner within and across varied brain regions still remains a challenge. In this work, we leverage the power of deep learning to build a rich model of neural microarchitecture across multiple, diversified brain areas. We then use low-rank matrix factorization to project the model's features onto an interpretable, lower-dimensional space. Our results show that the subsequent embeddings possess biologically meaningful structure which makes them useful in the study of brain structure at multiple scales. We demonstrate the use of this approach in the discovery of microstructural patterns and motifs within brain areas, as well as in revealing relationships between multiple heterogeneous brain regions.

1. Introduction

Mapping out the underlying microstructure of the brain is essential for many tasks in neuroscience (Mazziotta et al., 1995). For instance, in studies of disease (Rondina et al., 2018; Pflanz et al., 2020), aging (Tian & Ma, 2017), or development (Lebel & Deoni, 2018), a rich description of the microstructure is a pre-requisite for being able to compare brains across different conditions. Detailed maps of brain structure have also led to important discoveries regarding relationships between structure and function (Mountcastle, 1998) and provide a necessary sign post when targeting specific brain regions for subsequent studies.

The study of brain structure and organization has traditionally relied on human reasoning to define regions of interest (ROIs) (Brodmann, 1909), where neuroanatomists typically characterize parts of an imaged sample in terms of their anatomical compositions and then build a model of how the architecture changes across different brain regions. Moving forward however, given the ever-increasing sizes of neuroimaging datasets, as well as to further our knowledge of neural structure and organization beyond what is already understood, there is a need for automated solutions that can discover substructures within brain areas, along with new architectural primitives across different ROIs.

In the recent past, convolutional neural networks (CNNs) have proven to be particularly well suited for automated feature engineering and pattern recognition. These systems are designed to build and learn hierarchical, textural representations that are useful for solving downstream problems, directly from raw image data (Zeiler & Fergus, 2013; Olah et al., 2017; 2018; Lin & Maji, 2016). They have been shown to conclusively outperform classifiers trained on hand-crafted features, and are now routinely applied in modeling brain structure in macroscale datasets (Bernal et al., 2019; Lin et al., 2018). More recently, they have also been used with high resolution neuroanatomical data to solve problems ranging from tumor detection to pixel-level semantic segmentation in connectomics.

Deep learning applied to image understanding in neuroscientific microscopy, however, has so far primarily focused on supervised approaches for segmentation, either at the scale of brain regions (Chen et al., 2019; Iqbal et al., 2019; Tan et al., 2020), or individual components like neurites (Funke et al., 2018; Januszewski et al., 2018). And though these approaches are capable of learning rich features from images to solve their respective tasks, they are trained strictly to find specific components, thus failing to provide any tangible ways to discover new areas or structures of interest.

In this work we introduce a deep learning-based approach for modeling microstructure in brain imagery (Figure 1) that can be used to automatically discover regions within a brain sample that share similar characteristics in their local morphology or cytoarchitecture.

Our solution starts with a simple observation; If we train a network to do well on a brain area classification task using only *local* views of the brain's structure, then the network is forced to pay attention to the anatomical features of the dataset and build rich feature assemblies (e.g., patterning of axons, density and morphology of cells) along the way.

[1]School of Electrical & Computer Engineering, Georgia Institute of Technology [2]Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University. Correspondence to: Aishwarya H. Balwani <abalwani6@gatech.edu>.

Machine Learning Interpretability for Scientific Discovery (MLI4SD) Workshop, International Conference on Machine Learning, Vienna, Austria, 2020.

