Genetic Algorithms for Evolving Deep Neural Networks

Omid E. David
Bar-Ilan University
Ramat-Gan 52900, Israel
mail@omiddavid.com

Iddo Greental
Tel Aviv University
Tel Aviv 69978, Israel
iddo.greental@gmail.com

ABSTRACT

In recent years, *deep learning* methods applying unsupervised learning to train deep layers of neural networks have achieved remarkable results in numerous fields. In the past, many genetic algorithms based methods have been successfully applied to training neural networks. In this paper, we extend previous work and propose a GA-assisted method for deep learning. Our experimental results indicate that this GA-assisted approach improves the performance of a deep autoencoder, producing a sparser neural network.

Categories and Subject Descriptors: I.2.6 [Artificial Intelligence]: Learning—*Connectionism and neural nets*

General Terms: Algorithms.

Keywords: Genetic algorithms, Deep learning, Neural networks, Autoencoders

1. INTRODUCTION

While the motivation for creating deep neural networks consisting of several hidden layers has been present for many years, supported by a growing body of knowledge on the deep architecture of the brain and advocated on solid theoretical grounds [1, 3], until recently it was very difficult to train neural networks with more than one or two hidden layers.

Recently, *deep learning* methods which facilitate the training of neural networks with several hidden layers have been the subject of increased interest, owing to the discovery of several novel methods. Common approaches employ either *autoencoders* [2, 10] or *restricted Boltzmann machines* [5, 8, 9] to train one layer at a time in an unsupervised manner.

In the past, genetic algorithms have been applied successfully to training neural networks of shallow depths (one or two hidden layers) [11]. In this paper we demonstrate how genetic algorithms can be applied to improve the training of deep autoencoders.

2. DEEP AUTOENCODERS

In this section, we briefly describe autoencoders and explain how they are used in the context of deep learning.

GECCO'14, July 12–16, 2014, Vancouver, BC, Canada.
ACM 978-1-4503-2881-4/14/07.
http://dx.doi.org/10.1145/2598394.2602287 .

An autoencoder is an unsupervised neural network which sets the target values (of the output layer) to be equal to the inputs, i.e., the number of neurons at the input and output layers is equal, and the optimization goal for output neuron i is set to $y_i = x_i$, where x_i is the value of the input neuron i. A hidden layer of neurons is used between the input and output layers, and the number of neurons in the hidden layer is usually set to fewer than those in the input and output layers, thus creating a bottleneck, with the intention of forcing the network to learn a higher level representation of the input. The weights of the encoder layer (W) and the weights of the decoder layer (W') can be *tied* (i.e., defining $W' = W^T$).

Autoencoders are typically trained using backpropagation. When an autoencoder's training is completed, we can discard the decoder layer, fix the values of the encoder layer (so the layer can no longer be modified), and treat the outputs of the hidden layer as the inputs to a new autoencoder added on top of the previous autoencoder. This new autoencoder can be trained similarly. Using such layer-wise unsupervised training, deep stacks of autoencoders can be assembled to create deep neural networks consisting of several hidden layers (forming a *deep belief network*). Given an input, it will be passed through this deep network, resulting in high level outputs. In a typical implementation, the outputs may then be used for supervised classification if required, serving as a compact higher level representation of the data.

3. GA-ASSISTED DEEP LEARNING

Genetic algorithms (GA) have been successfully employed for training neural networks [11]. Specifically, GAs have been used as substitute for the backpropagation algorithm, or used in conjunction with backpropagation to improve the overall performance.

We now propose a simple GA-assisted method which (according to our initial results presented in the next section) improves the performance of an autoencoder, and produces a sparser network.

When training an autoencoder with tied weights (i.e., the weights of the encoding layer are tied to those of the decoding layer), we store multiple sets of weights (W) for the layer. That is, in our GA population each chromosome is one set of weights for the autoencoder. For each chromosome (which represents the weights of an autoencoder), the root mean squared error (RMSE) is calculated for the training samples (the error for each training sample is defined as the difference between the values of the input and output layers). The fitness for chromosome i is defined as $f_i = 1/RMSE_i$. After calculating the fitness score for all the chromosomes, they are sorted from the fittest to the least fit. The weights of the high rank-

ing chromosomes are updated using backpropagation, and the lower ranking chromosomes are removed from the population. The removed chromosomes are replaced by the offsprings of the high ranking chromosomes. The selection is performed uniformly with each of the remaining chromosomes having an equal probability for selection (regardless of the fitness values of the chromosomes, i.e., the fitness score is used only for determining which chromosomes are removed from the population). Given two parents, one offspring is created as follows: Crossover is performed by randomly selecting weights from the parents, and mutation is performed by replacing a small number of weights with zero.

