Towards Convolutional Neural Networks Compression via Global Error Reconstruction

通过全局误差重建对卷积神经网络压缩

Abstract

In recent years, convolutional neural networks (CNNs) have achieved remarkable success in various applications such as image classiﬁcation, object detection, object parsing and face alignment. Such CNN models are extremely powerful to deal with massive amounts of training data by using millions and billions of parameters. However, these models are typically deﬁcient（缺乏） due to the heavy cost in model storage, which prohibits their usage on resource-limited applications like mobile or embedded devices. In this paper, we target at compressing CNN models to an extreme without signiﬁcantly losing their discriminability. Our main idea is to explicitly（明确） model the output reconstruction error between the original and compressed CNNs, which error is minimized to pursuit（追求） a satisfactory rate-distortion after compression. In particular, a global error reconstruction method termed GER is presented, which ﬁrstly leverages an SVD-based low-rank approximation to coarsely compress the parameters in the fully connected layers in a layerwise manner. Subsequently, such layer-wise initial compressions are jointly optimized in a global perspective via back-propagation. The proposed GER method is evaluated on the ILSVRC2012 image classiﬁcation benchmark, with implementations on two widely-adopted convolutional neural networks, i.e., the AlexNet and VGGNet-19. Comparing to several state-of-the-art and alternative methods of CNN compression, the proposed scheme has demonstrated the best rate-distortion performance on both networks.

近年来，卷积神经网络（CNN）在诸如图像分类，对象检测，对象解析和人脸检测的各种应用中取得了显着的成功。这样的CNN模型极其强大，通过使用数百万和数十亿的参数来处理大量的训练数据。然而，由于模型存储的成本高昂，因此这些模型存在明显不足，这限制了它们在诸如移动或嵌入式设备的资源有限的应用上的使用。在本文中，我们的目标是将CNN模型压缩到极限，而不会显着地失去它们的可辨性。我们的主要思想是明确模型在原始和压缩CNN之间的输出重建误差，该误差被最小化以在压缩之后追求令人满意的速率失真。特别地，提出了称为GER的全局误差重建方法，其首先利用基于SVD的低秩近似来以分层方式粗略地压缩完全连接的层中的参数。随后，这种逐层初始压缩在全局视角中通过反向传播被联合优化。提出的GER方法是在ILSVRC2012图像分类基准上进行评估，在两个广泛采用的卷积神经网络，即AlexNet和VGGNet-19上已经实现。与CNN压缩的几种现有技术和替代方法相比，所提出的方案已经证明了在两个网络上的最佳速率失真性能。

1 Introduction

In recent years, convolutional neural networks have demonstrated impressive performance in various computer vision applications, for example, image classiﬁcation [A. Krizhevsky and Hinton, 2012; Y. Lecun and Haffner,1998;Simonyan and Zisserman,2014;C.Szegedy and Rabinovich, 2015; Zeiler and Fergus, 2014; Y. Jia and Darrell, 2014; K. He and Sun, 2015], object detection [R. Girshick and Malik, 2014; K. He and Sun, 2014] and image retrieval [Y. Gong and Lazebnik, 2014]. Despite the long history of neural network research in the literature [Fukushima, 1980], the signiﬁcant success of CNNs is mainly driven by the advanced computing resources available nowadays. For instance, to train a discriminative CNN model like AlexNet [A. Krizhevsky and Hinton, 2012] or VGGNet [Simonyan and Zisserman, 2014], it is typically required to set hundreds of millions of parameters, which are tuned using massive labelled or unlabelled data via approximated optimization ( e.g. stochastic gradient descent) through GPU or distributed settings [J. Deng and Li, 2009]. To that effect, various implementations of CNNs are introduced in the literature, e.g., AlexNet [A. Krizhevsky and Hinton, 2012], VGGNet [Simonyan and Zisserman, 2014], GoogleNet [C. Szegedy and Rabinovich, 2015] etc. Despite the state-of-the-art performances reported in challenging tasks like ImageNet ILSVRC [J. Deng and Li, 2009], the storage cost of CNNs is essentially huge, which typically require a large number of parameters (⇠ 108) [A. Krizhevsky and Hinton, 2012; Zeiler and Fergus, 2014; P. Sermanet and LeCun, 2013]. For instance, an 8-layer-AlexNet with 600,000 nodes costs 240MB storage, while a 19-layer-VGGNet with 1.5M nodes costs 548MB. Under such a circumstance, the existing CNNs cannot be directly deployed on devices that require compact storage,such as mobile phones or embedding devices. On the contrary, it is shown that CNNs with million-scale parameters typically tend to be heavily over-parameterized[M.Deniland Freitas, 2013]. Therefore, not all parameters and structures are essentially necessary in producing a discriminative CNN model. On the other hand, it is quantitatively shown in [Ba and Caruana,2014] that , neither shallown or simpliﬁed CNNs provide comparable performance to deep CNNs with million scale parameters. Therefore, a natural thought is to discover and discard the parameter redundancy in deep CNNs without signiﬁcantly decreasing the classiﬁcation accuracy.

