Modeling Variability in Brain Architecture with Deep Feature Learning

Aishwarya H. Balwani
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, USA
abalwani6@gatech.edu

Eva L. Dyer
Coulter Department of Biomedical Engineering
Georgia Institute of Technology
Atlanta, USA
evadyer@gatech.edu

***Abstract*—The brain has long been divided into distinct areas based upon its local microstructure, or patterned composition of cells, genes, and proteins. While this taxonomy is incredibly useful and provides an essential roadmap for comparing two brains, there is also immense anatomical variability within areas that must be incorporated into models of brain architecture. In this work we leverage the expressive power of deep neural networks to create a data-driven model of intra- and inter-brain area variability. To this end, we train a convolutional neural network that learns relevant microstructural features directly from brain imagery. We then extract features from the network and fit a simple classifier to them, thus creating a simple, robust, and interpretable model of brain architecture. We further propose and show preliminary results for the use of features from deep neural networks in conjunction with unsupervised learning techniques to find fine-grained structure within brain areas. We apply our methods to micron-scale X-ray microtomography images spanning multiple regions in the mouse brain and demonstrate that our deep feature-based model can reliably discriminate between brain areas, is robust to noise, and can be used to reveal anatomically relevant patterns in neural architecture that the network wasn't trained to find.**

***Index Terms*—Deep learning, convolutional neural networks, brain architecture, feature extraction, unsupervised learning.**

I. Introduction

Much of comparative neuroanatomy relies on our ability to effectively model the architecture of the brain, both at the level of individual anatomical structures (e.g. myelinated axons, blood vessels and cells) as well as entire brain areas (e.g. cortex, thalamus and striatum). Abstractions of various structures and regions of the brain then allow us to quantitatively characterize and compare brains across different ages, disease states or neurological conditions.

Unfortunately, effectively modeling brain architecture in a way that truly captures the heterogeneity in the distribution of various components across different brain areas is a difficult problem. Addtionally, problems in imaging such as noise and blur, as well as partial or misaligned fields of view only exacerbate the problem of being able to develop an expressive enough model of general brain structure [1]. As a results, traditional models of brain areas based on finding landmarks in different regions of the brain and feature extraction with pre-defined bases such as wavelets [2], [3] have worked with limited success, in large part because of the extensive experimentation and domain expertise required to develop the hand crafted features used in these models. With the advent of deep learning however, we now have ways of reliably learning features of neural microstructure directly from brain imagery [4].

In this work, we introduce an approach that leverages the power of deep neural networks as feature extractors to model the structural variability within and across brain areas (Figure 1). We approach this problem through the lens of discrimination, training a deep convolutional neural network (CNN) to distinguish different brain areas in a given sample. The resulting network provides us with a set of rich features which can be used to study the structure and continuous variability in different brain areas of interest.

Using this deep feature-based modeling approach, we then ask questions such as:

- How are different brain regions related to one another? By having measures of how brain areas are related to each other, we can identify as well as quantify their hierarchical organization [5].
- How likely is a sample to come from a specific brain area? An answer to this question can be used to potentially find outliers, or other diseased or abnormal regions in a sample [6], [7].
- Can further sub-divisions exist within a brain area of interest and if so, how can we find them? An answer to this question can be applied to further find architectural sub-divisions and micro-organizational patterns such as barrels and columns [8].

We apply our method to a microtomography dataset that spans multiple different brain areas, and show that our proposed deep feature learning based approach can indeed be used to not only discriminate between but also quantify the relationship between different brain areas directly from brain imagery. We also show the superiority of our method over using pre-trained networks for the same task by demonstrating that the latter couldn't provide the same quality of features for our tasks. Finally, we used our method to reveal insights in neural architecture by further subdividing a region of interest

Both AHB and ELD were supported by a Sloan Research Fellowship and NSF-IIS CRII award 1755871.

