## 文章标题

Compressing Convolutional Neural Networks in the Frequency Domain

## 文章总结

**待解决的问题**：CNN模型尺寸增大对分类器的存储和存储器需求增大，如何进行压缩来节省内存

**解决方案**：提出一个新的网络架构—频率敏感散列网络（FreshNets），利用卷积层和深度学习模型之间完全连接层的固有冗余，来节省内存和存储消耗。

**具体内容**：首先利用离散余弦变换（DCT）将滤波器权重转换到频域，并使用低成本哈希函数来随机地将频率参数分组成哈希桶，分配了相同散列桶的所有参数共享使用标准反向传播学习的单个值。为了进一步减小模型大小，将较少的哈希桶分配给高频分量

**结果**：评估FreshNets上的八个数据集，并表明它导致比几个相关基线更好的压缩性能。

## 文中综述的国内外情况

1. 深层神经网络中学习参数有重要冗余，参考M. Denil, B. Shakibi, L. Dinh, N. de Freitas, et al. Predicting parameters in deep learning. In Advances in Neural。研究的是在分解成两个低轶矩阵之后，在完全连接的层中学习参数

2. 通过学习一级基本过滤器来实现进一步的加速，参考E. L. Denton, W. Zaremba, J. Bruna, Y. LeCun, and R. Fergus. Exploiting linear structure within convolutional networks for efﬁcient evaluation. In Advances in Neural Information Processing Systems, pages 1269–1277, 2014。主张将滤波器权重的四维张量分解为不同一阶四维张量的总和，并显示一些令人鼓舞的结果。

用于加速卷积神经网络评估的另一种技术是在傅里叶频率中计算卷积

3.参考[38] M. Mathieu, M. Henaff, and Y. LeCun. Fast training of convolutional networks through ffts. arXiv preprint arXiv:1312.5851, 2013。对频率进行过采样将滤波器转换到大小为nxn的频域。

4. Y. Gong, L. Liu, M. Yang, and L. Bourdev. Compressing deep convolutional networks using vector quantization. arXiv preprint arXiv:1412.6115, 2014.，对量化值应用kmean聚类

5. S. Han, H. Mao, and W. J. Dally. A deep neural network compression pipeline: Pruning, quantization, huffman encoding. arXiv preprint arXiv:1510.00149, 2015.

递归地训练神经网络并基于它们的权重大小修剪不重要的连接。

其他集中在矩阵或张量分解压缩

6. Z. Yang, M. Moczulski, M. Denil, N. de Freitas, A. Smola, L. Song, and Z. Wang. Deep fried convnets. arXiv preprint arXiv:1412.7149, 2014.采用快速变换

7. Y.-D. Kim, E. Park, S. Yoo, T. Choi, L. Yang, and D. Shin. Compression of deep convolutional neural networks for fast and low power mobile applications. arXiv preprint arXiv:1511.06530, 2015 一般张量分解和秩选择压缩整个网络

8. O. Rippel, J. Snoek, and R. P. Adams. Spectral representations for convolutional neural networks. In Advances in Neural Information Processing Systems, pages 2440–2448, 2015.