近年来，卷积神经网络已经在各种计算机视觉应用中证明了令人印象深刻的性能，例如，图像分类[A. Krizhevsky和Hinton，2012; Y. Lecun和Haffner，Rabinovich，2015; Zeiler和Fergus，2014; Y. Jia和Darrell，2014; K. He and Sun，2015]，物体检测[R. Girshick and Malik，2014; K. He and Sun，2014]和图像检索[Y. Gong和Lazebnik，2014]。尽管文献中的神经网络研究有着悠久的历史[Fukushima，1980]，但CNN的显着成功主要是由现在可用的高级计算资源驱动的。例如，训练一个有区别的CNN模型像AlexNet [A. Krizhevsky和Hinton，2012]或VGGNet [Simonyan和Zisserman，2014]，通常需要设置数亿个参数，这些参数通过GPU或分布式的近似优化（例如随机梯度下降）设置[J.邓和李，2009年]。为此，CNN的各种实施方案引入了本文，例如AlexNet [A. Krizhevsky和Hinton，2012]，VGGNet [Simonyan和Zisserman，2014]，GoogleNet [C. Szegedy和Rabinovich，2015]等。尽管在像ImageNet ILSVRC这样的具有挑战性的任务中报告了最先进的性能。 Deng和Li，2009]，CNN的存储成本基本上很大，这通常需要大量的参数（⇠108）[A. Krizhevsky和Hinton，2012; Zeiler和Fergus，2014; P. Sermanet和LeCun，2013]。例如，具有600,000个节点的8层AlexNet的存储成本为240MB，而具有1.5M节点的19层VGGNet的成本为548MB。在这种情况下，现有的CNN不能直接部署在需要紧凑存储的设备上，例如移动电话设备。相反，显然，CNN的数学规模参数通常往往被超参数化[M.Deniland Freitas，2013]。因此，并非所有参数和结构在产生有区别的CNN模型中是必需的。另一方面，在[Ba和Cuanuana，2014]中定量地显示，无论是否提供的CNN都提供了相应的性能等级参数。因此，一个自然的想法是发现和抛弃参数的冗余，而不显着降低分类的准确性

The compression of CNNs has attracted a few research attentions very recently, which can be further categorized into three groups, i.e., parameter sharing, parameter pruning and matrix decomposition.

As for parameter sharing, Gong et al. [Y. Gong and Bourdev, 2014] employed vector quantization over parameters to reduce the redundancy in the parameter space. Chen et al. [W. Chen and Chen, 2015] proposed a HashedNets model which uses a low-cost hash function to group weights between two connected layers into hash buckets to share parameters. Cheng et al. [Y. Cheng and Chang, 2015] proposed to replace the conventional linear projection in fully connected layers with the circulant projection, which reduces the storage cost and enables the use of Fast Fourier Transform to accelerate the computation. As for parameter pruning, Srinivas and Babu [Srinivas and Babu, 2015] explored the redundancy among neurons, and proposed a data-free pruning to remove redundant neurons. Han et al. [S. Han and Dally, 2015] aimed at reducing the total amount of parameters and operations in the entire network. The above pruning approaches can give signiﬁcant reductions in both parameter size and computation workload. As for matrix decomposition , Denilet al. [M.DenilandFreitas, 2013] adopted low-rank decomposition to compress the weights in the fully connected layers in a layer-by-layer manner. Novikov et al. [A. Novikov and Vetrov, 2015] converted the dense weight matrices of the fully connected layers to the Tensor Train format,such that the number of parameters is reduced by huge factor while preserving the expressive power ofthelayer.

