Spatially Sparse Mixture of Experts

Daniel Byrne[1] and John Santerre[1]

Master of Science in Data Science, Southern Methodist University, Dallas TX 75275
USA {byrned,santeerej}@smu.edu

Abstract. The difficulty in training and using deep neural networks begins and ends with the fact that the design principle on which many are founded, fullly connected layers (FFL), contributes to a quadratic increase in the number of calculable parameters as the number of input dimensions and parameters increases. Traditionally, researchers have simply increased the number of parameters to increase the capacity and learning ability of the neural network. However, research into sparsely connected networks and pruning suggests that this need not be the case. Furthermore, research into small-world networks and biological systems suggests that spatially aware local expert clusters can combine in situationally uniquie ways with distant expert clusters through sparse long distance connections to form conditionally active subnetworks adept at solving complex problems in a computationally efficient way. This paper introduces a new neural network architecture, Parametric t-SNE gated Mixture of Sparse Experts, that combines these insights into a cohesive model. The model uses a 2D Parametric t-SNE gating network to drive the gate inputs of a mixture of sparsely interconnected expert clusters. These clusters in contrast to traditional Mixture of Experts networks, are not mutually exclusive and independant, but share a low fanout collection of interexpert routes. These expert clusters themselves are additionaly tuned locally to limit the number of interferring intracluster connections. This model achieves a compression ratio of ? percent over traditional FCL neural networks and a ? improvement over SGMoE neural networks while achieving the similar performance metrics. The network should also be faster to train and should give results faster during intpretatation due to the decreased computational budget.

Keywords: DNN· t-SNE· Mixture of Experts · pruning · sparse · brain modeling · small-world

1 Introduction

Deep learning's explosive growth in learning over the past few years has been fueled by the technique's success at embedding high dimensional data into multilevel encoded representations of increasing abstraction to solve complex functions and or to model natural processes such as Natural Language Processing (NLP) and Computer Vision. In common feed forward Deep Neural Networks, layers are typically fully connected. That is every node at any one layer is connected to every other node in adjacent layers. This fully connected nature supplies each successive layer with any and all embeddings of its preceding layer in any possible combination. It is then the task of the training procedure to effectively reduce this over abundance of overlapping connections by iteratively applying the delta rule to adjust the weights on the connections. The means tend to cluster around zero indicating that a large number of connections are redundant or destructively interfering causing the training procedure to zero them out. The net result of training then is a collection of weights and neurons that when applied an input stimulus effectively creates an alternative representation or classification of the input data at the network's output.

However, training such over specified models is time consuming compute resource intensive. The difficulty in training deep neural networks begins and ends with the fact that the design principle on which many are founded, full connectivity, contributes to a quadratic increase in calculable parameters as the number of input dimensions and parameters increases. In a typical DNN, there may be millions or billions of parameters and similarly large set of labelled training examples.

In a multi-class neural neural network, competing classifications with similar and dissimilar features will inject noise relative to each competing target classification. Neural networks then have to be trained to discriminate the important overlapping features while at the same time filtering out the less important noise to iterate towards a adequately minimized loss functions which exhibits the ability to discriminate patterns from different classes.

For example the theoretical ideal set of weights that consistently identifies cats in images is most likely in conflict with set of all possible weights of connections tuned to correctly classify trees. There will be some overlap, and that overlapping region is nonlinear objective function the neural network is trying to optimize. However, arriving at one of the many possible number of near optimal solutions to this equation is not trivial.

There have been a number of solutions to address this problem in the literature. The techniques vary, but the end results are that interfering patterns are filtered, and unique signals are amplified.

Dropout is a method by which a tunable percentave of random hidden units and their connections are eliminated from the network with each training case. During test all available neurons are included, but the trained weights on their connections are multiplied by the dropout probability p. Dropout gives major improvements in training time and accuracy by encouraging each individual neurons to learn a feature by degrading or filtering the impact other hidden units have on its connection weights.

DropConnect is similar to Dropout in that it introduces dynamic sparsity within the model, but instead of dropping neurons and their connections, it drops only connections based again on a settable probability parameter.