提出在频域中学习更快的收敛学习的权重参数。

## 使用的方法--频率敏感散列网络（FreshNets）

文章第二节讲述哈希和离散余弦变换原理内容

测试期间为任何现有的CNN代码添加一个低成本哈希函数和逆DCT变换，用于滤波器的重建；通过简单的反向传播学习散列权重

概述FreshNets培训程序，包括初始化阶段，前馈阶段和反向传播阶段，所有训练操作可以通过修改反向传播的计算程序在现有的CNN包中实现。

## 实验及输出结果

所有数据集采用表1所示神经网络架构

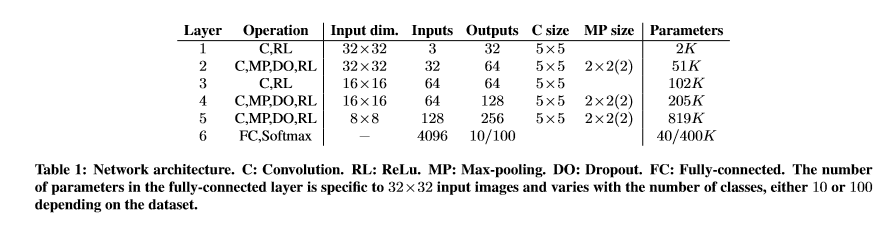
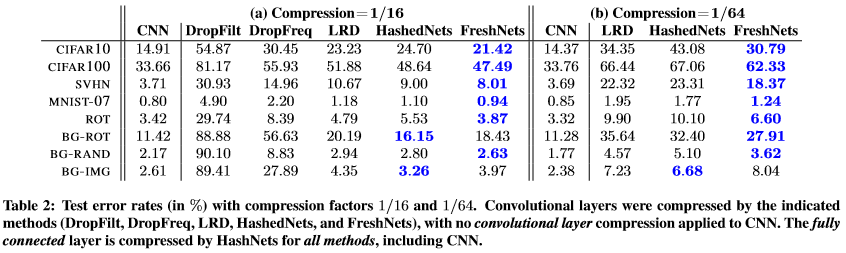
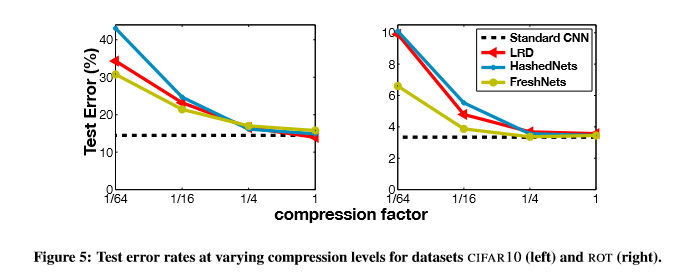


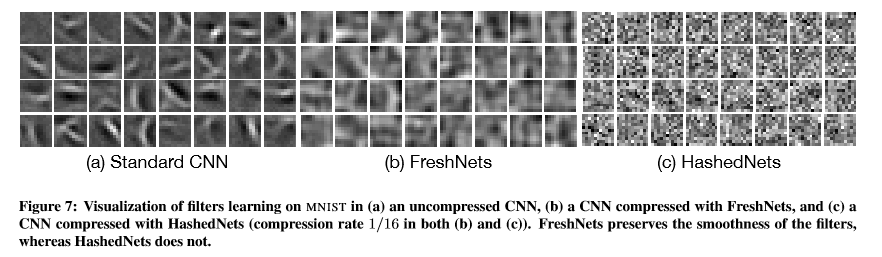
表2（a）和（b）分别示出了在压缩比为1/16和1/64时所有方法的综合评价



在图5的插入条形图中，我们报告了由频率忽略方案的测试误差归一化的测试误差



通过可视化（a）一个标准的，未压缩的CNN，（b）FreshNets和（c）HashedNets（在空间域中的权重共享）的滤波器权重（第一层）来研究图7中学习的卷积滤波器的平滑度。



## 使用的数据集

我们对八个基准数据集进行实验：CIFAR10 [29]，CIFAR100 [29]，SVHN [39]和五个具有挑战性的MNIST [31,34,44]变体。 CIFAR10数据集包含60000个32×32像素的图像，具有三个颜色通道。从十个类中选择图像，每个类由6000个唯一实例组成。

CIFAR100数据集还包含60000个32×32图像，但是更具挑战性，因为图像是从100个类中选择的（每个类有600个图像）。对于两个CIFAR数据集，指定50000个图像用于训练，剩余10000个图像用于测试。为了提高CIFAR100的精度，我们增加水平反射和裁剪，导致0.8M的训练图像。

SVHN数据集是从真实世界场景裁剪的数字（10类）的大集合，包括73257个训练图像，26032测试图像和531131更少的难看的图像用于额外的训练。

在我们的实验中，我们使用所有可用的训练图像，总共604388个训练样本。对于MNIST变体，每个变体通过旋转（ROT），背景重叠（BGRAND和BG-IMG）或其组合（BG-ROT）减少初始化（MNIST-07）我们预处理所有带有白化的数据集（除了CIFAR100和SVHN，这是非常大的）。

## 结论

在本文中提出FreshNets，一种学习卷积神经网络与显着压缩模型存储的方法。我们引入可忽略的效率开销（因为大多数运行时间仍然用于卷积运算），但获得一个显着更小的卷积神经网络。利用散列技巧无参数随机权重共享和利用卷积滤波器固有的平滑性，FreshNets以频率敏感的方式压缩参数，使得重要的模型参数（例如低频分量）被更好地保留。因此，FreshNets在高压缩率下保留预测精度比竞争基线更好。

## 小编总结

本文是对图片压缩进行的CNN压缩，采用哈希和离散余弦变换原理，在空间和频率入手，通过滤波器来学习研究卷积滤波器的平滑度。

但是和实验室项目研究内容不符合。