CNN的压缩最近吸引了一些研究关注，其可以进一步分为三组，即参数共享，参数修剪和矩阵分解。

至于参数共享，Gong et al。 [Y. Gong和Bourdev，2014]对参数使用矢量量化以减少参数空间中的冗余。 Chen et al。 [W. Chen和Chen，2015]提出了一个HashedNets模型，它使用低成本的散列函数将两个连接的层之间的权重分组为散列桶以共享参数。 Cheng et al。 [Y. Cheng and Chang，2015]提出在全连接层中用循环投影代替传统的线性投影，这降低了存储成本，并且能够使用快速傅立叶变换来加速计算。至于参数修剪，Srinivas和Babu [Srinivas和Babu，2015]探索神经元之间的冗余，并提出了一个数据自由精简删除冗余的神经元。 Han et al。 [S. Han and Dally，2015]，旨在减少整个网络中的参数和操作的总量。上述修剪方法可以显着减少参数大小和计算工作量。至于矩阵分解，Denilet al。 [M.DenilandFreitas，2013]采用低秩分解以逐层方式压缩全连接层中的权重。 Novikov et al。 [一个。 Novikov和Vetrov，2015]将完全连接的层的密集权重矩阵转换为Tensor Train格式，使得参数的数量通过巨大因子减少，同时保留了后台的表达能力。

However, the state-of-the-art methods [M. Denil and Freitas, 2013; Y. Gong and Bourdev, 2014; Srinivas and Babu, 2015]still rely on a layer-wise paramete rcompression,which do not provide an explicit modeling to the overall loss of classiﬁcation accuracy. In other words, such works can be regraded as a layer-wised, “implicit” and “local” compression for CNNs. In terms of the “implicit” compression, the existing works only considered approximating the parameters W between fully connected layers with ˆ W by minimizing their Euclidean distance W ˆ W 2 F . This setting is indeed problematic, which does not directly recover the output of CNNs (a.k.a., the learned deep features) used for classiﬁcation.

In terms of “local” compression, a more optimal solution is indeed to preserve the classiﬁcation accuracy in a global manner, i.e., compressing all parameters across the entire fully connected layers. Meanwhile, the global correlations among weights of inter-layers are simply ignore

in [M. Denil and Freitas, 2013; Y. Gong and Bourdev, 2014; Srinivas and Babu, 2015]. In particular, due to the nonlinear activation functions (e.g. sigmoid, tahn [Y. LeCun and M¨uller, 2012], or rectiﬁer linear unit (ReLU) [Nair and Hinton, 2010]), small quantization error between W and ˆ W in each layer might magniﬁed and propagated in the network, leading to large generalization error as shown in our experiments.

然而，最先进的方法[M. Denil和Freitas，2013; Y. Gong和Bourdev，2014; Srinivas和Babu，2015]仍然依赖于层次的参数压缩，这不能为分类精度的整体损失提供明确的建模。换句话说，这样的工作可以被重新分级为用于CNN的层次化的“隐式”和“本地”压缩。就“隐式”压缩而言，现有工作仅考虑通过最小化其欧几里得距离W W 2 F来近似W的完全连接层之间的参数W.这个设置确实是有问题的，它不直接恢复用于分类的CNN的输出（也称为学习的深层特征）。

 在“局部”压缩方面，更优化的解决方案确实是以全局方式保持分类精度，即，在整个完全连接的层上压缩所有参数。同时，层间权重之间的全局相关性被简单地忽略

 在[M. Denil和Freitas，2013; Y. Gong和Bourdev，2014; Srinivas和Babu，2015]。特别地，由于非线性激活函数（例如Sigmoid，Tahn [Y.LeCun和M¨uller，2012]或整流器线性单元（ReLU）[Nair和Hinton，2010]），W和W之间的小量化误差在每个层可能放大和传播在网络中，导致大的泛化误差，如我们的实验所示。

In practice, we conduct an initial compression of the weights in the fully connected layers by an SVD-based lowrank decomposition, which can relax the constrained term of the optimization function to being trackable. Subsequently, such layer-wise（逐层）, coarse（粗） compressions are further（进一步） jointly optimized among layers via back propagation to minimize the global error, which is done by a novel optimization method that using the stochastic gradient descent to well solve the non-convex optimization problem.

