Memory Efficient Nonuniform Quantization for Deep Convolutional Neural Network

Fangxuan Sun and Jun Lin

***Abstract*—Convolutional neural network (CNN) is one of the most famous algorithms for deep learning. It has been applied in various applications due to its remarkable performance. The real-time hardware implement of CNN is highly demanded due to its excellent performance in computer vision. However, the cost of memory of a deep CNN is very huge which increases the area of hardware implementation. In this paper, we apply several methods in the quantization of CNN and use about 5 bits for convolutional layers. The accuracy lost is less than 2% without fine tuning. Our experiment is depending on the VGG-16 net [1] and Alex net [2]. In VGG-16 net, the total memory needed after uniform quantization is 16.85 MB per image and the total memory needed after our quantization is only about 8.42 MB. Our quantization method has saved 50.0% of the memory needed in VGG-16 and Alex net compared with the quantization method in [3].**

***Index Terms*—deep learning, convolutional neural network, nonuniform, quantization, k-means clustering**

I. Introduction

CNN is one of the most famous algorithms of deep learning. It has been mainly applied in image classification and has been proved as a powerful method [2] [1]. Due to the fantastic performance of CNN, many other computer vision applications also employ CNN as a powerful tool to improve their performance. For examples, CNN has achieved great performance in image annotation [4], visual QA system [5], 3D interpreter [6] and other many areas. Moreover, CNN is also applied in speech recognition [7] and achieve some great results compared to the methods before.

The most common way to improve the performance of CNN on an application is to increase the depth of the convolution layer. In many cases, increasing the depth of convolutional layer can achieve a better result compared to the CNN with less layers. However, cause the higher computational complexity, the training and testing of deeper CNN is mainly depended on multiple GPUs which means CNN is hard to be used on mobile device. Moreover, the deep CNN is difficulty to be used on the local mobile processor not only because the high computational complexity but also due to the huge parameters generated. The deep CNNs can generate tens or even hundreds MB parameters during it's computation which makes the local mobile processor hard to handle. Due to the reasons above, an optimized hardware implement of CNN is highly demanded so that it can be processed on local processor with high speed and low power and the various application can be performed on mobile device. One approach to decrease the memory space needed is to use fixed point number rather than float point number in the computation of CNN which can lower the memory needed for CNN in applications.

In this paper, we apply several methods to implement deep CNN in fix point case. We use uniform quantization, nonuniform quantization and k-means clustering methods on VGG-16 [1] net respectively. We also do experiments on alex-net [2]. As we know, we are the first to use nonuniform quantization on the deep CNN and used only about 5 bit in most layers of VGG-16 net and keep the precision lost less than 2%. According to our experiment result, we have save about 84.4% memory compared with the original memory needed. In Alex net, we save 90.6% memory as well.

The rest of the paper is organized as follows. Distribution of CNN is in Section II. The method of fixed point quantization of deep CNN is presented in Section III. Comparison of the result can be seen in Section IV. At last, the conclusion is drawn in Section V.

II. Distribution of CNN

To use fixed point quantization of deep CNN, we need to know the distribution of the data of convolutional layers. [3] has done some experiments about the distribution of each layers in CIFAR-10 benchmark CNN. According to their experiments, the distribution of data in most layers are roughly Gaussian distribution. To find out whether the distribution in deeper CNN is Gaussian-like distribution as well, we check the data of some deeper networks like VGG-16 net and Alex net on the dataset of imagenet. The distribution of those layers is shown in Fig. 1.

The figure plotted above is the data extracted from the convolutional layer of VGG-16 net. The data presented is before the the computation of activation function which in VGG-16 is RELU function. As the figures above show, the distribution of VGG-16 net is Gaussian-like as well. We also do experiments on the Alex net and get the similar results. Because the data after RELU layer are all equal or greater than zero, our quantization doesn't take sign it into account.

