Neural network pruning with simultaneous matrix tri-factorization

Teja ~~Ro~~Roštan

Abstract

In this paper we present ~~a novel~~ an approach for pruning neural ~~network to reduce~~ networks, which significantly reduces the model size ~~and achieve better~~ while maintaining its generalization performance. We apply a simultaneous matrix tri-factorization to map weight matrices to a low-dimensional space, therefore ~~shrinking~~ reducing them and partially eliminating noise. Factorized models are thus more robust and have a better generalization ability. ~~..~~

1 Introduction

Deep neural networks are a popular tool that is being used to solve ~~a certain problem~~widely different problems. The advantages of neural networks are that they are relatively easy to use and can approximate any function, regardless of its linearity. They are widely used for complex or abstract problems such as image, sound and text recognition. However, they are computationally intensive to train and are known for black box problem as they will not tell you why they reached a certain conclusion. Success of neural networks largely depends on their architecture. While the size of the input layer and the output layer is known, the number of hidden layers and the number of nodes in each hidden layer depends on the complexity of the problem [1]. Generally, a network with large number of hidden nodes is able to learn fast and avoids local minima, but when a network is oversized, the network may overfit the training data and lose its generalization ability while still having unnecessary calculations as they are using more nodes than necessary. Better generalization performance can be achieved only by small networks. They are easier to interpret but their training may require a lot of effort. Also too small networks are very sensitive to initial conditions and learning parameters and ~~are prone to not learn the given problem~~and do not generalize well. The most popular approach to obtain the most optimal architecture of neural network is pruning. Pruning is defined as a network trimming within the assumed initial architecture, which is larger than necessary. Pruning algorithms are used to remove the redundant connections while maintaining the networks performance. So one can use the larger networks for training and its generalization can be improved by the process of pruning [1].

More recent researches have tackled upon an issue of deep neural network and deep convolutional neural networks which is that they involve many layers with millions of parameters, making the size of the network model to be extremely large to store. This prohibits the usage on resource limited hardware especially mobile devices or other embedded devices even though deep neural networks are increasingly used in applications suited for mobile devices [10].

In this work we present a novel approach using low-dimensional matrix factorization. Because we have more than one weight matrix and because the weight matrices between the layers in a neural network are dependent with their neighbor matrices, we used an upgraded approach of matrix factorization, named simultaneous matrix tri-factorization, ~~or in other words,~~ also known as data fusion. Pruning neural network with simultaneous matrix tri-factorization was names as matrix factorization-based brain pruning (MFBP).

2 Related work

~~In article [7] they said that giving~~

(author?) [7] state that only a few weight values for each feature ~~it is possible~~ are needed to accurately predict the remaining values while many of them do not need to be learned at all. They exploited the fact that the weights in learned networks tend to be sparse and structured. Another article [1] have shown that, in any case the overall time required for training a large network and then pruning it to a small size compares very favorably with that of simply training a small network.

Because there is significant redundancy in the parametrization of networks, many researchers found solutions to prune neural networks with possible accuracy loss in order to reduce the model size extensively. But were able to fine-tune the compressed layers with added learning iterations to recover the performance and improve the accuracy back.

Compressing the most storage demanding dense connected layers is possible by neural network pruning with low-rank matrix factorization methods [4, 20, 19], where network pruning has been used both to reduce model size and to reduce over-fitting [11]. State-of-the-art approaches are Optimal Brain Damage [15] and Optimal Brain Surgeon [12] which open the rich field of studies using matrix factorization to prune the networks.

Besides neural network pruning with matrix factorization many alternatives have been used in numerous ways to optimize neural network architecture. One of the latest study [10] used vector quantization methods for which they said have a clear gain over existing matrix factorization methods. Alternative approach [24] is application of singular value decomposition (SVD) on the weight matrices to decompose them and reconstruct the model based on the sparseness of the original matrices. There were also studies which used evolutionary pruning, more precisely, Genetic Algorithms [16] to examine potential redundancy in data and therefore prune the neural network. A simple solution to reduce the model size and preserve the generalization ability is to train models that have a constant number of simpler neurons which was presented in article [6].