978-1-7281-4300-2/19/$31.00 ©2019 IEEE

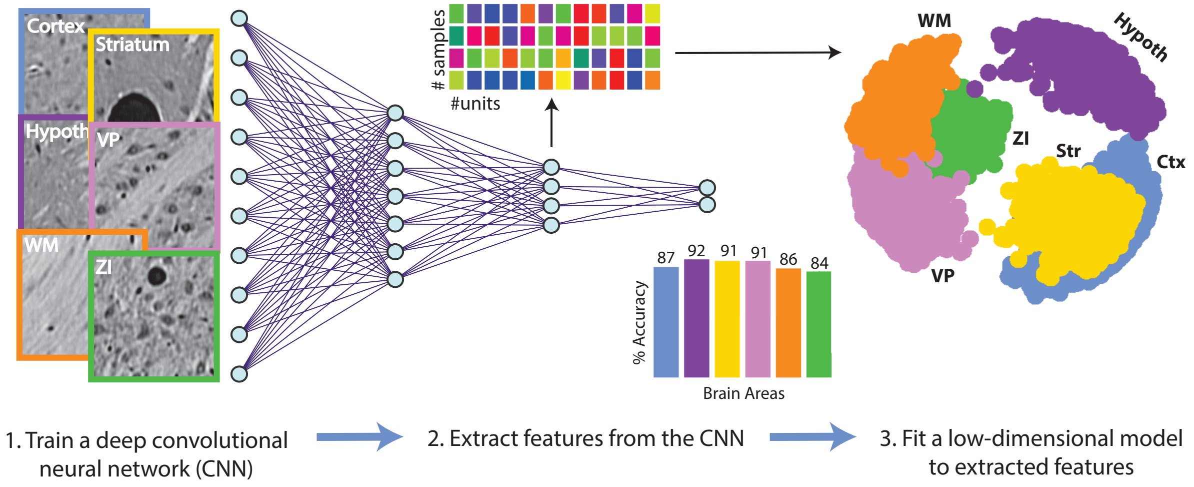
Cortex
Striatum
Hypoth
VP
WM
ZI
# samples
#units
87
92
91
91
86
84
% Accuracy
Brain Areas
WM
Hypoth
ZI
Str
Ctx
VP
1. Train a deep convolutional neural network (CNN)
2. Extract features from the CNN
3. Fit a low-dimensional model to extracted features

Fig. 1. *Overview of deep feature learning approach to modeling brain architecture.* Our pipeline consists of three main steps: 1. Training a deep neural network, 2. Extracting features from the trained DNN and 3. Fitting a (low dimensional) model on the extracted features.

and found that our approach could reveal structure in brain architecture beyond what the network was trained to find. This opens up many possibilities for finding new subdivisions in the brain and providing automated ways to extract more fine-scale features from brain structure.

II. Related Work

Feature extraction using deep neural networks has been studied extensively by the machine learning community in the past. Works such as [9]–[11] were some of the first to explore the use of deep networks as feature extractors. However, they were restricted to the unsupervised setting and did not leverage any class labels when training networks for feature extraction.

More recently, works such as [12], [13] approached deep feature extraction from a supervised standpoint and utilized class labels when training the network. However, these works mainly looked at these features from the lens of transfer learning and generalizability. Zeiler and Fergus [14] showed that features extracted from a convolutional neural network trained on the Imagenet dataset significantly outperformed hand-crafted ones when combined with a classifier such as a support vector machine (SVM) on a datasets such as Caltech-101, 256 and PASCAL-VOC 2012. Razavian et al. [15] furthered this line of work and performed an extensive study on the use of CNN features off the shelf using the OverFeat feature extractor [16] and were perhaps the first to show that astonishingly, classification pipelines consisting of deep extracted features appended with an SVM could consistently outperform trained state-of-the-art systems in a variety of visual classification tasks on various datasets.

Since then, works in different domains such as speech [17] and botany [18] have utilized pre-trained networks as feature extractors. These ideas have found application in neuroanatomy as well, with Chen et al. [19] using features extracted from a pre-trained CNN appended with a linear classifier to discriminate between different brain areas.

In our work, we train a CNN *on the task* we care about, rather than using a pre-trained network as feature extractor and fit a generic classifier on our extracted features as our final model for the different brain areas. We find that this approach outperforms (albeit slightly) both, i) A CNN trained end-to-end on the task, and ii) Classification pipelines with features extracted from pre-trained networks and an appended classifier. We also further explore the use of the extracted deep features and the information encoded in them to sub-divide and discover fine-grained structure within brain areas.