The proposed method is evaluated on the ILSVRC2012 image classiﬁcation benchmark, with testing on two widely adopted CNNs, i.e., AlexNet [A. Krizhevsky and Hinton, 2012] and VGGNet-19 [Simonyan and Zisserman, 2014]. We have demonstrated that the proposed GER compression scheme can lead to the state-of-the-art rate-distortion performance by comparing with several state-of-the-art and alternative schemes in CNN compression [M. Denil and Freitas, 2013;Y.GongandBourdev,2014;X.ZhangandSun,2015]. The main contribution of this paper is three-fold:

在实践中，我们通过基于SVD的低分辨率分解对完全连接的层中的权重进行初始压缩，这可以减轻已有优化函数的约束项。随后，通过反向传播在层之间进一步联合优化这些层次的粗略压缩，以最小化全局误差，其通过使用随机梯度下降以良好地解决非凸优化问题的新颖的优化方法来完成。

所提出的方法在ILSVRC2012图像分类基准上进行评估，对两个广泛采用的CNN进行测试，即AlexNet [A. Krizhevsky和Hinton，2012]和VGGNet-19 [Simonyan和Zisserman，2014]。我们已经证明，所提出的GER压缩方案可以通过与CNN压缩中的几种现有技术和替代方案进行比较来导致现有技术的速率 - 失真性能[M. Denil和Freitas，2013; Y.GongandBourdev，2014; X.ZhangandSun，2015]。本文的主要贡献是三重：

We introduce an explicit objective function to directly minimize the reconstruction error of outputs before and after network compression, which differs from the existing works that indirectly minimize the difference between the original and compressed parameters

We globally model the inter-layer correlation during the network compression, which can address the issue of layer-wise accumulated compression error.

We introduce an effective optimization method to solve the corresponding non-convex optimization, which ﬁrstly uses an SVD-based low-rank decomposition to relax the constrained term, and then adopts a stochastic gradient descent to learn optimal parameters.

我们引入明确的目标函数直接最小化网络压缩之前和之后的输出的重建误差，这不同于间接最小化原始和压缩参数之间的差异的现有工作

我们在网络压缩期间对层间相关性进行全局建模，这可以解决逐层累积压缩误差的问题。

我们引入了一种有效的优化方法来求解相应的非凸优化，首先使用基于SVD的低秩分解来放宽约束项，然后采用随机梯度下降来学习最优参数。

2. Initial Compression of CNN Based on Low-rank Decomposition

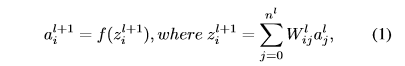
基于低秩分解的CNN的初始压缩

2.1 初步

定义特征矩阵X作为压缩全连接CNN输入（维度d，特征向量数量n）n是在原始CNN样本中（CNN-like AlexNet）的最后卷积层输出

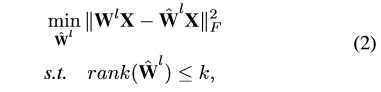
参考研究是2012年A. Krizhevsky和Hinton

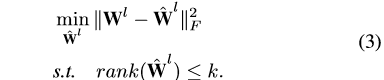
压缩链接的CNN第l层的向前传播的公式是:



2.2 逐层低轶近似的线性响应

首先考虑第l层和第l+1层之间的原始权重w的近似low-rank，为了找到近似的低轶子空间，最小化以下响应的重建误差:



对于相同输入信号X上的两个线性变换的误差，可以重写：：  


通过奇异向量分解（SVD）解决[Golub和Loan，2012]

算法1 逐层交替优化

输入：训练数据点X;在非线性变换之前的第1层和第1 + 1层之间的输出Z1;第1层和第1 + 1层之间的输入C1;限制秩k，最大值T.

