Database Schema for Human Brain Scale Neural Networks

Richard Wang, Simon Alford, Lauren Milechin, Ryan Robinett, and Jeremy Kepner
Massachusetts Institute of Technology

***Abstract*—Deep neural networks have shown great potential in solving various problems through machine learning. With recent high levels of accuracy, there is a new possibility to model and understand the human brain by running a neural network comparable to the number of neurons and connections in the human brain. In many cases, memory limits neural network size. Fortunately, the human brain's number of neurons and sparse connections implies that we can model the human brain as a large, sparse neural network. With the Dynamic Distributed Dimensional Data Model (D4M) software and the Apache Accumulo database, it is possible to store a neural network, with as many neurons and connections as the human brain, through the help of parallel computing in supercomputers. This work describes a schema for storing a human brain scale neural network. Using Graphulo, the stored neural network will be able to execute large inference on the same scale as the human brain.**

***Index Terms*—sparse neural networks, human brain, Accumulo database**

I. Introduction

Neural networks have become larger and sparser to maximize the accuracy of machine learning [1]. Major companies have managed to create and run large neural networks that are reasonably accurate [2]. Still, the largest existing neural networks have far fewer connections than the human brain with its estimated 86 billion neurons and over 100 trillion connections [3]–[5]. Constructing a neural network at this scale requires massive parallel databases and schemas. The Apache Accumulo database is designed to hold large, sparse data and is a reasonable starting point to explore building a human brain scale neural network [6].

With about 100 billion neurons and 100 trillion connections, the human brain has inherent sparsity. Sparse matrices are the result of taking regular neural network weight matrices and setting weight values close to 0 to 0. This process of approximating the weight value is equivalent to severing the connection between the two neurons that the weight values connected. With sparse matrices, matrices of larger dimensions can be stored using less memory compared to denser matrices of the same dimensions and can take advantage of sparse matrix optimized math libraries such as the GraphBLAS [7], [8].

This material is based in part upon work supported by the NSF under grant number DMS-1312831 and CCF-1533644. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

In general, the resulting sparse matrix does not guarantee that every input node has a path to every output node. Recently, there have been studies in generating sparse weight matrices that have path connectedness, a property with which every input neuron has a path connecting to every output neuron [1]. Through the use of Radix-Net Topologies, a neural network composed of sparse weight matrices can be carefully monitored through the weights' path-connectedness [9].

With Graphulo, matrix multiplication, the operation that allows for neural network inference, can be executed inside the Accumulo database. Using D4M, matrix multiplication is performed through the product of associative arrays [10]. As a result, associative arrays provide an effective way to run the forward propagation of weight matrices. However, in order to utilize this method of forward propagation, weight matrices need to be labeled properly. Although the brain has been mapped through various schemas, there does not exist a schema that has enough brain labels for all the neurons in the brain. Since this potential model of the human brain uses associative array multiplication, a naming scheme for more brain regions is necessary for more efficient matrix multiplication within the Accumulo database.

To fully insert all of the 100 billion neurons and 100 trillion connections into a database, parallel computing and vectorization are also necessary to reduce the amount of time it takes to ingest weights.

II. Approach

To partition the individual weight matrices of the human brain into reasonable sizes, further separations are necessary. The 100 trillion (10^{14}) connections could be reached by putting approximately 300 billion ($3 * 10^{11}$) connections into 300 instances of Accumulo. Within each instance of Accumulo, the 300 billion connections could be achieved through 300 weight matrices. These would be 10^6 by 10^6 weight matrices with 1000 connections for each neuron. This means that each weight matrix would have sparsity 10^{-3}. Using the Radix-Net Topology, more specifically extended from the Extended Mixed-Radix (EMR) Topology, two of these weight matrices are constructed with $\mathcal{N}^* = (1000, 1000)$ [9].