III. Methods, Experiments and Results

1) Details of multi-area brain dataset

To test our idea of using deep learning-based feature extraction for modeling brain architecture, we obtained a large ($\sim 5\text{mm}^3$, 5.9x10^9voxels) thalamocortical slice from a mouse brain that spans both cortical (somatosensory) and deep-brain areas (thalamus, striatum) and acquired three-dimensional image volumes with X-ray microtomography as described in [20]. This dataset provided sufficient resolution for our analysis, revealing diverse structures (e.g., myelinated axons, cells, and blood vessels) that provided details necessary for a trained anatomist to divide the sample into distinct brain areas.

This 3D image volume was then divided into regions-of-interest using multiple 2D sections that spanned five different brain areas, as well as the internal capsule (white matter) (Figure 2A). After annotating the 2D slices, we then extracted 128 x 128 patches (pixel size is 1.17 microns) from the different brain areas: somatosensory cortex (Ctx), hypothalamus (HypoTh), striatum (Str), the ventral posterior (VP) region of the thalamus, and zona incerta (ZI), as well as white matter (WM) (Fig. 1). We created a training, validation, and test set by selecting ∼2000 images/class for the train and 1000 images/class for the validation and test sets. Samples from the three datasets, are separated by 50 microns to ensure that the slices would be sufficiently different to avoid model

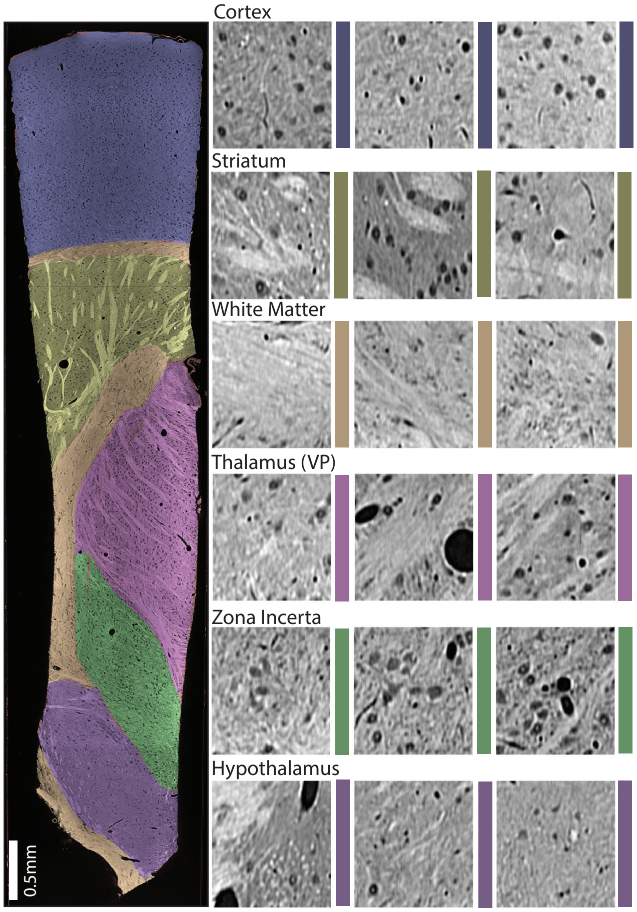

Fig. 2. On the left, we display an example of an annotated slice of the thalamocortical volume. Different colours signify different brain areas. On the right, we show examples of 128x128 image patches sampled from the different brain areas.

overfitting. The images in the training set were sampled such that they belonged strictly to the interior of a class and did not include portions of the boundary, or other classes. However, we made no such restrictions on the data sampled to be part of the validation and test sets. Our data curation strategy ensured that our training, validation and test sets all spanned the different classes uniformly, represented the heterogeneity in the classes adequately, and were sufficiently different from each other so as to ensure a certain degree of generalizationa and avoid overly optimistic results.

2) Training a CNN to discriminate across brain areas

We trained a feed forward convolutional neural network to discriminate between the six different classes in our dataset. We used the ReLU activation function throughout and our network had 7 layers - the first four being convolutional layers with kernels of size 7x7, 5x5, 3x3, 3x3 and 16, 32, 64, 128 filters respectively, followed by three fully connected layers with 1024, 64, 6 nodes respectively. The loss function used is simple cross entropy and the optimizer used was Adam with a learning rate of 1e-4.

