Parametric t-SNE Gated, Stochastically Initiated, Mixture of Sparse Experts

Daniel Byrne, Joanna Duran, Stacey Smith and John Santerre

February 14, 2020

Abstract

Deep neural networks (DNN) are founded on the principle of fully connected layers (FFL) which quadratically increase in the number of calculable parameters as the number of input dimensions and parameters increase. This presents a difficulty during the training and use of datasets with a large number of dimensions. Traditionally, researchers have simply increased the number of parameters to increase the capacity and learning ability of the network. However, research into sparsely connected networks, pruning, small-world networks and biological systems suggests that there may be other options. This paper introduces a new neural network that combines these insights into a cohesive model. The model uses a 2D Parametric t-Stochastic Neighbor Embedding(Pt-SNE) gating network to drive the gate inputs of a mixture of sparsely interconnected expert clusters. The expert clusters are not mutually exclusive nor independent, they share a low fan-out collection of inter-expert routes. They are also locally tuned to limit the number of interfering intra-expert connections. The team anticipates the result being a less dense, less computationally complex network that allows for information sharing from far flung expert networks through minimal influence connections.

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

1 Introduction

Deep learning's explosive growth has been fueled by the technique's success at embedding high dimensional data into multilevel encoded representations to solve complex functions and to model natural processes. This functionality has made it ideal for applications such as Natural Language Processing (NLP) and Computer Vision.

Deep Neural Networks are commonly designed with fully connected layers; that is, every node in a layer is connected to every node in adjacent layers. This architecture supplies each successive layer with all embeddings of its preceding layer in every possible combination. The training procedure has to effectively reduce this overabundance of overlapping connections by iteratively applying the delta rule to adjust the weights on the connections. The means tend to cluster around zero which indicate that a large

number of connections are redundant or destructively interfering causing the training procedure to zero them out. The result is a collection of weights and neurons that when applied an input, creates an alternative representation or classification at the network's output.

2 Problem

The issue that arises in training such over specified models is that the process is very time consuming and compute resource intensive. In a typical DNN, there may be millions or billions of parameters and similarly large set of labelled training examples. Another issue is that in a multi-class neural network, competing classifications with similar and dissimilar features will inject noise relative to each competing target classification. Neural networks have to be trained to discriminate the important features while at the same time filtering out the noise to iterate towards an adequately minimized loss functions which exhibits the ability to discriminate patterns from different classes.

An example is the set of weights that consistently identifies cats in images is likely in conflict with a set of weights tuned to correctly classify trees. There will be some overlap, and that overlapping region is a nonlinear objective function that the neural network is trying to optimize. Trying to reach a near optimal solution is not an easy task considering there are many possible combinations.

3 Advances in Techniques

There have been a number of solutions to address the issues of complexity and noise. The techniques vary, but the end results are that interfering patterns are filtered, and unique signals are amplified. Techniques researched are enabling auto-encoders, pooling layers, Optimal Brain Damage, Dropout, DropConnect, StochasticNets , pattern separation, PathNet and Sparsely Gated Mixture of Experts models.

Embedding auto-encoders has 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 decompression technique is to filter out nonessential, noisy, data points.

Pooling layers are often utilized to reduce the size of the feature maps in convolutional Neural Networks. The feature maps coming out of a filter bank are reduced before feeding to successive layers. This has the effect of reducing the number of interfering connections, reducing the computational complexity, and filtering noisy features.

Optimal Brain Damage is a method devised by Yann Le Cun, John Denker and Sara Solla in which they used second order derivatives to identify and prune unimportant weights from a network which they showed improved the network's generalization.

Dropout is a method which randomly eliminates a percentage of hidden units and their connections during each training case. Dropout is primarily used as a simple way of preventing Neural Networks from overfitting. During testing 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 the impact other hidden units.

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. This technique is not as widespread as Dropout but is gaining popularity.

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, Huizi Mao and William Dally introduce Deep Compression which chooses to learn both the weights on connections and the existence of connections during training. The 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 ratios 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 Deep Compression by starting with a sparse randomly connected network and ending with a sparse network. Pruning is thus inherent in their design.

Pattern separation a fundamental function of the brain and in their study Cayco-Gajic, Clopath and Silver study the pattern separation abilities of the cerebellum determined that feed forward 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 de-correlate their influences.

Decebal Constantin Mocanu et al. also took 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).

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 sub-networks 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.

4 Proposed Solution

The team will introduce new neural network architecture, Parametric t-SNE Gated, Stochastically Initiated, Mixture of Sparse Experts, that combines insights gleaned from these prior approaches into a cohesive model.

We propose a modification of this gated Mixture of Experts (MOE) model in which the gating network enforces a spatial separation between expert clusters using Pt-SNE. Pt-SNE is a method by which the Mutual Information components of similar Gaussian distributions are converted into Euclidean 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 MOE architectures which isolate adjacent clusters, we propose adding a stochastically sparse number of connections between expert clusters. This information sharing between clusters should reduce the greedy nature of training algorithms tendency to overusing some clusters while minimally using others. Clusters are also sparsely initialized using methods devised in Stochastic Net.

The team anticipates the result being a less dense, less computationally complex network that allows for information sharing from far flung expert networks through minimal influence connections.

References

[1] Danielle Smith Bassett and ED Bullmore. Small-world brain networks. *The neuroscientist*, 12(6):512–523, 2006.

[2] Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A Rusu, Alexander Pritzel, and Daan Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*, 2017.

[3] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *Advances in neural information processing systems*, pages 1135–1143, 2015.

[4] Olaf Sporns, Dante R Chialvo, Marcus Kaiser, and Claus C Hilgetag. Complex networks: small-world and scale-free architectures. *Trends in Cognitive Sciences*, 9(8):418–425, 2004.

[5] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.