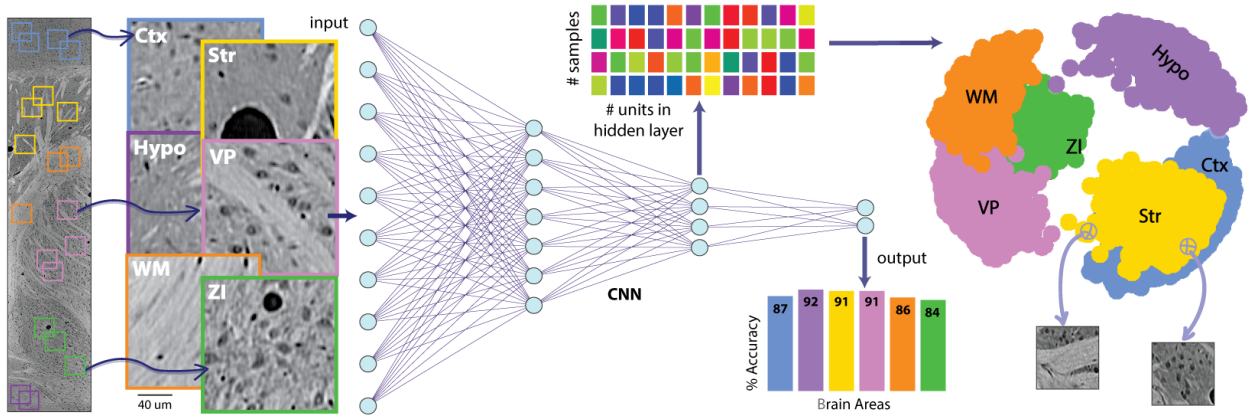

Figure 1. Deep feature learning approach for modeling brain microarchitecture and neuroanatomical discovery. On the left, we show how patches from different ROIs are selected from a large brain sample, and used to train a deep CNN that can classify these images in terms of their underlying ROI. Once the network is trained, we extract activations from its last hidden layer across many test samples and embed them into a low-dimensional space via low-rank matrix factorization. These new low-dimensional features are then used to further find patterns and regions of interest in a large brain sample.

After training a CNN to solve this task, we then treat it as a feature extractor (Bengio et al., 2013; Sermanet et al., 2014; Razavian et al., 2014) and collect the activations (i.e., representations) of units in its last hidden layer across many inputs. Because the network has been trained to provide meaningful information about the diverse brain structures it has previously seen, its representations provide useful cues about the relationships between different brain areas, as well as the changes or demarcations within them (e.g., layers, barrels and barreloids in cerebral cortex).

The network's representations are however, still rather high dimensional, entangled, and uninterpretable. To address these problems we project the representations onto a lower, k-dimensional space via non-negative matrix factorization (NMF) (Lee & Seung, 2001). NMF imposes constraints that encourage compositional localization and disentanglement in the transformed space, thus resulting in embeddings that i) provide a more fine-grained lens into how different regions in the sample are organized within the network, and ii) demonstrate improved microstructural disentanglement along their components, making them suitable as features in unsupervised downstream tasks aimed at region discovery.

We applied our framework for microstructure discovery to a large-scale thalamocortical sample that spans six different brain areas (Agmon & Connors, 1991) and was imaged with synchrotron X-ray microtomography (micro-CT) (Prasad et al., 2020a;b). On a macrostructural scale, our framework allowed us to identify directions in k-dimensional space that strongly aligned with certain ROIs in the thalamocortical slice, as well as some which revealed co-expression patterns that aligned with specific neuroanatomical features (e.g., regions with myelinated axons and little to no cells) across the entire image slice. On a microstructural scale, we combined our framework with a downstream clustering task, and successfully discovered both laminar differences and barrel fields in the cortex, without being given any prior knowledge about these motifs.

Our findings point to the fact that deep learning-based representations can be utilized to find finer sub-divisions and biological features in data without explicit supervision to do so. They speak to our framework's potential for application in the discovery of microarchitectural motifs in relatively unexplored and under studied brain areas, as well as open up possibilities for such methods to be translated into approaches for modeling continuous variability in brain structure and progression of neurodegenerative disease.

2. Methods

Dataset. In order to build an expressive model of local neural microarchitecture, it was imperative that we used image data that was structurally heterogeneous, as well as of sufficiently high resolution. We therefore used a 3D X-ray microtomography dataset that contains a varied set of microarchitectural structures including cell bodies of differing morphology and density, myelinated axons, and blood vessels, all resolved at 1.17 micron isotropic. From the raw data we sampled 150x150 micron images from each of the six manually annotated ROIs in the slice. Our resulting dataset provided us with the necessary scale and resolution required to capture the micorarchitectural anatomical differences across the brain areas of interest.