III. Fixed point quantization of deep CNN

A. Methods for Quantization

Our quantization is organized as follows:

- According to the distribution of data in VGG-16 net, we take a general bit-width for all layers and test the accuracy. Assuming the threshold of accuracy lost is δ and the original accuracy of VGG-16 net is X, and the

J. Lin is with the school of electronic science and engineering, Nanjing University, China (jlin@nju.edu.cn)

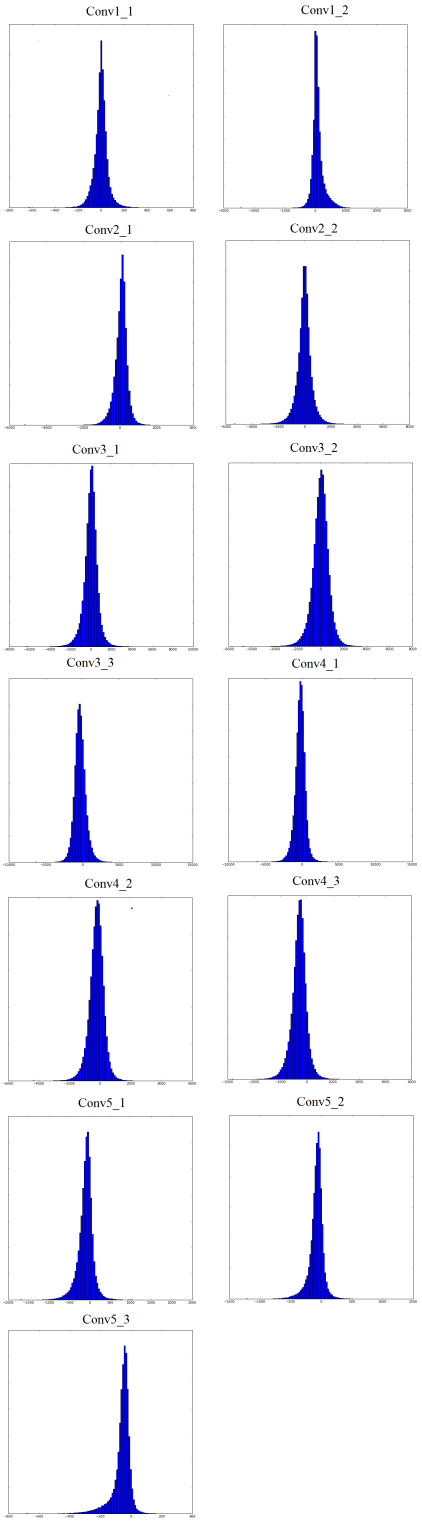

Fig. 1. Distribution of activations in VGG-16 net

accuracy with this threshold is $X^{'} = X - \delta$. We take $X^{'}$ as our floor during the experiment [3].

- After the uniform quantization, we apply nonuniform quantization method in which each layer has a different quantization bit. After a large amount of experiments, we decide the quantization bit for each layer.
- To further decrease the bit cost, we use discrete nonuniform quantization. A group of numbers with uniform space are applied to quantity the activation data. The bit cost is to judge how many numbers we use each layer. The generation of numbers is depended on the maximum and minimum of the data of each layer.
- Considering the discrete quantization does net make full use of the data of each layer. We employ K-means clustering method in the pre-process of the discrete numbers. The number after K-means is more suitable to the distribution of the VGG-16 net. We also use a upper bound of the data of each layer so that the large number won't affect the clustering greatly. The comparison of Discrete and K-means can be seen in Fig. 2