Another examined strong method uses the significance of neurons by evaluating the information on weight variation and consequently prune the insignificant nodes. Removing all connections whose weight is lower than a threshold is introduced in [11]. There the first phase learns which connections are important and removes the unimportant ones using multiple iterations. Hashing is also an effective strategy for dimensionality reduction while preserving generalization performance [23, 21]. The strategy used on neural networks named HashedNets [5] uses a low-cost hash function to randomly group connection weights into hash buckets where all connection inside share a single and tuned parameter value.

Compressing the parameters to reduce model size brings the focus upon how to prune the dense connected layers since the vast majority of weights reside in these layers which results in significant savings and by replacing the fully connected layers of the network with an Adaptive Fastfood transform, introduced in article [25], and results in a deep fried convnet. The Fastfood transform allows for a theoretical reduction in computation also. However, the computation in convolutional neural networks is dominated by the convolutions, and hence the deep fried convnets are not necessarily faster in practice.

3 ~~Method description~~Approximation of network weights with simultaneous matrix tri-factorization

Deep neural network is a feed-forward, artificial neural network with more than one or two hidden layers between the input and output layer 1. The number of nodes in every hidden layer is chosen manually. Our main goal was to use more nodes than necessary to prune the unnecessary ones later. The pruning was achieved using a low-dimensional approximation of original weight matrices to estimate which nodes are batter to prune. To approximate the weight matrices, we used simultaneous matrix tri-factorization.

~~Deep neural network with densed connected layers.~~

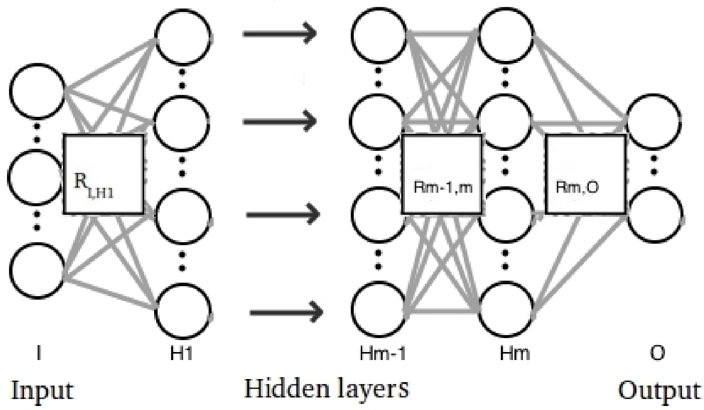

Figure 1: Deep neural network with dense connected layers. Relation matrix $W_{i,j}$ stores the weights of connections between neurons at layer i and j.

To do (1)

3.1 ~~Approximation of matrix with matrix factorization~~

Matrix factorization is a technique to search linear representation with factorizing. Approximation of matrix with matrix factorization is used to approximate the data in low-dimensional space in order to find latent features. ~~The theorem of the method matrix factorization follows in **??**.~~

~~Given is a matrix $X \in \Re^{m \times n}$ and a positive integer $k \ll min\{m, n\}$. Find matrices $U \in \Re^{m \times k}$ and $V \in \Re^{k \times n}$, such that $UV \approx X$. The product UV is named as a matrix factorization of a matrix X. UV is an approximated factorization if rang $r = k$ **??**. Graphical visualization of matrix factorization.~~

With ordinary artificial neural network, we have only one hidden layer and therefore two weight matrices with sharing dimension. Because of this property, we are able to concatenate the matrices through sharing dimension and apply a matrix factorization.

With deep neural networks we have a multi-layer architecture where only neighbour weight matrices share the same dimension. We can apply co-dependency between neighbour weight matrices but we can not apply dependency between, for example, first and third weight matrix. Our goal was to consider all relations that exist between weight matrices in deep neural network. Simultaneous matrix tri-factorization applies our criteria.

