Robust and Compact Deep Networks Using Information Theory/ Feature-channel reduction in Deep Convolutional neural network

Prad Kadambi, pkadambi@asu.edu, ASUID 1204813671

Gaurav Srivastava, gsrivas4@asu.edu, ASUID 1209591107

Abstract—The focus of this project is to create compact neural networks. For this we have adopted two problem statements. The problem that we will finally attempt to solve will be based on some initial analysis.

Problem 1: Deep neural networks (DNNs) have proved to be vital to achieve human like performance in many classic machine learning problems such as image recognition, natural language processing, and speech processing. However, significant computational challenges exist in implementing deep networks due to the large model size. Various information theory based [1] and sparsity based [7] methods have been applied to compress deep networks. We will define Fisher information based network regularization term within the cost function to create robust and compact networks.

Problem 2: The second Idea is to make Deep convolutional neural network compact by finding the redundant information within channels and feature maps. Some works [8], [9], [10], [11] have attempted to relate the information across feature maps and channels. Extension of this work will be to find the correlation between these feature maps. study and utilization of 3D convolutions and applying KL divergence across activations in adjacent layers.

Index Terms— Information theory, Fisher information Matrix (FIM), Deep Neural Network (DNN), Graphics Processing Unit (GPU), Field Programmable Gate Array (FPGA), Application Specific Integrated Circuit (ASIC), Convolutional Neural Network (ConvNet)

1 Literature survey

Modern deep neural networks have achieved tremendous performance on many learning tasks. However many such networks like GoogLeNet contain millions of parameters. (GoogLeNet required more than 60 million parameters). This has lead to the requirement of specialized hardware such as GPUs, FPGAs, and ASICs to support the massively parallel computation required for DNNs. Considerable interest exists in reducing the size of these deep networks in order to: 1) port models to mobile platforms, 2) reduce overfitting by pruning extraneous parameters, 3) make networks robust to adversaries.

Problem 1: Previous work such, as optimal brain damage [3], has focused on removing weights that present low values for the Hessian: if small changes in network parameters do not notably change the output, they are removed. In [4] an information theoretic metric is used to compare a compressed "student" network to a larger teacher network. A cost term consisting of the cross-entropy between the outputs of the student and teacher networks ensures that the smaller network can approximate the larger one.

Problem 2: Previous work to generalize activation values between feature-channel has been done in [8]. It uses a maxout strategy to propagate the resultant activation values across features by taking maximum amongst them. Other papers [9], [10], [11] are using variations of 3d convolution for image recognition and classification purposes.

2 Problem statement

Problem 1: In [1] the Fisher Information Matrix (FIM) - a parameter describing how much information each weight has about the network output - is estimated when training the network. After training, the parameters are clustered based on their corresponding value in the FIM diagonal, and parameters with small FIM entries are pruned. We attempt to use the FIM to answer a larger subset of problems: Can we train deep networks to ensure their outputs are robust to changes in weights by using information theoretic constraints? We expand on Tu et al's work by computing the FIM *within* the training process.

A network robust to changes in weights has significant advantages. If weights are deleted, or tampered by adversarial agents, the predictions can be guaranteed to stay the same. Weights that are small (close to zero) can be removed, and weights can be quantized, all without significant loss of performance.

Problem 2: The attempt will be made to make convolutional neural network sparse. Some of the ConvNets have huge feature maps, such as VGGnet and ResNet have upto 500 feature maps. It is a problem of interest to study if all of the feature maps are contributing to the classification task. Reduction in the number of these feature maps can result in ConvNets which can be trained faster with less power. The approach to make ConvNets

sparse will be made by attempting to embed more information across features. This can be done by theoretical pruning of signals based on information theory.

3 PLANNED WORK

Problem 1: To ensure that we train a robust network, we must add a cost to networks that have large responses to changes in their parameters. The natural gradient descent (NGD) algorithm[5] is one such method. NGD minimizes a cost function such that successive iterations of parameters have the same K-L divergence. Estimating the K-L divergence between two sets of parameters is an ill posed problem, so a Taylor expansion of the K-L divergence is related to the FIM [2].
Our first objective will be to come up with a functioning learning algorithm that utilizes the Fisher information as a regularization parameter within the context of the Adam algorithm. We use Adam because it has some similarities to NGD, and Adam employs the FIM in its preconditioning [6].
Once we are able to find FIM estimates during training we will find a method to prune and quantize weights during training. We apply this method to the network described in [1]. Time permitting, we will then turn our optimization study to the recent deep networks that have succeeded in ImageNET.

Problem 2: To find the relationship between the features of a ConvNet, one mathematical tool we intend to use is KL divergence, to find the layers which are least contributing to the output. This can result in pruning of those layers or a portion of those layers. Another direction of attempt will be to find an extension of 3D convolution which will extract more relationships across the feature maps.

4 PLANNED ANALYSIS AND EXPERIMENTS

Problem 1: On the MNIST dataset, we will compare the accuracy of our compressed network to the network in [1]. We will find the accuracy of the network as a function of the percentage of removed parameters. Then, we compare the average number of bits used by the parameters of the network versus the accuracy. We compare robustness by subjecting the weights to increasing amounts of additive noise and finding the accuracy for different amounts of noise. The analysis will also be expanded to the CIFAR-10 dataset, and the compression rate will be compared with other state of the art network compression methods.

Problem 2:The project for feature reduction within ConvNet, will use CIFAR-10 or Imagenet dataset. As the goal of the task is to reduce the redundancy between features, a network will be taken which has higher feature depth. Thus VGGnet or ResNet are candidates on which these experiments will be applied. The metric of good reduction of features would be to make reduction in feature maps without incurring considerable reduction in validation and test error.

5 REFERENCES

[1] Tu, M., Berisha, V., Woolf, M., Seo, J. S., & Cao, Y. (2016). Ranking the parameters of deep neural networks using the fisher information. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016 - Proceedings* (Vol. 2016-May, pp. 2647-2651). [7472157]

[2] Razvan Pascanu and Yoshua Bengio. Revisiting natural gradient for deep networks. In ICLR, 2014

[3] Yann LeCun, John S Denker, Sara A Solla, Richard E Howard, and Lawrence D Jackel. “Optimal brain damage.” In *Advances in Neural Information Processing Systems*, volume 2, pages 598–605, 1989.

[4] Hinton, G. Vinyals, O. and Dean, J. “Distilling knowledge in a neural network.” In Deep Learning and Representation Learning Workshop, , 2014

[5] Amari, S. (1997). Neural learning in structured parameter spaces - natural Riemannian gradient. In In Advances in Neural Information Processing Systems, pages 127–133. MIT Press

[6] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations, 2015.

[7] S. Han, H. Mao, and W. Dally. Deep compression: Compressing DNNs with pruning, trained quantization and huffman coding. arxiv:1510.00149v3, 2015a.

[8] Ian J. Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, Yoshua Bengio, Maxout Networks, Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28

[9] Daniel Maturana1 and Sebastian Scherer, 3D Convolutional Neural Networks for Landing Zone Detection from LiDAR, Robotics Institute, Carnegie Mellon University

[10] Daniel Maturana1 and Sebastian Scherer, VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition, Robotics Institute, Carnegie Mellon University

[11] Adhish Prasoon, Kersten Petersen, Christian Igel, François Lauze, Erik Dam, Mads Nielsen, Deep Feature Learning for Knee Cartilage Segmentation Using a Triplanar Convolutional Neural Network, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013 pp 246-253