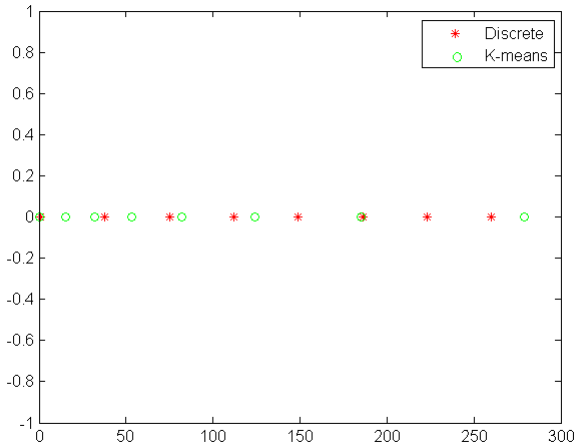

Fig. 2. Discrete and K-means

B. Uniform Quantization

We first do the uniform quantization of VGG-16 net according to the distribution above. The accuracy of the full precision is 88.5%. We use the same quantization bit for all layers without fine-tuning. The result can be seen in I.

TABLE I
UNIFORM QUANTIZATION OF VGG-16 NET

Quantization bit	14	13	12	11	10	9
Top5 Accuracy(%)	88.4	88.4	88.5	87.0	79.6	56.0

According to the activation data, some are even larger than 15000, but our quantization experiments show that we only need 12 bit to quantity the activation data.

C. Nonuniform Quantization

As we see in the distribution, the range of activation data in different layers is totally different, so we choose to apply nonuniform quantization on different layers. We use different quantization bit in different layers and do a series of experiments, part of the result is shown in Table. II III.

Compared with the uniform quantization, the conv1_1, conv1_2 layer which have the most data use only 8 bit rather than 12 bit.

TABLE II
NONUNIFORM QUANTIZATION OF VGG-16 NET PART 1

Top5 Accuracy(%)	Index	conv1_1	conv1_2	conv2_1	conv2_2	conv3_1	conv3_2	conv3_3	conv4_1	conv4_2	conv4_3
83.96	1	8	10	10	11	11	11	11	11	11	10
86.44	2	8	10	10	11	11	11	11	11	11	10
85.61	3	8	10	10	11	11	11	11	11	11	10
78.20	4	8	10	10	11	11	11	11	11	11	10
86.20	5	8	10	10	11	11	11	11	11	11	10
84.97	6	8	10	10	11	11	11	11	11	11	9
83.33	7	8	10	10	11	11	11	11	10	11	10
81.07	8	8	10	11	10	11	11	11	11	11	10
86.22	9	8	8	10	11	11	11	11	11	11	10

TABLE III
NONUNIFORM QUANTIZATION OF VGG-16 NET PART 2

Top5 Accuracy(%)	Index	conv5_1	conv5_2	conv5_3	fc6	fc7	fc8
83.96	1	10	9	8	5	4	4
86.44	2	10	9	8	7	4	3
85.61	3	10	9	7	7	3	2
78.20	4	10	9	5	7	3	2
86.20	5	11	9	8	7	3	2
84.97	6	10	9	8	7	3	2
83.33	7	10	9	8	7	3	2
81.07	8	10	9	8	7	3	2
86.22	9	10	9	8	7	3	2

D. Discrete Nonuniform Quantization

To further reduce the bit cost, we apply discrete nonuniform quantization in VGG-16 net. We use a group of numbers with uniform space to quantity the activation data. And the bit cost is to judge how many numbers we use each layer. To verify this method, we also have done some experiments, part of them is shown in Talbe. IV V.

As we see from the result above, the discrete nonuniform quantization save nearly half of the memory required compared with the methods above.

E. K-means clustering Quantization

In discrete nonuniform quantization, we apply a group of number with uniform space. To make it more suitable for the VGG-16 net, we also try to use K-means clustering methods [8] to pre-precess the activation data.

K-means clustering methods aims to partition n samples into k clusters where each sample belongs to the cluster with the nearest mean, serving as a prototype of the cluster. [9] applied K-means clustering methods in the quantization of weights of deep CNN. In this paper, we use K-means clustering methods in the nonuniform quantization in the activation data of VGG-16 net to improve the performance of discrete nonuniform quantization. The result of K-means clustering nonuniform quantization on VGG-16 net and Alex net is in Table. VI.