3.1 ~~Simultaneous matrix tri-factorization~~

Configuration of relation matrices W_{ij} from figure 1 for simultaneous matrix tri-factorization. In our case, the relation matrices W_{ij} present weight matrices of neural network. Configuration is set on diagonal because the neighbour weight matrices share the dimension from shared hidden layer.

Theorem 3.1 *Simultaneous tri-factorization of multiple matrices simultaneously factorize all available relation matrices R_{ij} into $G_i \in \Re^{m \times k}$, $G_j \in \Re^{n \times h}$ and $S \in \Re^{k \times h}$ and regularize their approximations through constrained matrices θ_i and θ_j, such that $R_{ij} \approx G_i S_{ij} G_j^T$ [26] 2.*

The theorem of simultaneous matrix tri-factorization 3.1. We can reduce the number of nodes (parameters) in network as long as the number of parameters in G_i and G_j is less than the number of parameters in X. If we would like to reduce the number of parameters in X by a fraction of p [19], we require the equation 1 to hold.

$$mk + kh + hn < pmn \tag{1}$$

To do (2)

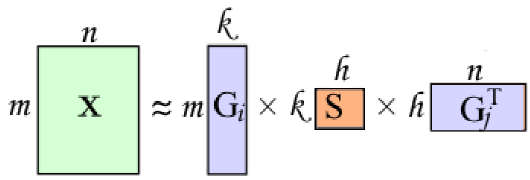

Figure 2: Graphical visualization of simultaneous matrix tri-factorization.

In a figure ~~??~~ 1 is shown a neural network with two hidden layers and their relation weight matrices ~~R_{ij}~~ W_{ij} between them. The weight matrices are collected from neural network and configured in a matrix of relations ~~R~~ W as shown in figure ~~??~~1. A block in the i-th row and j-th column (~~R_{ij}~~W_{ij}) of matrix ~~R~~ W represents the relationship between object type ξ_i and ξ_j. In case of a neural network, these represent neurons at layers i and j, respectively. The block matrix ~~R~~ W is tri-factorized into block matrix factors G and S. A factorization rank k_i is assigned to ξ_i during inference of the factorized system. Factors S_{ij} define the relations between ~~object types~~ layers ξ_i and ξ_j, while factors G_i are specific to ~~objects of type~~ layers ξ_i and are used in the reconstruction of every relation with this ~~object type~~layer. In this way, each ~~relation matrix R_{ij}~~ weight matrix W_{ij} obtains its own factorization $G_i S_{ij} {G_j}^T$ with factor G_i (G_j) that is shared across relations which involve ~~object types~~ layers ξ_i (ξ_j). The objective function minimized by penalized matrix tri-factorization ensures good approximation of the input data and adherence to must-link and cannot-link constraints [26].

To do (3)

With approximations we determined which weights are better to prune. We pruned weights which hold followed criteria and were forced to a zero value to be considered as pruned: ~~$abs(original_weight) - abs(approximated_weight) < threshold.$~~

$abs(original_weight) - abs(approximated_weight) < threshold$

To do (4)

4 Experimental setup

We evaluated matrix factorization-based brain pruning on MNIST (Mixed National Institute of Standards and Technology dataset) dataset. The MNIST database of handwritten digits 0-9, available in [14], has a training set of 60~~,~~ 000 instances and a test set of 10~~,~~ 000 instances. The digits have been size-normalized and centered in a fixed-size 28x28 images.