输出：近似低秩矩阵W1。

1：初始化C0 byX，辅助变量Y1 byZ1，并且t = 1。

2：重复

3：步骤1：通过GSVD固定Y1 + 1和更新W1

4：步骤II：通过分析y1 + 1 ij修正W1和更新Y1 + 1，并求解1维最优化Eq。 9viaEq。 10和Eq。 11。

5：t：= t + 1;

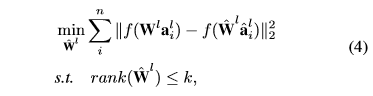
6：直到会聚或达到最大最大值T

2.3 扩展到非线性响应

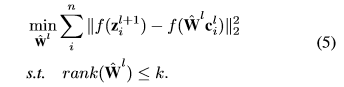
For non-linear transition that commonly occurs in CNNs, the result of approximated matrix is not equivalent to that of the original one. Therefore, the effect of non-linear transition should be taken into account when designing low-rank approximation of parameter matrix W. Taking the widely used ReLU transition for instance,

对于通常在CNN中出现的非线性转变，近似矩阵的结果不等于原始结果。 因此，在设计参数矩阵W的低阶近似时，应该考虑非线性跃迁的效应。以广泛使用的ReLU跃迁为例，

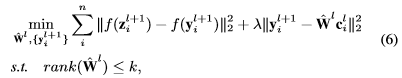
为了最小化ReLUlike响应的重建误差，我们有



方程4的优化问题可以通过使用交替求解器的分层优化来求解，更具体的是，这种优化层到层是自下向上。重写：



由于非线性和低轶约束的共存而难以解决，为获得可行解，修改为：



采用交替求解器求解方程6，下面详细的说明这种替代优化：

参考[GowerandDijksterhuis,2004;TakaneandHwang,2007] 减少秩回归问题

参考[Takane and Jung，2006]广义奇异向量分解（GSVD）

3. 全局误差重建进行跨层压缩

The initial compression of CNN using low-rank decomposition is to coarsely approximate ˆ Wl of each layer in a bottom up manner. As discussed above, the compression errors are accumulatedlayer-by-layer,resultingtolargeoverallerrorin the output layer of CNN. To address this issue, the proposed global error reconstruction (GER) targets at jointly optimizing among layers as shown in Fig. 2.

In particular, if the original CNN model has m fully connected layers, we minimize the global reconstruction error of ReLU-likenon-linear responses as follows:

使用低秩分解的CNN的初始压缩是以自底向上的方式粗略近似每层的W。 如上所述，压缩误差是逐层累积的，从而导致CNN的输出层中的全部误差。 为了解决这个问题，提出的全局误差重建（GER）目标在层之间联合优化，如图2所示。

特别地，如果原始CNN模型具有m个完全连接的层，我们将ReLU-likenon线性响应的全局重建误差最小化如下：

4 Experiments

To evaluate the performance of GER scheme, we conduct comprehensive experiments on ILSVRC2012 image classiﬁcation benchmark. We deploy the proposed GER on two widely used CNNs (a.k.a. AlexNet [A. Krizhevsky and Hinton, 2012] and VGGNet-19 [Simonyan and Zisserman, 2014]), with comparisons to the state-of-the-art scheme proposed very recently [M.Denil and Freitas, 2013; Y.Gong and Bourdev, 2014; X.Zhang and Sun,2015

实验

为了评估GER方案的性能，我们对ILSVRC2012图像分类基准进行了全面实验。 我们在两个广泛使用的CNN（也称为AlexNet [A.Krizhevsky和Hinton，2012]和VGGNet-19 [Simonyan和Zisserman，2014]）上部署了提议的GER，与最近提出的最先进的计划 [M.Denil和Freitas，2013; Y.Gong和Bourdev，2014; X.Zhang和Sun，2015

4.1 Experimental Setting

Dataset.

We test the proposed GER based CNN compression on the ILSVRC2012 image classiﬁcation benchmark. The dataset contains more than 1 million training images from 1,000 object classes. It also has a validation set of 50,000 images, where each object class contains 50 images. We randomly select 100,000 images (100 from each class) from the training set for training, and test on the validation set.

Implementation Details.

We implement GER on two CNNs i.e., AlexNet [A. Krizhevsky and Hinton, 2012] and VGGNet-19 [Simonyan and Zisserman, 2014]. The VGGNet-19 contains 16 convolutional layers and 3 fully connected layers, and the AlexNet contains 5 convolutional layers and 3 fully connected layers. The compressed networks are trained using Caffe [Y. Jia and Darrell, 2014] and run on NVIDIA GTX TITAN X graphics card with 12GB. The learning rate starts at 0.01 and is halved every 10-epochs; the weight decay is set to 0.0005 and the momentum is set to 0.9.

Baselines.

