A Survey of Model Compression and Acceleration for Deep Neural Networks

深度神经网络模型压缩加速度研究综述

日期：2019-02-07

作者：Yu Cheng (http://tongtianta.site/search?search_txt=Yu%20Cheng) 、Duo Wang (http://tongtianta.site/search?search_txt=Duo%20Wang) 、Pan Zhou (http://tongtianta.site/search?search_txt=Pan%20Zhou) 、Tao Zhang (http://tongtianta.site/search?search_txt=Tao%20Zhang)

论文：http://arxiv.org/pdf/1710.09282v7.pdf (http://arxiv.org/pdf/1710.09282v7.pdf)

Yu Cheng, Duo Wang, Pan Zhou, Member, IEEE, and Tao Zhang, Senior Member, IEEE

Yu Cheng，Duo Wang，Pan Zhou，IEEE会员，张涛，IEEE高级会员

Abstract—Deep convolutional neural networks (CNNs) have recently achieved great success in many visual recognition tasks. However, existing deep neural network models are computationally expensive and memory intensive, hindering their deployment in devices with low memory resources or in applications with strict latency requirements. Therefore, a natural thought is to perform model compression and acceleration in deep networks without significantly decreasing the model performance. During the past few years, tremendous progress has been made in this area. In this paper, we survey the recent advanced techniques for compacting and accelerating CNNs model developed. These techniques are roughly categorized into four schemes: parameter pruning and sharing, low-rank factorization, transferred/compact convolutional filters, and knowledge distillation. Methods of parameter pruning and sharing will be described at the beginning, after that the other techniques will be introduced. For each scheme, we provide insightful analysis regarding the performance, related applications, advantages, and drawbacks etc. Then we will go through a few very recent additional successful methods, for example, dynamic capacity networks and stochastic depths networks. After that, we survey the evaluation matrix, the main datasets used for evaluating the model performance and recent benchmarking efforts. Finally, we conclude this paper, discuss remaining challenges and possible directions on this topic.

摘要 - 深度卷积神经网络（CNNs）最近在许多视觉识别任务中取得了巨大成功。然而，现有的深度神经网络模型在计算上是昂贵且存储器密集的，阻碍了它们在具有低内存资源的设备中或在具有严格延迟要求的应用中的部署。因此，一种自然的想法是在深度网络中执行模型压缩和加速，而不会显着降低模型性能。在过去几年中，这一领域取得了巨大进展。在本文中，我们调查了最近开发的压缩和加速CNN模型的先进技术。这些技术大致分为四种方案：参数修剪和共享，低秩分解，转移/紧凑卷积滤波器和知识蒸馏。将在开始时描述参数修剪和共享的方法，之后将引入其他技术。对于每个方案，我们提供有关性能，相关应用，优势和缺点等的深入分析。然后我们将介绍一些最近的其他成功方法，例如动态容量网络和随机深度网络。之后，我们调查评估矩阵，用于评估模型性能的主要数据集和最近的基准测试工作。最后，我们总结本文，讨论有关该主题的其余挑战和可能的方向。

Index Terms—Deep Learning, Convolutional Neural Networks, Model Compression and Acceleration,

索引项-深度学习，卷积神经网络，模型压缩和加速，

I. INTRODUCTION
一，导言

In recent years, deep neural networks have recently received lots of attention, been applied to different applications and achieved dramatic accuracy improvements in many tasks. These works rely on deep networks with millions or even billions of parameters, and the availability of GPUs with very high computation capability plays a key role in their success. For example, the work by Krizhevsky et al. [1] achieved breakthrough results in the 2012 ImageNet Challenge using a network containing 60 million parameters with five convolutional layers and three fully-connected layers. Usually, it takes two to three days to train the whole model on ImagetNet dataset with a NVIDIA K40 machine. Another example is the top face verification results on the Labeled Faces in the Wild (LFW) dataset were obtained with networks containing hundreds of millions of parameters, using a mix of convolutional, locally-connected, and fully-connected layers
近年来，深度神经网络最近受到了很多关注，已经应用于不同的应用并且在许多任务中实现了显着的准确性改进。这些工作依赖于具有数百万甚至数十亿参数的深度网络，具有极高计算能力的GPU的可用性在其成功中起着关键作用。例如，Krizhevsky等人的工作。 [1]在2012年ImageNet挑战赛中取得了突破性成果，使用的网络包含6000万个参数，包含五个卷积层和三个完全连接的层。通常，使用NVIDIA K40机器在ImagetNet数据集上训练整个模型需要两到三天。另一个例子是使用卷积，局部连接和完全连接层的混合，使用包含数亿个参数的网络获得野外Labeled Faces（LFW）数据集的顶面验证结果