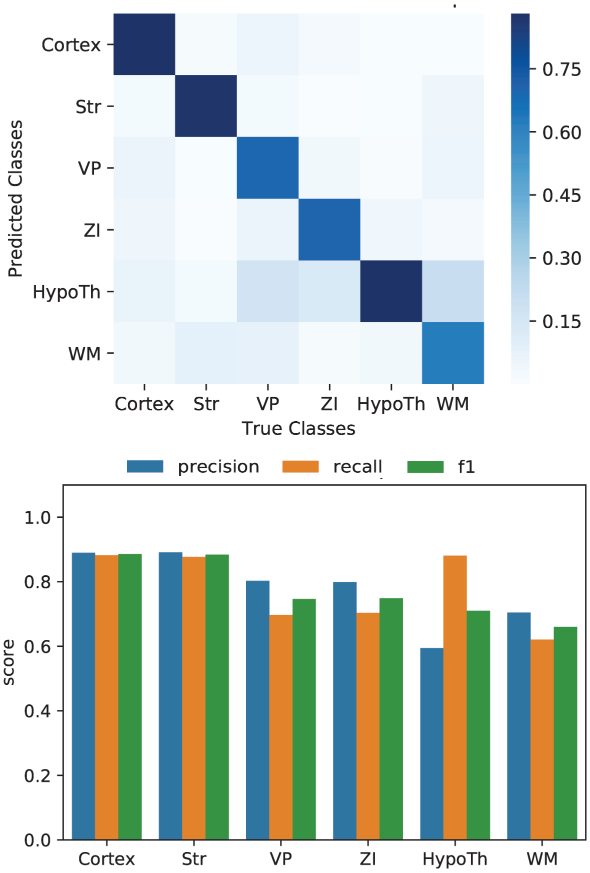

Fig. 3. *Results of brain area prediction at scale using a convolutional neural network architecture.* Top: Confusion matrix for the performance of the CNN. Bottom: The precision, recall, and f1-scores for the network split across different brain areas.

Our best performing model was chosen on the basis of its performance on the validation set, and achieved an accuracy of 88.88% on the test set. To confirm our model's ability to perform well on large scale data, we tested the model on an entire slice of the thalamocortical volume which had more than five million 128x128 image patches in total. Even at this scale, we obtained consistent, accurate results across a range of different brain areas (Figure 3).

The overall accuracy for the trained CNN at scale was 80%, demonstrating that the features learned by the network are sufficient to resolve differences across a diverse set of areas, even when they exhibit significant overlap. Moreover as seen from the confusion matrix in Fig. 3, the trained CNN can easily distinguish between classes that are significantly different from each other in distribution but is slightly confused between areas that share boundaries and have similar structural compositions. Our results therefore suggest that CNNs can be used to learn features that reliably separate the brain areas of interest in our study and that the mistakes made by the network would be similar to those a human would make.

3) *Feature extraction from deep neural networks*

After obtaining a network that can reliably discriminate between different brain areas, we examined the activations at the network's last hidden layer (Fig. 1). Representations formed at this layer provide rich and concise abstractions of the high dimensional inputs fed to the network and allow us to ask questions about the different latent factors of the data. To study the population level responses of the network, we collected representations for all our samples (across all six classes) and arranged them in a matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$ where N is the total number of samples and D is the dimensionality of the extracted representations (i.e. number of units in a layer). Studying joint network activations across a multitude of inputs allows us to develop an understanding of the global characteristics of the trained CNN, revealing insights into the behaviour of the network across a variety of test inputs.