We compare the proposed GER scheme with 4 state-of-the-art approaches published very recently, including PQ-based compression (PQ) [Y. Gong and Bourdev, 2014],Low-rankdecomposition(LRD)[M.DenilandFreitas, 2013], Layer-wise optimization by alternating solver (AS) [X. Zhang and Sun, 2015], and Binary-based compression (BIN) [Y. Gong and Bourdev, 2014]. As for the alternative approach, we compare GER with GER-IC, which is the SVD based initial compression based only on Sec.2.

Evaluation protocol.

The classiﬁcation error on the validation set was employed as the evaluation protocol. We used both the top-1 classiﬁcation error and the top-5 classiﬁcation error to evaluate different compression methods. Then we measure the compression performance in terms of the rate distortion, which reﬂects the balance between compression rate and classiﬁcation error.

4.1实验设置

数据集。

我们在ILSVRC2012图像分类基准测试所提出的基于GER的CNN压缩。该数据集包含来自1,000个对象类的超过100万个训练图像。它还具有50,000个图像的验证集，其中每个对象类包含50个图像。我们从用于训练的训练集中随机选择100,000个图像（来自每个类别的100个），并且对验证集进行测试。

实施细节。

我们在两个CNN上实现GER，即AlexNet [A. Krizhevsky和Hinton，2012]和VGGNet-19 [Simonyan和Zisserman，2014]。 VGGNet-19包含16个卷积层和3个完全连接的层，AlexNet包含5个卷积层和3个完全连接的层。压缩网络使用Caffe [Y. Jia和Darrell，2014]，运行NVIDIA GTX TITAN X显卡12GB。学习率从0.01开始，每10个周期减半;重量衰减设置为0.0005，动量设置为0.9。

基线。

我们将所提出的GER方案与最近公布的4种最先进的方法进行比较，包括基于PQ的压缩（PQ）[Y. Gong和Bourdev，2014]，低秩分解（LRD）[M.DenilandFreitas，2013]，通过交替求解器（AS）的层次优化[X. Zhang和Sun，2015]和二进制压缩（BIN）[Y. Gong和Bourdev，2014]。至于替代方法，我们比较GER与GER-IC，这是基于SVD的初始压缩只基于Sec.2。

 评估方案。

使用验证集上的分类误差作为评价方案。我们使用第1分类错误和第5分类错误来评估不同的压缩方法。然后我们测量压缩性能的速率失真，反映了压缩率和分类误差之间的平衡。

Rate-Distortion Comparison.

We use a different rank k from 25 to 210 to achieve different compression rates. For PQ, we ﬁx the number of centers to 256 (8 bits) and vary the segment dimension s =1 ,2,4,8. For LRD and Layer-wise optimization by alternating solver, we use the same compression criterion with GER that varies k from25 to210. For BIN, which has no parameter to tune,the compression rate is ﬁxed as 32.

Both the top-1 and top-5 classiﬁcation errors are shown in Fig. 3, which show a consistent trend in rate-distortion. As for intra-layer approximation, GER-IC achieves a similar classiﬁcation error to LRD. Instead, by explicitly modeling the reconstruction error in a global way, GER greatly improves rate-distortion by comparing to GER-IC. GER also performs much better than LRD and AS for compressing the fully connected layers. To explain, GER merits in its “explicit” compression, which effectively combines the initial layer-wise compression and cross-layer global compression, while LRD and AS are “implicit” compression which only consider the local intra-layer relationship. Note that PQ has achieved better performance comparing to that of LRD and AS. However, it is shown in Fig. 3 that PQ is hard to achieve high compression rate, which might be due to the limited codebook size. In contrast, GER achieves the best rate-distortion by comparing to other baselines. Finally, as discovered by Gong et al. [Y. Gong and Bourdev, 2014], the simplest BIN works well when we ﬁxed the compression rate to 32. Therefore, The base binary quantization is also a good choice when the goal is to compress data very aggressively. However, it is hard to be adaptive when we want to control the compression rate, which in turn is one of the key advantages for our scheme. We further show that the classiﬁcation error with a ﬁxed compression rate in Tab.1 which also shows GER can achieve the best performance by comparing to other baselines especially, when implementing on VGGNet-19.