We used a modern neural network, presented in [17]. There are two main contributions to a modern neural network. One is changing of activation function. Instead of sigmoid function it uses a rectifier (Rectified linear unit (ReLU) $f(x) = max(0, x)$, where x is the input to a neuron. What the rectifier does is that if the input to a neuron is below zero the activation function does nothing. If the input is above zero it does activate. This activation function has been argued to be more biologically plausible [8]. It induces the sparsity in the hidden neurons. Another advantage of rectifier is that it does not face gradient vanishing problem as with sigmoid or tanh function. It has ben also shown that can be deep neural networks trained efficiently using rectifier even without pre-training. The other contribution is regularizing the model with dropout [22]. Dropout is one of the biggest improvements in the field of neural networks in recent years as it addresses the main problem in deep learning that is overfitting. The purpose of dropout is to add some noise by dropping out a random number of some neuron activations in a given layer. By dropping them is meant to set them to zero or as in our case to prune them. With every iteration a different random set of neurons are chosen to drop, therefore it prevents co-adaptation of neurons. There was also a change at update rule. Instead of a standard stochastic gradient descent (SGD) backpropagation method we used RMSprop (A mini-batch version of rpop). The idea behind SGD is to approximate the real update step by taking the average of the all given instances or as in our case mini batches. The problem of SGD is that it is sensitive to outliers which can destroy all the gradient information collected before [9]. On the other hand, the RMSprop keeps a running average of its recent gradient magnitudes and divides the next gradient by this average so that loosely gradient values are normalized [13]. RMSprop follows: $MeanSquare(w,t) = 0.9MeanSquare(w,t-1) + 0.1(\delta E/\delta w^{(t)})^2$ [13].

To evaluate our experiments, we implemented algorithm on Python with the help of Theano [2, 3]. Theano is a Python library that is suitable for building an optimized neural network. We chose it as it gives a comprehensive control over neural network formation which is suitable for our problem. Another reason we used Theano is because the implementation of modern neural net described above is available online as open source. Data fusion algorithm which performs simultaneous matrix tri-factorization is available in a python library Scikit-fusion [26]. To measure our results, we used a machine learning library Scikit-learn [18].

To estimate and analyze our results, we trained and tested four neural networks: ordinary neural network with two hidden layers without pruning, ordinary neural network with two hidden layers with pruning, deep neural network with five hidden layers without pruning and another deep neural network with five hidden layers with pruning. Every neural network had 60 iterations available to learn. The non-pruned networks learned on 60 iterations where every iteration had 50, 000 train instances packed in mini-batches of 128. The pruned network had 40 iterations to learn without pruning. The other 20 iterations consisted of every second iteration of pruning (overall 10 iterations of pruning) and another half of iterations for fine-tuning. Fine-tuning was used to adapt the non-pruned weights which have been affected with pruning, in other words, to recover the non-pruned weight values which have been biased by the pruned weights before pruning. In this case there were also all 50, 000 train instances available at every iteration in mini-batches.

5 Results

The reported results are measured with area under ROC curve (AC) on test set and shown in figure 3. The model size compression rate in % results are coming...

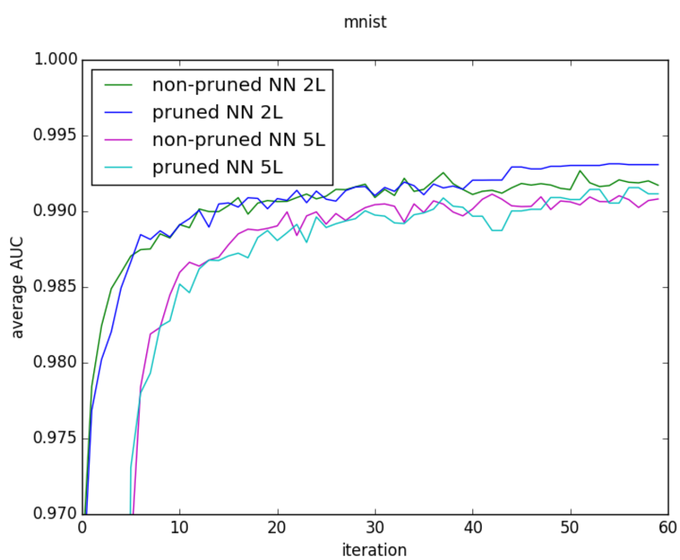

Figure 3: AUC results of neural networks. With pruned neural networks, the pruning starts at 40th iteration.

6 Discussion and conclusion

To be written...

Acknowledgments

References

References

[1] M Gethsiyal Augasta and T Kathirvalavakumar. Pruning algorithms of neural networksa comparative study. *Central European Journal of Computer Science*, 3(3):105–115, 2013.

[2] Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard, and Yoshua Bengio. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

[3] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June 2010. Oral Presentation.