Training of the Deep Neural Network. Next, we trained a feed-forward CNN to perform six-way brain area classification using the cross-entropy loss function. Images in the training set were drawn from a single slice (z=159)

while those for validation and testing were drawn from brain slices 50 (z=209) and 100 (z=259) microns away from the training image, respectively. We found that the trained network performed well on both the validation (89.77%) and test (88.88%) datasets in terms of accuracy, thus providing evidence that the classifier could generalize to new images.

Extracting Network Activations. After training the network to discriminate between brain areas using local views, we then sought to use it to explore finer-scale microarchitectural characteristics within our data. To do so, we froze the network's weights and collected representations from its last hidden layer for image patches obtained by densely sampling across the full test slice (z=259). We then arranged these representations into a matrix $\mathbf{X} \in R^{d \times n}$ where $\mathbf{X} = f_\theta(\mathbf{D})$, $\mathbf{D}$ is our dataset with n samples, and $f_\theta(\cdot)$ denotes the transformation from image space to the d-dimensional space defined by last hidden layer of the network. In our experiments, $d = 64$ and $n \approx 5.5$M.

Decomposing Network Representations via Matrix Factorization. Once we had our set of network representations $\mathbf{X}$, we then needed to project the matrix onto a space where the transformed representations were more structured and biologically interpretable. We achieved this through non-negative matrix factorization, a low-rank approximation technique that factorizes the matrix $\mathbf{X}$ into two matrices $\mathbf{U}$ and $\mathbf{V}$ such that the residual $||\mathbf{X} - \mathbf{UV}||_\mathbf{F}$ is minimized, subject to $\mathbf{U}, \mathbf{V} \geq 0$. Here, the matrix $\mathbf{U} \in R^{d \times k}$ forms our basis that projects the d-dimensional data onto a lower k-dimensional space, and $\mathbf{V} \in R^{k \times n}$ is the matrix of coefficients obtained for all n examples in the dataset. Each of the columns of $\mathbf{V}$ explain how aligned each data sample in $\mathbf{X}$ is to the different basis vectors given by the columns of $\mathbf{U}$. A salient property of NMF is that its resulting factors exhibit an inherent spectral clustering on the columns (i.e., the samples) of the data matrix $\mathbf{X}$ and form sparse, localized embeddings (Ding et al., 2005) that are easy to interpret. This is a result of the non-negativity constraints that NMF imposes on its factors, thus making the technique extremely well suited for our task of simultaneously disentangling and reducing the dimensionality of the representations in $\mathbf{X}$.

Selecting Predictive Non-Negative Factors. Factors obtained using NMF unfortunately do not have a natural ordering that dictates their importance in either explaining the variance of or reconstructing the data. We therefore developed a greedy algorithm to choose a subset of p non-negative factors from the original set of k, such that those in the subset shared minimal information with each other while still being able to reconstruct as much of the entire sample as possible. The algorithm takes as inputs the original set of non-negative factors, number of factors p one wants to sub-select, annotations that demarcate pre-defined regions of interest in the sample and a tunable hyperparameter λ (default = 0.5) that encourages the selection of more localized or more uniformly distributed factors for higher and lower values respectively. It then computes two sets of scores, i.e., the coverage and leakage scores, for all combinations of the different non-negative factors and ROIs in the sample. For a factor i and ROI a whose set of points are given by the set A, the coverage score, $s_c = \sum 1_A . 1_{v_i' > 0}$ is the number of positive coefficients for the ROI associated with factor i. The leakage score, $s_\ell = \lambda \sum 1_{A^\mathbf{C}} . 1_{v_i' > 0}$ for the same factor and ROI is the number of positive coefficients outside the ROI and associated with factor i, scaled by λ. The difference between the two gives us the total score, $s_t = s_c - s_\ell$. Once the total scores for all k factors and r ROIs are calculated and ordered into a (k, r) matrix, the algorithm iteratively selects the factor with the highest score, given that score belongs to neither a factor nor an ROI selected previously.

3. Experiments and Results

After training the network and constructing our matrix of representations $\mathbf{X}$, we obtained the non-negative factors and coefficients across all ~ 5.5M samples taken together ($k = 15$). We then conducted two experiments, the first across all brain areas where we studied the neuroanatomical features that the different factors were aligned to, and the second within the cortex, where we looked at different structural motifs and sub-divisions within the area.