4) *Testing the extracted representations in transfer learning*

Furthermore, the activations at the last hidden layer of a deep neural network have been shown to be useful as off-the-shelf features for generic computer vision tasks [14]–[16], and the same approach has been applied to medical imaging tasks as well [19]. We therefore chose to examine the suitability of these last-layer representations from our trained network as features for a generic classifier (e.g. a simple logistic regression classifier or an SVM) in a simple brain area discrimination task, and how they fared in comparison to those extracted from (ImageNet) pre-trained networks. Our analysis (Fig. 4) revealed that with appropriate supervision, the features extracted from InceptionNet and other pre-trained networks were capable of obtaining high accuracy in predicting the intended classes. It should be noted that the accuracies of all simple supervised classifiers trained on the extracted features from all networks was quite comparable to the final accuracy of the end-to-end trained CNN. Impressively, in some cases, the accuracy of the simple classifiers was even slightly *higher* than that of the end-to-end trained CNN. These findings are in line with previous results in both machine learning research and medical imaging and brain histology problems similar to our application, where pre-trained nets have been used successfully as fixed-feature extractors.

However, the features extracted from fixed pre-trained networks failed to perform well on a simple unsupervised clustering task, thus showing that they inherently contain a limited amount of relevant information about the structure of brain areas. Here we fit a gaussian mixture model (GMM) on the training data (i.e. extracted features for samples in the training set) and predicted the clusters of the test data (extracted features for samples in the test set) using the GMM. We found that extracted features from all the pre-trained networks had very low V-Measure scores and were significantly outperformed by the features extracted from the trained CNN. Furthermore, in our second set of experiments where we added some noise to the images, we found that the features extracted from the pre-trained networks failed to enable separation of the different classes, while the features extracted from the CNN were far more resilient to even moderate levels of noise and showed a less drastic degradation in performance (Fig. 4). These experiments speak to the lack of separability in the extracted features of the fixed pre-trained networks and make a strong case for those extracted from a deep network that is *trained* on the task.

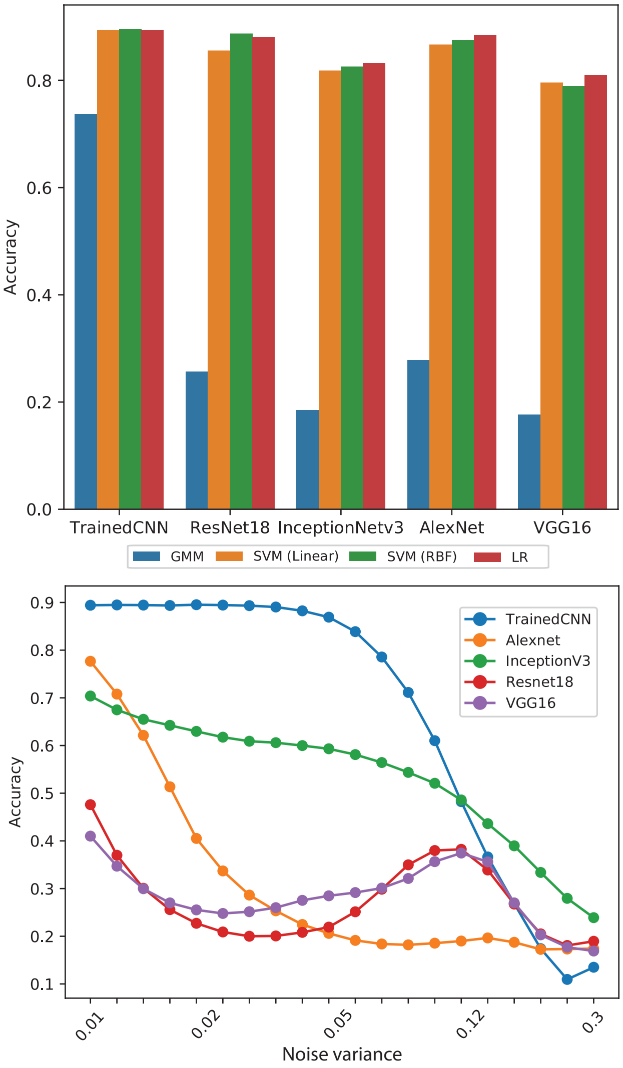

Fig. 4. *Performance of different classifiers and clustering methods when trained using features extracted from a trained CNN and different* pre-trained *convolutional neural networks.* On top, the performance on the test set for supervised classifiers (SVM-Linear, SVM-RBF, linear regression (LR)) and for an unsupervised clustering method (GMM). The score for the GMM is the V-Measure since accuracy cannot be generally defined for clustering techniques. On the bottom, we show the performance of the LR classifier when presented with noisy data.

