## 文章标题

Towards Convolutional Neural Networks Compression via Global Error Reconstruction

## 文章总结

卷积神经网络（CNN）模型强大，使用数百万和数十亿参数来处理大量的训练数据。这样出现的问题是模型存储成本高昂，本文目标是在保证不会显著地失去CNN模型的可辨性下，将CNN模型压缩到极限。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]。

## 文中综述的国内外情况

1. 训练一个有区别的CNN模型AlexNet，参考的是A. Krizhevsky和Hinton，2012

VGGNet模型 [Simonyan和Zisserman，2014]

GoogleNet [C. Szegedy和Rabinovich，2015]

2. CNN压缩分为三组：参数共享，参数修剪和矩阵分解

（1）参数共享

Gong et al. [Y. Gong and Bourdev, 2014]对参数使用矢量化以减少参数空间中的冗余

Chen et al. [W. Chen and Chen, 2015] 提出了一个HashedNets模型，使用低成本散列函数将两个连接的层之间的权重分组为散列块共享参数

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

（2）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。

（3）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. [A. Novikov and Vetrov, 2015] 将完全连接的层的密集权重矩阵转换为Tensor Train格式

## 使用的方法

提出GER全局误差重建方法

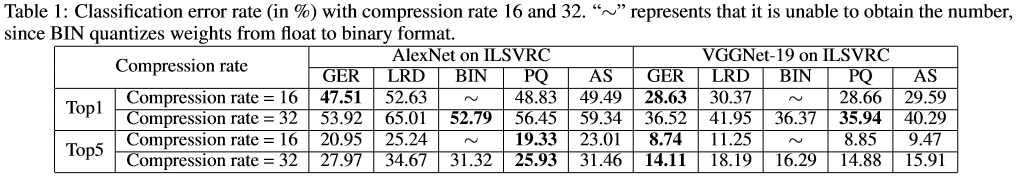
1）利用基于SVD的低轶近似来以分层方式粗略地压缩完全连接的层中参数

2）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

## 实验及输出结果

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。

结果报告于图1中



## 使用的数据集

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。

结论

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]

## 小编总结

本文主要是压缩优化参数，通过GER方案通过对CNN压缩几个最近方案比较，但是只有实验表1给出的数据来说明实验的情况，对压缩内存有一定的借鉴，对于实验室项目来说暂时没有很大的联系。