Identifying Anatomically linked Macrostructure. Using the strategy described in Section 2, we selected the top six mutually uncorrelated factors (Figure 2C) across the sample that aligned with the labeled brain areas in the dataset. We found that factors 12, 7 and 14 aligned with the cortex, striatum and ZI, respectively, while factors 11 and 5, 6 revealed co-expression patterns that aligned with specific neuroanatomical features viz., regions with myelinated axons with little to no cells for F5 and a general thalamic distribution in F11. Examination of Factor 11's coefficients highlighted parts of the sample that had patchy and diffuse axonal expression, mainly parts of VP and ZI innervated with myelinated axons. However, unlike F5 that highlighted WM specifically, F11 also seemed to require a joint distribution of cells. Further studying the distribution of F11 revealed that the factor showed activity in the thalamic regions (VP, ZI) and to some degree, the cortex. While joint co-expression of regions in the thalamus wasn't particularly startling, to see the same factor highlight parts of cortex was surprising. When we further looked into F11's expression in the cortex and examined the component correlation across manually annotated layers, we observed that the areas highlighted were in layers 4 and 6, both of which do indeed exhibit diffuse expression of axons. Picking out these regions of sparse axonal innervation by eye is not easy and the network appeared to have identified a good solution to this problem. This analysis therefore provided us