TABLE VI
K-MEANS CLUSTERING QUANTIZATION

	Without bound		with bound	
Quantization bit of VGG-16	4	5	4	5
Top5 Accuracy of VGG-16(%)	62.8	85.8	69.57	86.58
Quantization bit of Alex net	2	3	2	3
Top5 Accuracy of Alex net(%)	34.41	77.55	39.03	78.23

In the experiments of Table. VI, we apply the same quantization bit to all the convolutional layers. However, the K-means clustering method doesn't improve the performance of discrete nonuniform quantization greatly. In some cases, the accuracy under the same quantization bit is even less than before. In the experiment, we find out that the number computed by K-means method is much larger than the discrete nonuniform quantization with uniform space. This is because the large number contribute more to the classifier generally. So we do some pre-process to the data by adding a up bound on the data so that some large number won't do help to the K-means clustering. The experiment shows that K-means clustering method with bound has a better result than those without bound. We don't apply K-means clustering in full connected layers because the accuracy will drop sharply if we decrease the bit given in Table. II III.

IV. RESULT AND COMPARISON

In the last section, we apply several methods in the quantization of deep CNN. The comparison of their result can be seen in Table. VII. In Table. VII, we also show the memory reduction compared with method in [3].

TABLE VII
RESULT AND COMPARISON

Quantization method	Accuracy	Memory Saved	Methods in [3]
Uniform(%)	87.0	65.6	110.2
Nonuniform(%)	86.22	68.8	100
Discrete nonuniform(%)	86.25	82.4	56.4
K-means clustering(%)	86.58	84.4	50.0

V. CONCLUSION

In this paper, we apply several methods in the quantization of deep CNN. We compare different methods on the quantization of the VGG-16 net and Alex net and use only 5 bit on

TABLE IV
DISCRETE NONUNIFORM QUANTIZATION OF VGG-16 NET PART 1

Top5 Accuracy(%)	Index	conv1_1	conv1_2	conv2_1	conv2_2	conv3_1	conv3_2	conv3_3	conv4_1	conv4_2	conv4_3
56.60	1	8	4	5	5	5	5	5	5	5	5
79.43	2	8	4	5	5	5	5	5	5	5	5
83.46	3	8	4	5	5	5	5	5	5	5	5
85.81	4	8	5	6	4	6	6	6	6	6	6
86.19	5	8	5	6	5	5	4	6	6	6	6
86.25	6	8	5	6	5	5	5	5	5	5	5

TABLE V
DISCRETE NONUNIFORM QUANTIZATION OF VGG-16 NET PART 2

Top5 Accuracy(%)	Index	conv5_1	conv5_2	conv5_3	fc6	fc7	fc8
56.60	1	5	4	4	3	2	2
79.43	2	5	4	4	4	2	2
83.46	3	5	4	4	6	3	2
85.81	4	6	5	5	6	3	2
83.46	5	6	5	5	6	3	2
86.25	6	6	5	5	6	3	2

the quantization of VGG-16 net which save the memory cost with more than 6 times. The accuracy lost is less than 2%. To our best knowledge, we are the first to use nonuniform quantization and K-means clustering quantization on the deep CNN.

REFERENCES

[1] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[3] Darryl D Lin, Sachin S Talathi, and V Sreekanth Annapureddy, "Fixed point quantization of deep convolutional networks," *arXiv preprint arXiv:1511.06393*, 2015.

[4] Andrej Karpathy and Li Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.

[6] Jiajun Wu, Tianfan Xue, Joseph J Lim, Yuandong Tian, Joshua B Tenenbaum, Antonio Torralba, and William T Freeman, "Single image 3d interpreter network," *arXiv preprint arXiv:1604.08685*, 2016.

[7] Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," *arXiv preprint arXiv:1512.02595*, 2015.

[8] John A Hartigan and Manchek A Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[9] Song Han, Huizi Mao, and William J Dally, "Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding," *CoRR, abs/1510.00149*, vol. 2, 2015.