In order to achieve the desired number of brain regions, existing naming schemes were combined. For the cerebrum, there exist many labels in the set of accessible brain images [11]. With the usage of the different layers (types of neurons) of the cerebrum, the number of brain regions is multiplied even

more [12]. As for the cerebellum, there exist many different naming schemes, such as the anatomical and functional naming schemes [13]. Like the cerebrum, even more regions are generated through the cortical layers of the cerebellum [13]. While the cerebrum has often been regarded as the portion of the brain that deals with higher-level thinking, the cerebellum, which is smaller in size than the cerebrum, has more than twice as many neurons as the cerebrum [4]. As a result, the idea of five different microzones, as seen in cats, allows us to generate even more regions for the cerebellum and maintain the ratio of cerebrum labels to cerebellum labels [13], [14].

Thus, to obtain a useful naming schema and ontology for the brain, the following brain regions were used and combined:

TABLE I
ONTOLOGY FOR DIVIDING UP THE HUMAN BRAIN INTO 1272 REGIONS

Hemisphere	Functional	Anatomical	Layer	Microzone
L R		S_fnt Rm_fnt Cm_fnt P_oprc Tran P_orb L_orb M_orb Precn Paracn Nthl Parhpo Fsfrm S_tmpl M_tmpl I_tmpl T_tmpl Postcn Spmrgl Insla S_prtl I_prtl Precu Lngl Pclcrn Cu L Ra Ca P Ls	Pxfm Exgn Prmd Ingn Gngl Mtfm	
L R	Vrms Spno Cbro	l1 l2 l3 l4 l5 l6 l7Af l7At l7B l8A l8B l9	mlclr prknj grnlr	mz_1 mz_2 mz_3 mz_4 mz_5

Once the brain regions are all created, a list of all the brain regions allows the weight matrices to be properly labeled. Proper labels, using consecutive elements of the brain region list, are integral to associative array multiplication. Vectorization allows us to reduce the number of for-loops and, as a result, decreases the amount of time that it takes to generate the associative array labels (row ids, column ids, and values).

Ingesting large data efficiently requires finding the optimal number of parallel processes. We tested different numbers of workers, which all created, labeled, and inserted the optimal size weight matrices. To test that, we tried 1, 2, 4, and 8 workers and graphed the ingest rate. In doing so, we can extrapolate the total time needed to ingest all of the brain regions.

III. RESULTS

By combining the existing labels in brains with the different layer labels, we were able to obtain 1272 brain region names. Since these region names were obtained through the cross sections of existing regions, each existing label was shortened to reduce the size of the labels. For all of the brain regions, they are separated by the left and right hemispheres. A forward slash between the regions separates the various cross sections for that brain region.

The 372 cerebrum region names were obtained using the vast existing names in both hemispheres with the 6 layers in the cerebrum. For the cerebrum, the brain regions are in the form {hemisphere}/{anatomical region}/{layer}, using the components listed in the top half of table 1 [11].

The 900 cerebellum region names were obtained through cross sections of the hemispheres, functional names, anatomical names, layers, and five suspected microzones. For the cerebellum, the region names are in the form {hemisphere}/{functional region}/{anatomical region}/{layer}/{microzone}, using the components listed in the bottom half of table 1 [15].

By varying the number of workers, the ingest rate nearly doubled from 2 to 4 workers. Changing from 4 to 8 workers, the ingest rate increased by approximately 20%. Also, transposing the weight matrix proved to be necessary in order to convert the sparse matrix formatting between compressed sparse columns (CSC), which Julia uses, and compressed sparse rows (CSR), which is used in Accumulo.

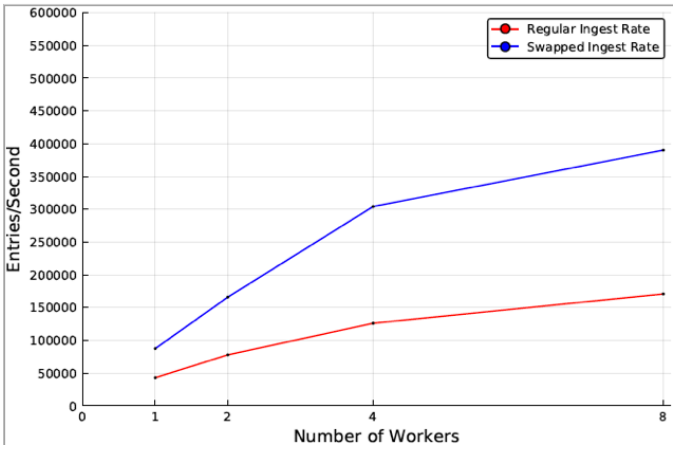

Fig. 1. Insert rate versus number of inserting processes

Lastly, the weight matrices were ingested with their brain labels. As the number of workers increases, the speed of ingesting increases due to parallel computing. In addition, by taking the transpose (swapped) of the matrices, the matrices are ingested in the correct format at faster rates.