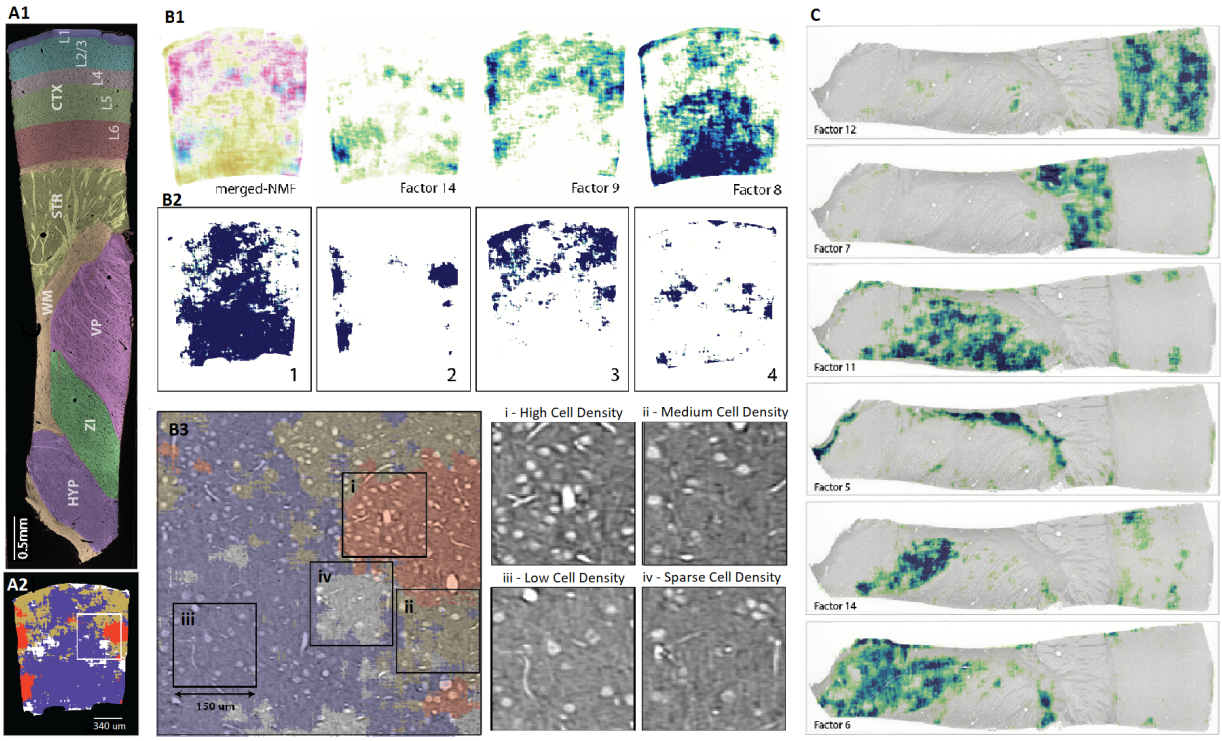

Figure 2. Discovery of neural structure and organization at multiple scales. In (A1) we show the manual annotations for all brain areas in the test sample, and the cortex is further divided into its different layers. (A2) shows the results from clustering the NMF factors of samples in the cortex. Panel (B) of the figure demonstrates our results for discovery of neuroanatomical motifs at the microstructural level in the cortex. In (B3) we zoom into the bounding box shown in (A2) and identify regions of high cellular density (i), medium cellular density (ii), low cellular (iii), and sparse cellular density (iv) that we find after clustering the 15D NMF cortical features. (B2) shows the four different clusters in (A2) individually and see that different factors highlight different characteristics of interest (e.g. barreloids in B2-2). (B1) shows the cortical embeddings for the top-3 NMF factors that are selected to be uncorrelated across the layers while still spanning most of the cortex, individually and combined in an RGB map. In panel (C) we show heatmaps obtained by visualizing the coefficients corresponding to the top-6 predictive NMF factors selected across the entire test slice. As is clearly visible, the factors very distinctly align with different ROIs (e.g. factors 12, 7 and 14) or specific neuroanatomical features (e.g. factors 5, 11)

with new insights into how different brain areas are micro-architecturally related, and how patterning of axons and cells appeared to be aligned with specific factors.

Revealing Laminar Divisions and Regions of Varying Cell Density in Cortex. Looking at the top three NMF factors of the cortex (Figure 2B1) revealed that they roughly mapped onto different cortical layers. Moreover, certain divisions in Layers 4 and 5 were particularly pronounced, and their patterns of architecture agreed with the descriptions of patterning of barrels and barreloids in somatosensory cortex (Petersen, 2007). Further analysis of the cortical factors by fitting a Gaussian mixture model (GMM) to them (Figure 2A2) revealed that a subset of the resulting components (Figure 2B2) explained much of the cortex. Significant chunks of layers 2/3 and 5/6 were split roughly across components 1 and 3, which primarily represented areas with moderate cell density and no axons. We also found that parts of the image that had high axon count in conjunction with high cell density were consistently grouped into component 1, both in Layer 4 and 6. Inspection of the raw image data at higher resolution additionally supplemented these findings and also clearly showed that the clusters were formed on the basis of underlying anatomical patterns (Figure 2B3). All of these results reaffirmed that our learnt representations were biologically meaningful and they can be used to discover anatomical patterns in the data.

4. Conclusion

In this work we described a representation learning framework that leverages the inherent expressiveness and low-dimensionality of a deep neural network's latent space to discover neuroanatomical primitives both within and across brain regions. Given the framework's generality and its ability to effectively model brain structure and organization, the described methods can easily be adapted to other open problems in comparative neuroanatomy and digital pathology, thus providing a promising path forward in discovering patterns in brain architecture with limited to no supervision.

References

Agmon, A. and Connors, B. Thalamocortical responses of mouse somatosensory (barrel) cortexin vitro. *Neuroscience*, 41(2-3):365–379, 1991.

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

Bernal, J., Kushibar, K., Asfaw, D. S., Valverde, S., Oliver, A., Martí, R., and Lladó, X. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. *Artificial Intelligence in Medicine*, 95: 64–81, 2019.

Brodmann, K. *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Barth, 1909.

Chen, Y., McElvain, L. E., Tolpygo, A. S., Ferrante, D., Friedman, B., Mitra, P. P., Karten, H. J., Freund, Y., and Kleinfeld, D. An active texture-based digital atlas enables automated mapping of structures and markers across brains. *Nature Methods*, 16(4):341, 2019.

Ding, C., He, X., and Simon, H. D. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pp. 606–610. SIAM, 2005.

Funke, J., Tschopp, F., Grisaitis, W., Sheridan, A., Singh, C., Saalfeld, S., and Turaga, S. C. Large scale image segmentation with structured loss based deep learning for connectome reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1669–1680, 2018.

Iqbal, A., Khan, R., and Karayannis, T. Developing a brain atlas through deep learning. *Nature Machine Intelligence*, 1(6):277, 2019.

