Poster Abstract: MicroBrain: Compressing Deep Neural Networks for Energy-efficient Visual Inference Service

Shiming Ge[1*], Zhao Luo[1,2], Qiting Ye[1,2], Xiao-Yu Zhang[1*]
[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing 10019, China.
Email: {geshiming, luochao, yeqiting, zhangxiaoyu}@iie.ac.cn

***Abstract*—The deployments of deep neural network models on mobile or embedded devices have been hindered due to their large number of weights. In this work, we develop a deep neural network (DNN) model compression service termed MicroBrain to reduce the resource usage for energy-efficient visual inference. By automatically analyzing the trained DNN models, we propose a high-performance DNN model compression approach to perform resource control via four modules. The proposed service, along with the compression approach, can provide 20-30x compression rate with negligible accuracy loss to condense DNN models, which facilitates their deployments on mobile devices for energy-efficient visual inference. We conduct an evaluation on two representative models, AlexNet and VGG-16, for object recognition and face verification tasks, which demonstrate the effectiveness of our proposed approach.**

I. Introduction

With the rapid development of modern computing power and big data techniques, deep neural networks (DNNs) have pushed artificial intelligence limits in a wide range of visual inference tasks. For example, visual recognition method [1] achieves 7.3% top-5 test error on the ImageNet LSVRL-2014 classification dataset [2], while face verification system [3] achieves almost 99% accuracy on face benchmark LFW [4]. Beyond the remarkable performance, there is increasing concern that the larger number of parameters consumes considerable resources (e.g., storage, memory and energy), which hinders their deployments on mobile or embedded devices. Therefore, it is necessary to compress the large models to extend their deployments.

Recently, some DNN model compression methods are proposed. Han *et al.*developed a method to prune unimportant connection and then retrain the weights to reduce storage and computation [5]. Gong *et al.*applied k-means clustering to the weights or conducting product quantization, and achieved 16-24x compression of the network with only 1% loss of accuracy on ImageNet classification task [6]. Wu *et al.*proposed Quantized CNN to simultaneously speedup the computation and reduce the storage and memory overhead of CNN models [7]. This method obtains 4-6x speedup and 15-20x compression with 1% loss of accuracy on ImageNet. Kim *et al.*used Tucker decomposition to achieve 5.46x and 1.09x compression for AlexNet and VGG-16 respectively [8].

*Shiming Ge and Xiao-Yu Zhang are the corresponding authors

In this paper, we proposed **MicroBrain**, a deep neural network compression service, to condense deep models for enabling energy-efficient visual inference. With a hybrid four-module approach, we provides a scalable model compression with small accuracy degradation. Our main contributions are three-folds. First, we design a scalable DNN model compression service to facilitate a range of visual inference tasks on mobile devices, which equips the interest-of-thing (IoT) devices with a smart "brain". Second, we propose a hybrid model compression approach with negligible accuracy degradation to adaptively control the resource usages, which facilitates energy-efficient visual inference tasks with DNNs. Finally, we apply two representative models to validate our approach and investigate the impact of DNN model compression on both the accuracy and resource usage of two visual inference tasks including object recognition and face verification.

II. Framework

Our framework is motivated by three observations: 1) The convolutional layers dominate most of the computations, while fully connected layers contain most of the weights. 2) DNNs are over-parameterized for training. 3) The memory bandwidth of a DNN model in inference greatly impact the energy consumption. As illustrated in Fig. 1, **MicroBrain** compresses the trained deep models to facilitate the deployment on various mobile devices via four major modules.

The approximation module. Different from existing works which focus on the convolutional layers and need to retrain or fine-tune, we adopt singular value decomposition (SVD)-based low rank algorithm to approximate only the fully connected layers. We check the construction accuracy on the final feature output to determine the optimal rank, which enables drastic reduction of the number of model weights without retraining.

The quantization module. We adopt dynamical fixed point algorithm to perform quantization. First, the dynamic range of the weights is analyzed to find a good fixed point representation. Then, a search is used to find the optimal bit numbers for convolutional weights, fully connected weights, and layer outputs by evaluating a trade-off between small number representation and the accuracy of construction.