Embedding auto-encoders have been successful in image noise reduction and modeling the time series relationships of stock trading data. Auto encoders work by compressing input data of n dimensions into a smaller p dimensional hidden layer which is then re-encoded into an n dimensional output layer through gradient descent training. The effect of this compression and decompresssion technique is to filter out nonessential, noisey, data points.

In Convolutional Neural Networks, pooling layers are often utilized to reduce the size of the feature maps coming out of a filter bank before feeding the resultant filtered feature map to successive layers. This has the effect of reducing the number of interfering connections, reducing the computational complexity, and filtering noisy features.

LeCun et al. devised a method dubbed Optimal Brain Damage in which they used second order derivitives to identify and prune unimportant weights from a network which they showed improved the network's generalization.

While Dropout and DropConnect reduce the number of parameters during training, but reinstate them during testing, a true reduction in the network connections has not been achieved. Song Han et al. chose to learn both the weights on connections and the existence of connections during training. Their method first trains a dense network, prunes the low weight connections, and then retrains the network to fine tune the remaining connections.

Deep Compression has achieved compression rations of 98 percent with no loss of accuracy following a similar process, but in the final step they used Huffman encoding to further reduce the size of the network.

Inspired by research on Rat Brain Tissue which found that synaptic formation for specific functional brain regions can be modeled as a random formation, StochasticNets modified this above approach by starting with a sparse randomly connected network and ending with a sparse network. Pruning is thus inherent in their design.

Decebal Constantin Mocanu et al. again taking inspiration from the structure of biological networks argue that DNNs should never have fully connected layers and subsequently demonstrate the effectiveness of sparse representations on several diverse neural network architectures (RBMs, MLPs, and CNNs).

N. Alex Cayco-Gajic in their study on the pattern seperation abilities of the cerebellum determined that feedforward networks such as those in the brain and in neural networks can separate patterns by projecting them onto a larger population of neurons and spatially encoding the inputs to decorrelate their influences.

Research on a broad cross-section of network architectures has determined this pattern of densely connected clusters with sparse connections to remote clusters occurs organically in biological, social, and technological systems thus reinforcing the notion that this type of architecture is an optimal solution in a number of diverse paradigms.

Further reinforcing this concept that regional specialization is important in neural network architectures, PathNet uses agents to identify low cost paths through the network that minimize the cost of a user specific targets solution in the global solution space. Agents must learn to work with other agents, sharing parameters or zeroing out interfering paths as necessary.

Thus that leads to the intuition that perhaps the networks we should be trying to develop are composite networks comprised of a mixture of local dense expert clusters weakly connected to distant, non-overlapping, clusters.

Such a technique has been implemented by Hinton et al. in their Sparsely Gated Mixture of Experts model. This model uses a separate gating network which permits access to segregated subnetworks trained to be experts on subdomains of the global target domain. However, their approach did not spatially segregate the training inputs leading to overuse of certain experts thus increasing their correlations and thus increasing the noise.

This paper introduces a new neural network architecture, Parametric t-SNE gated Mixture of Sparse Experts, that combines insights gleaned from these prior approaches into a cohesive model.

In this paper we propose a modification of this gated MOE model in which the gating network enforces a spatial seperation between expert clusters using Parametric t-SNE. Parametric t-Stochastic Neighbor Embedding, Pt-SNE, is a method by which the Mutual Information component of similar Gaussian distributions are converted into Euclidian distances to convert a high dimensional signal into a lower dimensional space while preserving the original structure of the data.

Also, in contrast to traditional Mixture of Experts architectures which isolates adjacent clusters, we propose adding a stochasitcally sparse number of connections between expert clusters. This information sharing between clusters should reduce the greedy nature of training algorithms tendency to overuse some clusters while minimally using others.

Clusters are also pruned or sparsely initialized using methods devised in StochasticNet and/or Brain Damage.