Januszewski, M., Kornfeld, J., Li, P. H., Pope, A., Blakely, T., Lindsey, L., Maitin-Shepard, J., Tyka, M., Denk, W., and Jain, V. High-precision automated reconstruction of neurons with flood-filling networks. *Nature Methods*, 15 (8):605–610, 2018.

Lebel, C. and Deoni, S. The development of brain white matter microstructure. *NeuroImage*, 182:207–218, 2018.

Lee, D. D. and Seung, H. S. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 14*, pp. 556–562, 2001.

Lin, T.-Y. and Maji, S. Visualizing and understanding deep texture representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2791–2799, 2016.

Lin, W., Tong, T., Gao, Q., Guo, D., Du, X., Yang, Y., Guo, G., Xiao, M., Du, M., Qu, X., and Initiative, T. A. D. N. Convolutional neural networks-based mri image analysis for the alzheimer's disease prediction from mild cognitive impairment. *Frontiers in Neuroscience*, 12:777, 2018.

Mazziotta, J. C., Toga, A. W., Evans, A., Fox, P., and Lancaster, J. A probabilistic atlas of the human brain: theory and rationale for its development. *NeuroImage*, 2(2): 89–101, 1995.

Mountcastle, V. B. *Perceptual Neuroscience: The Cerebral Cortex*. Harvard University Press, 1998.

Olah, C., Mordvintsev, A., and Schubert, L. Feature visualization. *Distill*, 2(11):e7, 2017.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., and Mordvintsev, A. The building blocks of interpretability. *Distill*, 3(3):e10, 2018.

Petersen, C. C. The functional organization of the barrel cortex. *Neuron*, 56(2):339–355, 2007.

Pflanz, C. P., Charquero-Ballester, M., Majid, D. A., Winkler, A. M., Vallée, E., Aron, A. R., Jenkinson, M., and Douaud, G. One-year changes in brain microstructure differentiate preclinical huntington's disease stages. *NeuroImage: Clinical*, 25:102099, 2020.

Prasad, J., Balwani, A., Johnson, E., Miano, J., Sampathkumar, V., Andrade, V. D., Fezzaa, K., Du, M., Vescovi, R., Jacobsen, C., Kording, K. P., Gursoy, D., Roncal, W. G., Kasthuri, N., and Dyer, E. A three-dimensional thalamocortical dataset for characterizing brain heterogeneity: X-ray microCT images (Tiff). *figshare*, 2020a.

Prasad, J. A., Balwani, A. H., Johnson, E. C., Miano, J. D., Sampathkumar, V., de Andrade, V., Fezza, K., Du, M., Vescovi, R., Jacobsen, C., Kording, K. P., Gürsoy, D., Gray-Roncal, W., Kasthuri, N., and Dyer, E. L. A three-dimensional thalamocortical dataset for characterizing brain heterogeneity. *bioRxiv*, 2020b.

Razavian, A. S., Azizpour, H., Sullivan, J., and Carlsson, S. Cnn features off-the-shelf: An astounding baseline for recognition. *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 512–519, 2014.

Rondina, J. M., Ferreira, L. K., de Souza Duran, F. L., Kubo, R., Ono, C. R., Leite, C. C., Smid, J., Nitrini, R., Buchpiguel, C. A., and Busatto, G. F. Selecting the most relevant brain regions to discriminate alzheimer's disease patients from healthy controls using multiple kernel learning: A comparison across functional and structural imaging modalities and atlases. *NeuroImage: Clinical*, 17:628–641, 2018.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. *2nd International Conference on Learning Representations*, 2014.

Tan, C., Guan, Y., Feng, Z., Ni, H., Zhang, Z., Wang, Z., Li, X., Yuan, J., Gong, H., Luo, Q., et al. Deepbrainseg: Automated brain region segmentation for micro-optical images with a convolutional neural network. *Frontiers in Neuroscience*, 14, 2020.

Tian, L. and Ma, L. Microstructural changes of the human brain from early to mid-adulthood. *Frontiers in Human Neuroscience*, 11:393, 2017.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks (2013). *arXiv preprint arXiv:1311.2901*, 2013.