5) *Low-dimensional structure and modularity in network activations*

Consequently, while pre-trained networks appear to provide good performance as feature extractors in the presence of little to no noise on supervised tasks, our experiments in the unsupervised and noisy settings led us to hypothesize that there would be major differences in the structure of features in a pre-trained net versus a trained network. Thus, we computed the Gram matrix of the normalized extracted features given as $\mathbf{G} = \mathbf{X}\mathbf{X}^T$ (Figure 5). The ideal structure of such a matrix would be block diagonal, and therefore the structure of the matrix $\mathbf{G}$ indicates how characteristic the features are of a particular class as well as how suitable they are for a discriminative task. We quantified how far the covariance matrices for the representations using the trained CNN and two other pre-trained networks were from their ideal block diagonal structure $\mathbf{D}$ by computing $\frac{1}{N}||\mathbf{G} - \mathbf{D}||_F$ where $||.||_F$ is the Frobenius norm. As expected, we found that the covariance of the features of the trained CNN was the closest to being ideally block diagonal with an approximation error of 0.28. On the other hand the pre-trained networks had much higher approximation errors of 0.56 (Inception) and 0.72 (Resnet18), thus confirming our hypothesis that representations from the trained CNN and pre-trained networks show significant structural differences.

From an anatomical perspective, the similarities and differences between brain areas that we hoped to see were also present in the structure of the features' covariance matrix. For instance, cortex and striatum, as well as VP and ZI, which are examples of brain area pairs that are neighboring and have similarities in their structural distributions were highly correlated with one another, while those areas such as Cortex and VP are more dissimilar. Finally, the properties discussed previously also held true for the low-dimensional representations of the features extracted from the trained CNN, as seen in Figure 5 where we observed that neighboring and anatomically similar areas showed up close together with a certain amount of overlap while significantly different areas were very clearly separated. However, neither the covariance matrices nor the low-dimensional representations showed much structure for the pre-trained networks.

Our analyses of the features extracted from different networks therefore confirmed two facts: i) Features formed in a trained neural network are better structured and anatomically accurate than those from pre-trained networks, and ii) The features from the former can capture the relationships amongst different areas, are highly informative of brain structure and can be used to effectively model the continuous intra- and inter-area variability in brain architecture.

6) Further sub-dividing regions of interest

Taking this one step further, we examined whether features learned in a trained CNN could be used to further divide a brain area. We tested this idea on a slice in cortex, given that we know that it consists of six different layers that can be distinguished by cell density [21]. After densely sampling image patches in the region of interest and organizing their features along the cortical depth as seen in Fig. 6, we reduced the dimensionality of our extracted features using principal component analysis (PCA) to both, denoise the features as well as orient them in a way that would presumably align with key latent factors of the data. We then applied K-Means clustering to the low-dimensional features (k=4 clusters). We found that we could quite clearly identify regions near layers 1 and 2/3, as well as transition zones around them (Fig. 6). These preliminary results of the sub-division experiment show that areas of varying cell density can be determined through combining deep extracted features and unsupervised learning. This method thus shows promise in being able to reveal structure beyond what the network has been trained to see and classify in the images its trained on.