IV. SUMMARY

By taking the cross sections of existing brain regions, we can easily generate over a thousand unique brain region labels. With these labels, we can make the row ids and column ids for associative arrays so that large inference can be done within the Accumulo database. Consequently, we have found a way to create a 100 trillion connection sparse neural network with 9×10^4 path-connected weight matrices.

ACKNOWLEDGMENT

The authors wish to acknowledge the following individuals for their contributions and support: Megan Blackwell, Alan Edelman, Vijay Gadepally, Chris Hill, Hayden Jananthan, Charles Leiserson, Sid Samsi, and the MIT SuperCloud team.

REFERENCES

[1] A. Prabhu, G. Varma, and A. Namboodiri, "Deep expander networks: Efficient deep networks from graph theory," *arXiv preprint arXiv:1711.08757*, 2017.

[2] J. Hsu, "Biggest neural network ever pushes ai deep learning," Jul 2015. [Online]. Available: https://spectrum.ieee.org/tech-talk/computing/software/biggest-neural-network-ever-pushes-ai-deep-learning

[3] F. A. Azevedo, L. R. Carvalho, L. T. Grinberg, J. M. Farfel, R. E. Ferretti, R. E. Leite, W. Jacob, R. Lent, and S. Herculano-Houzel, "Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain." *PubMed Central*, Apr 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19226510

[4] S. Herculano-Houzel, "The human brain in numbers: A linearly scaled-up primate brain," *TPubMed Central*, Nov 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2776484/

[5] B. Voytek, "Are there really as many neurons in the human brain as stars in the milky way?" May 2013. [Online]. Available: https://www.nature.com/scitable/blog/brain-metrics/are_there_really_as_many

[6] J. Kepner, W. Arcand, D. Bestor, B. Bergeron, C. Byun, V. Gadepally, M. Hubbell, P. Michaleas, J. Mullen, A. Prout *et al.*, "Achieving 100,000,000 database inserts per second using accumulo and d4m," in *High Performance Extreme Computing Conference (HPEC), 2014 IEEE*. IEEE, 2014, pp. 1–6.

[7] J. Kepner and J. Gilbert, *Graph algorithms in the language of linear algebra*. SIAM, 2011.

[8] J. Kepner, M. Kumar, J. Moreira, P. Pattnaik, M. Serrano, and H. Tufo, "Enabling massive deep neural networks with the graphblas," in *High Performance Extreme Computing Conference (HPEC), 2017 IEEE*. IEEE, 2017, pp. 1–10.

[9] R. A. Robinett and J. Kepner, "Neural network topologies for sparse training," *arXiv preprint arXiv:1809.05242*, 2018.

[10] J. Kepner and H. Jananthan, *Mathematics of big data: Spreadsheets, databases, matrices, and graphs*. MIT Press, 2018.

[11] A. Klein and J. Tourville, "101 labeled brain images and a consistent human cortical labeling protocol," *Frontiers in Human Neuroscience*, 2012.

[12] B. Pansky, *Review of medical embryology*. Macmillan New York, 1982.

[13] R. Apps and R. Hawkes, "Cerebellar cortical organization: a one-map hypothesis," *Nature Reviews Neuroscience volume 10*, 2009.

[14] G. Andersson and O. Oscarsson, "Climbing fiber microzones in cerebellar vermis and their projection to different groups of cells in the lateral vestibular nucleus," *PMID: 689129*, 1978.

[15] A. D'Mello and C. Stoodley, "Cerebro-cerebellar circuits in autism spectrum disorder," *Frontiers in Neuroscience*, vol. 9, 11 2015.