[4] Andrey Bondarenko and Arkady Borisov. Artificial neural network generalization and simplification via pruning. *Information Technology and Management Science*, 17(1):132–137, 2014.

[5] Wenlin Chen, James T Wilson, Stephen Tyree, Kilian Q Weinberger, and Yixin Chen. Compressing neural networks with the hashing trick. *arXiv preprint arXiv:1504.04788*, 2015.

[6] Maxwell D Collins and Pushmeet Kohli. Memory bounded deep convolutional networks. *arXiv preprint arXiv:1412.1442*, 2014.

[7] Misha Denil, Babak Shakibi, Laurent Dinh, Nando de Freitas, et al. Predicting parameters in deep learning. In *Advances in Neural Information Processing Systems*, pages 2148–2156, 2013.

[8] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In Geoffrey J. Gordon and David B. Dunson, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS-11)*, volume 15, pages 315–323. Journal of Machine Learning Research - Workshop and Conference Proceedings, 2011.

[9] Eren Golge. Comparison: Sgd vs momentum vs rmsprop vd momentum+rmsprop vs adagrad, 2015. Avaliable: `http://www.erogol.com/` [Reached 17.1.2016].

[10] Yunchao Gong, Liu Liu, Ming Yang, and Lubomir D. Bourdev. Compressing deep convolutional networks using vector quantization. *CoRR*, abs/1412.6115, 2014.

[11] Song Han, Jeff Pool, John Tran, and William J Dally. Learning both weights and connections for efficient neural networks. *arXiv preprint arXiv:1506.02626*, 2015.

[12] Babak Hassibi, David G Stork, and Gregory J Wolff. Optimal brain surgeon and general network pruning. In *Neural Networks, 1993., IEEE International Conference on*, pages 293–299. IEEE, 1993.

[13] Geoffrey Hinton. Lecture 6a overview of mini-batch gradient descent, 2014. Avaliable: `www.cs.toronto.edu/` [Reached 17.1.2016].

[14] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[15] Yann LeCun, John S Denker, Sara A Solla, Richard E Howard, and Lawrence D Jackel. Optimal brain damage. In *NIPs*, volume 89, 1989.

[16] Xiangmei Li. Tuning the structure and parameters of a neural network by a new network model based on genetic algorithms. *International Journal of Digital Content Technology & its Applications*, 6(11), 2012.

[17] Alec Radford (newmu). Theano-tutorials, 2015. Avaliable: `http://www.github.com` [Reached 27.1.2016].

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[19] Tara N Sainath, Brian Kingsbury, Vikas Sindhwani, Ebru Arisoy, and Bhuvana Ramabhadran. Low-rank matrix factorization for deep neural network training with high-dimensional output targets. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6655–6659. IEEE, 2013.

[20] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[21] Qinfeng Shi, James Petterson, Gideon Dror, John Langford, Alex Smola, and SVN Vishwanathan. Hash kernels for structured data. *The Journal of Machine Learning Research*, 10:2615–2637, 2009.

[22] 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.

[23] Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1113–1120. ACM, 2009.

[24] Jian Xue, Jinyu Li, and Yifan Gong. Restructuring of deep neural network acoustic models with singular value decomposition. In *INTERSPEECH*, pages 2365–2369, 2013.

[25] Zichao Yang, Marcin Moczulski, Misha Denil, Nando de Freitas, Alex Smola, Le Song, and Ziyu Wang. Deep fried convnets. *arXiv preprint arXiv:1412.7149*, 2014.

[26] Marinka Zitnik and Blaz Zupan. Data fusion by matrix factorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(1):41–53, 2015.

To do…

- ☐ 1 (p. 3): Edit figure to what sketched in current version. Write W (for weights) instead of R in figure.
- ☐ 2 (p. 3): This is valid for one matrix (i.e., layer in the network). Write a general equation for the entire network.
- ☐ 3 (p. 4): Should be rewritten with factorization of weights matrices in mind.
- ☐ 4 (p. 4): This seems wrong and counterintuitive. Is it really so?