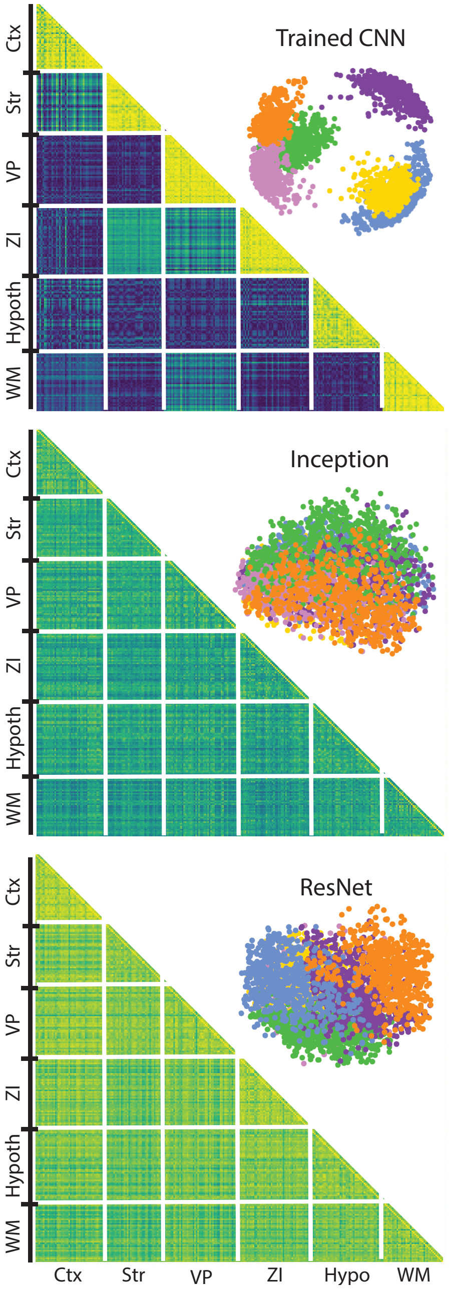

Fig. 5. *Visualization of low-dimensional and modular structure of network activations from trained CNNs.* From top to bottom: The covariance matrices and 3D PCA representations of the features extracted from the trained CNN (top), InceptionNet (middle), and ResNet (bottom).

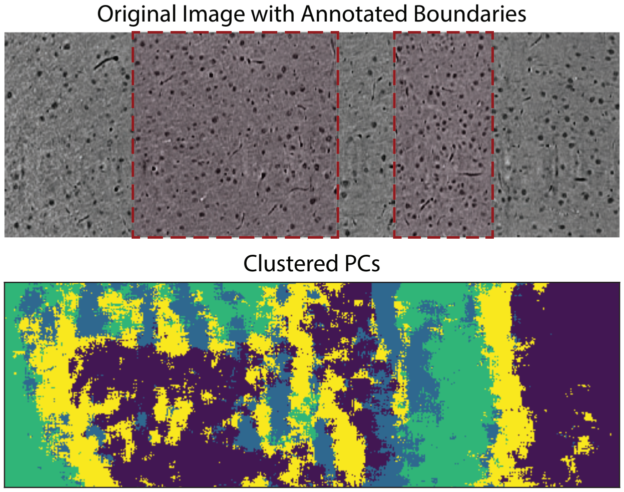

Fig. 6. *Using deep features to find regions of low and high cell density in cortex.* In the image on top, we show an X-ray slice which spans five out of six cortical layers, starting with white matter (Layer 6) on the left (not in picture) and progressing to Layer 1 (top of cortex) as we move to the right. When applying k-means to cluster the extracted features (bottom), our method divides the sample into regions with low cell density, higher cell density, and transition zones. In this example, two layers can be extracted, near Layer 2/3 (left) and Layer 1 (right).

IV. Conclusion

In this paper, we developed an approach that uses deep feature extraction to form a model of brain architecture. We applied our method to a X-ray micro-tomography dataset and showed that our method provides meaningful representations of different brain areas. We also showed that features learned in deep networks can be used to reveal subdivisions in tissue in a biologically plausible manner. This study opens up the possibility of using deep learning to extract features from neural datasets that can be used to further reveal fine-scale organization of brain structure without needing to specify these labels a priori.

Acknowledgments

We would like to acknowledge Judy Prasad and Narayanan (Bobby) Kasthuri at the University of Chicago and Vincent deAndrade from Argonne National Laboratory for their help in acquiring and generating the X-ray microCT dataset used in this study.

References

[1] K. Milligan, A. Balwani, and E. Dyer, “Brain mapping at high resolutions: Challenges and opportunities,” *Current Opinion in Biomedical Engineering*, 2019.

[2] A. Acebes, “Brain mapping and synapse quantification in vivo: It’s time to imaging,” *Frontiers in neuroanatomy*, vol. 11, p. 17, 2017.

[3] D. W. Shattuck, M. Mirza, V. Adisetiyo, C. Hojatkashani, G. Salamon, K. L. Narr, R. A. Poldrack, R. M. Bilder, and A. W. Toga, “Construction of a 3d probabilistic atlas of human cortical structures,” *Neuroimage*, vol. 39, no. 3, pp. 1064–1080, 2008.