Gradient descent methods such as backpropagation are susceptible to trapping at local minima. Our method assists backpropagation in this respect, reducing the probability of trapping at local minima. Additionally, mutating the weights to zero encourages sparsity in the network (fewer active weights). Sparse representations are appealing due to information disentangling, efficient variable-size representation, linear separability, and distributed sparsity [4].

Note that when training of an autoencoder is complete, the values of the best chromosome are selected for that autoencoder. These values are fixed and shared amongst all chromosomes when a new autoencoder layer is added on top of the previously trained layer. Thus, each chromosome contains only the values of the layer currently being trained.

4. EXPERIMENTAL RESULTS

For our experiments we used the popular MNIST handwritten digit recognition database [7]. In the MNIST dataset, each sample contains 784 pixels (28x28 image), each having a grayscale value between 0 to 255 (which we scale to a 0 to 1 range). Each sample also contains a target classification label (between 0 and 9), which is used for the subsequent supervised classification phase (using the high level representations generated by the unsupervised autoencoder).

Our deep neural network uses a stack of 5 layers. The first layer has 784 neurons, followed by four higher level layers consisting of 500, 250, 100, and 50 neurons. Each layer is trained separately, with the next layer added only once training is complete: first we train the 784-500 layer, then use the 500 output neurons as inputs to the 500-250 layer, and similarly for the 250-100 and 100-50 layers.

The GA implementation uses a population of 10 chromosomes. In each generation, the five worst chromosomes (half the population) are removed and replaced by the offsprings of the five best chromosomes. We used crossover and mutation rates of 0.8 and 0.01 accordingly.

To compare the performance of our GA-assisted method with traditional backpropagation, we ran both methods under similar conditions. First, we ran the traditional backpropagation version 10 times and selected the result with the least reconstruction error (best tuned). Next, we ran the GA-assisted method only once, allowing the same total runtime as the previous method. Comparing the reconstruction errors of the two approaches, the GA-assisted method consistently yielded a smaller reconstruction error, as well as a sparser network.

In order to compare the classification accuracy of the two methods, we ran 10,000 new test samples through the two trained networks and recorded the 50 output values for each sample. Recall that in this test phase the weights of the network are already fixed, hence an input sample of 784 values is passed through the layers of 500, 250, 100, and 50 neurons without modifying their weights. The representation quality of the networks can be compared by applying supervised classification to the higher level values produced by the 50 neurons of the output layer. We used SVM classification with a radial basis function (RBF) kernel. Using SVM, the traditional autoencoder achieved a 1.85% classification error, while the GA-assisted method's classification error was 1.44%.

5. CONCLUDING REMARKS

In this paper we presented a simple GA-assisted approach, which according to our initial results improves the performance of a deep autoencoder. While our implementation used an autoencoder, the same method is applicable to other forms of deep learning such as restricted Boltzmann machines (RBM).

In recent years, several improvements upon traditional autoencoders and RBM have been proposed which improve their generalization. Such improvements include *dropout* [6], which randomly disables some neurons during training, *dropconnect* [13], which randomly disables some weights during training, and *denoising autoencoders* [12], which randomly add noise by removing a portion of the training data. The improved performance of the GA-assisted autoencoder could arise from a similar principle, since mutation randomly disables some of the weights during training. It is important to compare the GA-assisted approach to the above mentioned alternative improvements in future research.

6. REFERENCES

[1] Y. Bengio and Y. LeCun. Scaling learning algorithms towards AI. *Large Scale Kernel Machines*. MIT Press, 2007.

[2] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. *Advances in Neural Information Processing Systems 19*, pages 153–160. MIT Press, 2007.

[3] Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

[4] X. Glorot, A. Bordes and Y Bengio. Deep sparse rectifier neural networks, *14th International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2011.

[5] G.E. Hinton, S. Osindero, and Y.W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

[6] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR* abs/1207.0580, 2012.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[8] G.E. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[9] H. Lee, C. Ekanadham, and A. Ng. Sparse deep belief net model for visual area V2. *Advances in Neural Information Processing Systems 20*, pages 873–880, MIT Press, 2008.

[10] M. Ranzato, C.S. Poultney, S. Chopra, and Y. LeCun. Efficient learning of sparse representations with an energy-based model. *Advances in Neural Information Processing Systems 19*, pages 1137–1144. MIT Press, 2007.

[11] J.D. Schaffer, D. Whitley, and L.J. Eshelman. Combinations of genetic algorithms and neural networks: a survey of the state of the art. *International Workshop on Combinations of Genetic Algorithms and Neural Networks*, pages 1–37, 1992.

[12] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371-3408, 2010.

[13] L. Wan, M.D. Zeiler, S. Zhang, Y. LeCun, and R. Fergus. Regularization of neural networks using dropconnect. *International Conference on Machine Learning*, pages 1058–1066. JMLR.org, (2013)