The result, we hope, is a spatially aware, sparsely connected and initiated, mixture of local experts with distant connections allowing for information sharing from far flung expert networks through minimal influence connections that should be less dense and less computationally complex than a fully connected comparative network. [2, 12, 20, 11, 7, 13, 25, 17, 16, 23, 5, 9, 21, 24, 8, 14, 22, 18, 10, 3, 19, 4, 6, 1, 15]

References

1. Bassett, D.S., Bullmore, E.: Small-world brain networks. The Neuroscientist **12**(6), 512–523 (2019/10/07 2006). https://doi.org/10.1177/1073858406293182, https://doi.org/10.1177/1073858406293182
2. Caswell, I., Shen, C., Wang, L.: Loopy neural nets: Imitating feedback loops in the human brain. Tech. Report (2016)
3. Cayco-Gajic, N.A., Clopath, C., Silver, R.A.: Sparse synaptic connectivity is required for decorrelation and pattern separation in feedforward networks. Nature Communications **8**(1), 1116 (2017). https://doi.org/10.1038/s41467-017-01109-y, https://doi.org/10.1038/s41467-017-01109-y
4. Fernando, C., Banarse, D., Blundell, C., Zwols, Y., Ha, D., Rusu, A.A., Pritzel, A., Wierstra, D.: Pathnet: Evolution channels gradient descent in super neural networks. CoRR **abs/1701.08734** (2017), http://arxiv.org/abs/1701.08734

5. Han, S., Mao, H., Dally, W.J.: Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint arXiv:1510.00149 (2015)
6. Han, S., Pool, J., Tran, J., Dally, W.: Learning both weights and connections for efficient neural network. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28, pp. 1135–1143. Curran Associates, Inc. (2015), http://papers.nips.cc/paper/5784-learning-both-weights-and-connections-for-efficient-neural-network.pdf
7. Hassibi, B., Stork, D.G., Wolff, G.J.: Optimal brain surgeon and general network pruning. pp. 293–299. IEEE (1993)
8. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 (2012)
9. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. pp. 4700–4708 (2017)
10. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural Computation **3**(1), 79–87 (Feb 1991). https://doi.org/10.1162/neco.1991.3.1.79, http://dx.doi.org/10.1162/neco.1991.3.1.79
11. Kiyono, S., Suzuki, J., Inui, K.: Mixture of expert/imitator networks: Scalable semi-supervised learning framework. vol. 33, pp. 4073–4081 (2019)
12. Krizhevsky, A., Hinton, G.E.: Using very deep autoencoders for content-based image retrieval. vol. 1, p. 2 (2011)
13. LeCun, Y., Denker, J.S., Solla, S.A.: Optimal brain damage. pp. 598–605 (1990)
14. Mocanu, D.C., Mocanu, E., Stone, P., Nguyen, P.H., Gibescu, M., Liotta, A.: Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. Nature communications **9**(1), 2383 (2018)
15. Mocanu, D.C., Mocanu, E., Stone, P., Nguyen, P.H., Gibescu, M., Liotta, A.: Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. Nature Communications **9**(1), 2383 (2018). https://doi.org/10.1038/s41467-018-04316-3, https://doi.org/10.1038/s41467-018-04316-3
16. Morris, G., Nevet, A., Bergman, H.: Anatomical funneling, sparse connectivity and redundancy reduction in the neural networks of the basal ganglia. Journal of Physiology-Paris **97**(4-6), 581–589 (2003)
17. Samsonovich, A.V., Ascoli, G.A.: A simple neural network model of the hippocampus suggesting its pathfinding role in episodic memory retrieval. Learning & Memory **12**(2), 193–208 (2005)
18. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017)
19. SPORNS, O., CHIALVO, D., KAISER, M., HILGETAG, C.: Organization, development and function of complex brain networks. Trends in Cognitive Sciences **8**(9), 418–425 (Sep 2004). https://doi.org/10.1016/j.tics.2004.07.008, http://dx.doi.org/10.1016/j.tics.2004.07.008
20. Sporns, O., Kötter, R.: Motifs in brain networks. PLoS biology **2**(11), e369 (2004)
21. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)
22. Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., Fergus, R.: Regularization of neural networks using dropconnect. pp. 1058–1066 (2013)

23. Watts, D.J., Strogatz, S.H.: Collective dynamics of ‘small-world’networks. nature **393**(6684), 440 (1998)
24. Zhigulin, V.P.: Dynamical motifs: building blocks of complex dynamics in sparsely connected random networks. Physical review letters **92**(23), 238701 (2004)
25. Zorins, A., Grabusts, P.: Artificial neural networks and human brain: Survey of improvement possibilities of learning. vol. 228, p. 231 (2015)