[4] A. Iqbal, R. Khan, and T. Karayannis, “Developing a brain atlas through deep learning,” *Nature Machine Intelligence*, vol. 1, no. 6, p. 277, 2019.

[5] J. A. Harris, S. Mihalas, K. E. Hirokawa, J. D. Whitesell, J. Knox, A. Bernard, P. Bohn, S. Caldejon, L. Casal, A. Cho, *et al.*, “The organization of intracortical connections by layer and cell class in the mouse brain,” *BioRxiv*, p. 292961, 2018.

[6] A. G. Huth, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant, “Pragmatic: A probabilistic and generative model of areas tiling the cortex,” *arXiv preprint arXiv:1504.03622*, 2015.

[7] P. Thompson, M. S. Mega, and A. W. Toga, “Disease-specific probabilistic brain atlases,” in *Proceedings IEEE Workshop on Mathematical Methods in Biomedical Image Analysis. MMBIA-2000 (Cat. No. PR00737)*, pp. 227–234, IEEE, 2000.

[8] D. Rolnick and E. L. Dyer, “Generative models and abstractions for large-scale neuroanatomy datasets,” *Current opinion in neurobiology*, vol. 55, pp. 112–120, 2019.

[9] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, “Self-taught learning: transfer learning from unlabeled data,” in *Proceedings of the 24th international conference on Machine learning*, pp. 759–766, ACM, 2007.

[10] R. Salakhutdinov and G. Hinton, “Deep boltzmann machines,” in *Artificial intelligence and statistics*, pp. 448–455, 2009.

[11] G. Mesnil, Y. Dauphin, X. Glorot, S. Rifai, Y. Bengio, I. Goodfellow, E. Lavoie, X. Muller, G. Desjardins, D. Warde-Farley, *et al.*, “Unsupervised and transfer learning challenge: a deep learning approach,” in *Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning workshop-Volume 27*, pp. 97–111, JMLR. org, 2011.

[12] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, “Decaf: A deep convolutional activation feature for generic visual recognition,” in *International conference on machine learning*, pp. 647–655, 2014.

[13] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Learning and transferring mid-level image representations using convolutional neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1717–1724, 2014.

[14] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *European conference on computer vision*, pp. 818–833, Springer, 2014.

[15] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “Cnn features off-the-shelf: An astounding baseline for recognition: Ieee conference on computer vision and pattern recognition workshops,” 2014.

[16] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, “Overfeat: Integrated recognition, localization and detection using convolutional networks,” *arXiv preprint arXiv:1312.6229*, 2013.

[17] E. Tatulli and T. Hueber, “Feature extraction using multimodal convolutional neural networks for visual speech recognition,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2971–2975, IEEE, 2017.

[18] S. H. Lee, C. S. Chan, S. J. Mayo, and P. Remagnino, “How deep learning extracts and learns leaf features for plant classification,” *Pattern Recognition*, vol. 71, pp. 1–13, 2017.

[19] Y. Chen, L. E. McElvain, A. S. Tolpygo, D. Ferrante, B. Friedman, P. P. Mitra, H. J. Karten, Y. Freund, and D. Kleinfeld, “An active texture-based digital atlas enables automated mapping of structures and markers across brains,” *Nature methods*, vol. 16, no. 4, p. 341, 2019.

[20] E. L. Dyer, W. G. Roncal, J. A. Prasad, H. L. Fernandes, D. Gürsoy, V. De Andrade, K. Fezzaa, X. Xiao, J. T. Vogelstein, C. Jacobsen, *et al.*, “Quantifying mesoscale neuroanatomy using x-ray microtomography,” *eneuro*, vol. 4, no. 5, 2017.

[21] T. G. Belgard, A. C. Marques, P. L. Oliver, H. O. Abaan, T. M. Sirey, A. Hoerder-Suabedissen, F. García-Moreno, Z. Molnár, E. H. Margulies, and C. P. Ponting, “A transcriptomic atlas of mouse neocortical layers,” *Neuron*, vol. 71, no. 4, pp. 605–616, 2011.