978-1-5386-2784-6/17/$31.00 ©2017 IEEE

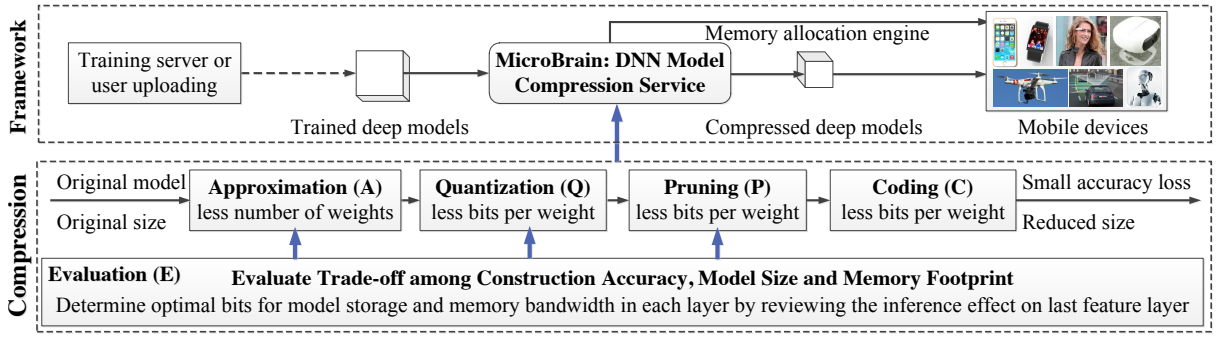

Fig. 1. Our scheme. Framework: **MicroBrain** carries out compression to the trained deep models and then deploys the compressed deep models on various mobile devices(*e.g.*, mobile phones, wearable devices, robots, etc). Compression: The hybrid deep compression algorithm consists of four major modules. Each module is performed by adaptively evaluating an optimal trade-off among inference accuracy, model representation and memory footprint.

TABLE I
RESOURCE USAGE BEFORE (B) AND AFTER (A) MODEL COMPRESSION.

Network	Data	Storage(b/a)	#MAC(b/a)	MAC ↓
AlexNet	ImageNet	241MB/11MB	1.14B/383M	33.6%
VGG-16	ImageNet	537MB/24MB	15.47B/5.83B	37.7%
VGG-16	LFW	553MB/17MB	15.48B/5.82B	37.6%

The pruning module. Considering most of the quantized weights are around zero-value, **MicroBrain** prunes small weights to give a sparse structure which can be stored with compressed sparse row or col in a index-value structure.

The coding module. Further, absolute index in the pruned sparse structure are replaced with relative index. Since the relative index are rarely above a threshold, we encode the weights and index with Huffman coding to compress further.

III. EVALUATION

To evaluate **MicroBrain**, we select two representative networks, AlexNet [9] and VGG-16 [1], for benchmarking. In the experiments, we evaluate the performance on compression rate and resource usage, and compared with other methods.

Compression performance. As shown in Tab.I, AlexNet on ImageNet can be compressed to 4.37% of its original size (241MB) with small accuracy loss (Top-5 accuracy of 78.6% compared to original 80.3%). For a larger network, VGG-16 on ImageNet, the model size can be compressed from 537MB to 24MB, while the accuracy loss is less than 1%. For VGG-16 on LFW [4] for face verification, **MicroBrain** condenses the model from 553MB to 17MB at the accuracy of 96.88% on LFW without embedding comparison, which achieves 32x compression rate with negligible accuracy loss (less than 0.4% compared to original 97.27% [3]).

Resource usage and energy efficiency. Beyond the compression performance, we also looked at the resource usage of the compressed model compared with its original model. As shown in Tab.I, over 95% of storage can be saved with our scheme, while the MAC operations are reduced about 2/3. As thus, **MicroBrain** can compress the DNN models to facilitate their deployment on mobile devices.

TABLE II
COMPRESSION PERFORMANCE COMPARISON WITH RECENT METHODS.

Method	Year	Idea	AlexNet	VGG-16
Gong [6]	2015	vector quantization	16-20x	–
Han [5]	2015	pruning	9x	13x
Wu [7]	2016	quantization	15-20x	–
Kim [8]	2016	Tucker decomposition	5.46x	1.09x
MicroBrain	–	hybrid	21.9x	22.4x

Comparison with other methods. As shown in Table. II, for AlexNet on ImageNet, quantization based methods [6] and [7] achieved 16-20x and 15-20x compression rate respectively. Han *et al.*. [5] used pruning to reduce the number of parameters of AlexNet by a factor of 9x and VGG-16 by 13x respectively, while Kim *et al.* [8] used Tucker decomposition to achieve 5.46x and 1.09x compression for AlexNet and VGG-16 respectively. Our **MicroBrain** achieves larger compression rate.

ACKNOWLEDGMENT

This work is supported in part by National Key Research and Development Plan (Grant 2016YFC0801005) and National Natural Science Foundation of China (Grant 61402463 and 61501457).

REFERENCES

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.
[2] O. Russakovsky, J. Deng, H. Su, and et al, "Imagenet large scale visual recognition challenge," *arXiv:1409.0575*.
[3] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, 2015.
[4] E. Miller, G. Huang, A. Chowdhury, and et al, "Labeled faces in the wild: A survey," *Advances in Face Detection and Facial Image Analysis*, 2016.
[5] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural networks," in *NIPS*, 2015.
[6] Y. Gong, L. Liu, M. Yang, and L. Bourdev, "Compressing deep convolutional networks using vector quantization," in *ICLR*, 2015.
[7] J. Wu, C. Leng, Y. Wang, and et al, "Quantized convolutional neural networks for mobile devices," in *IEEE CVPR*, 2016.
[8] Y. Kim, E. Park, S. Yoo, and et al, "Compression of deep convolutional neural networks for fast and low power mobile applications," in *ICLR*, 2016.
[9] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.