On the Single Layer Error.

we further analyze the classiﬁcation error by compressing each single layer while ﬁxing the rest layers as the original uncompressed version. The results are reported in Fig. 4. We found using all the baselines to compress the ﬁrst two fully connected layers (i.e. “FC6” and “FC7”) does not lead to the decreasing of accuracy. In contrast, compressing the last fully connected layer leads to large classiﬁcation error for all the baselines, except for the proposed GER. We explain this advantage by the fact that GER can automatically adjust the inter-layer error by tuning and reﬁning across all layers.

速率失真比较。

我们使用从25到210的不同秩k来实现不同的压缩率。对于PQ，我们将中心数量设置为256（8位），并且改变段尺寸s = 1,2,4,8。对于通过交替求解器的LRD和层次优化，我们使用与从25到210变化k的GER相同的压缩标准。对于没有调整参数的BIN，压缩率固定为32。

第1和第5分类误差都显示在图3中。它们显示了速率失真的一致趋势。对于层内近似，GER-IC实现与LRD类似的分类误差。相反，通过以全局方式明确地建模重建误差，GER与GER-IC相比， 大大地改善了速率失真。 GER还执行比LRD和AS好得多以压缩完全连接的层。为了解释，GER的“显式”压缩有效地结合了初始的逐层压缩和跨层全局压缩，而LRD和AS是“隐式”压缩，只考虑局部层内关系。注意，与LRD和AS相比，PQ获得了更好的性能。然而，如图1所示。 3，PQ难以实现高压缩率，这可能是由于码书大小有限。相比之下，GER通过与其他基线相比，实现了最佳的速率失真。最后，如Gong et al。 [Y. Gong和Bourdev，2014]，当我们将压缩率固定为32时，最简单的BIN效果很好。因此，当目标是非常积极地压缩数据时，基本二进制量化也是一个不错的选择。然而，当我们想要控制压缩率时，这是难以适应的，这又是我们的方案的关键优点之一。我们进一步表明，在Tab.1中固定压缩率的分类误差，也显示GER可以通过与其他基线相比，特别是在VGGNet-19上实现时获得最佳性能。

单层错误。

我们进一步分析分类错误通过压缩每个单一层，同时固定其余层作为原始的未压缩版本。结果报告于图1中。 4.我们发现使用所有基线压缩第一个两个完全连接的层（即“FC6”和“FC7”）不会导致精度的降低。相比之下，压缩最后完全连接层导致所有基线的大分类误差，除了建议的GER。我们解释这个优势的事实，GER可以通过调整和所有层之间的修改自动调整层间误差。

5 Conclusion

In this paper, we propose to compress the convolutional neural networks to reduce the model storage by a novel Global Error Reconstruction scheme, which can facilitate emerging applications in mobile or embedding devices with limited storage. GER ﬁrstly uses an SVD-based low-rank approximations to coarsely compress the parameters in the fully connected layers. Such layer-wise initial compressions are further jointly optimized among layers in a global way via backpropagation. Unlike previous approaches that only considered recovering the internal weight parameters,GER also explicitly models the reconstruction error between the outputs of both the original and compressed CNNs, which we significantly reduces the accumulated reconstruction error caused by the nonlinear activation. We have demonstrated that the proposed GER scheme can lead to state-of-the-art ratedistortion performance by comparing to several very recent schemes in CNN compression [M. Denil and Freitas, 2013; Y. Gong and Bourdev, 2014; X. Zhang and Sun,2015]. In our future work, we would extend and evaluate our compression scheme from the fully connected layer to the convolutional layer, and at the same time, further accelerate computation of convolutional layers.

5结论

在本文中，我们提出通过一个新颖的全局错误重建方案来压缩卷积神经网络以减少模型存储，这可以促进在有限内存下移动或嵌入设备的新兴应用程序。 GER首先使用基于SVD的低秩近似来粗略地压缩完全连接的层中的参数。这种逐层初始压缩进一步通过反向传播以全局方式在层之间联合优化。与以前仅考虑恢复内部权重参数的方法不同，GER还明确地建模了原始和压缩CNN的输出之间的重建误差，这显着地减少了由非线性激活引起的累积重建误差。我们已经证明，提出的GER方案可以通过与CNN压缩中的几个最近的方案相比较而导致最先进的额定失真性能[M. Denil和Freitas，2013; Y. Gong和Bourdev，2014; X. Zhang and Sun，2015]。在我们未来的工作中，我们将扩展和评估我们的压缩方案从完全连接层到卷积层，同时进一步加速卷积层的计算。