Yu Cheng is a Researcher from Microsoft AI & Research, One Microsoft Way, Redmond, WA 98052, USA.
Yu Cheng是Microsoft AI＆Research，One Microsoft Way，Redmond，WA 98052，USA的研究员。

Duo Wang and Tao Zhang are with the Department of Automation, Tsinghua University, Beijing 100084, China.
Duo Wang和Tao Zhang在清华大学自动化系，北京100084。

Pan Zhou is with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China.
潘周在华中科技大学电子信息与通信学院，武汉430074。

[2], [3]. It is also very time-consuming to train such a model to get reasonable performance. In architectures that rely only on fully-connected layers, the number of parameters can grow to billions [4].
[2]，[3]。训练这样的模型以获得合理的性能也是非常耗时的。在仅依赖于完全连接的层的架构中，参数的数量可以增长到数十亿[4]。

As larger neural networks with more layers and nodes are considered, reducing their storage and computational cost becomes critical, especially for some real-time applications such as online learning and incremental learning. In addition, recent years witnessed significant progress in virtual reality, augmented reality, and smart wearable devices, creating unprecedented opportunities for researchers to tackle fundamental challenges in deploying deep learning systems to portable devices with limited resources (e.g. memory, CPU, energy, bandwidth). Efficient deep learning methods can have significant impacts on distributed systems, embedded devices, and FPGA for Artificial Intelligence. For example, the ResNet50 [5] with 50 convolutional layers needs over 95MB memory for storage and over 3.8 billion floating number multiplications when processing an image. After discarding some redundant weights, the network still works as usual but saves more than 75% of parameters and 50% computational time. For devices like cell phones and FPGAs with only several megabyte resources, how to compact the models used on them is also important.
考虑到具有更多层和节点的更大的神经网络，降低其存储和计算成本变得至关重要，特别是对于诸如在线学习和增量学习的一些实时应用。此外，近年来在虚拟现实，增强现实和智能可穿戴设备方面取得了显着进步，为研究人员提供了前所未有的机会，可以解决将深度学习系统部署到资源有限的便携式设备（例如内存，CPU，能源，带宽）的基本挑战。

高效的深度学习方法会对分布式系统，嵌入式设备和人工智能FPGA产生重大影响。例如，具有50个卷积层的ResNet50 [5]在处理图像时需要超过95MB的存储空间和超过38亿次浮点数乘法。丢弃一些冗余权重后，网络仍可正常工作，但可节省75％以上的参数和50％的计算时间。对于仅具有几兆字节资源的手机和FPGA等设备，如何压缩其上使用的模型也很重要。

Achieving these goal calls for joint solutions from many disciplines, including but not limited to machine learning, optimization, computer architecture, data compression, indexing, and hardware design. In this paper, we review recent works on compressing and accelerating deep neural networks, which attracted a lot of attention from the deep learning community and already achieved lots of progress in the past years.

实现这些目标需要来自许多学科的联合解决方案，包括但不限于机器学习，优化，计算机体系结构，数据压缩，索引和硬件设计。在本文中，我们回顾了最近关于压缩和加速深度神经网络的工作，这引起了深度学习界的广泛关注，并且在过去几年中已经取得了很多进展。

We classify these approaches into four categories: parameter pruning and sharing, low-rank factorization, transferred/compact convolutional filters, and knowledge distillation. The parameter pruning and sharing based methods explore the redundancy in the model parameters and try to remove the redundant and uncritical ones. Low-rank factorization based techniques use matrix/tensor decomposition to estimate the informative parameters of the deep CNNs. The approaches based on transferred/compact convolutional filters design special structural convolutional filters to reduce the parameter space and save storage/computation. The knowledge distillation methods learn a distilled model and train a more compact neural network to reproduce the output of a larger network.

我们将这些方法分为四类：参数修剪和共享，低秩分解，转移/紧凑卷积滤波器和知识蒸馏。参数修剪和基于共享的方法探索模型参数中的冗余，并尝试去除冗余和不重要的参数。基于低秩分解的技术使用矩阵/张量分解来估计深CNN的信息参数。基于传输/紧凑卷积滤波器的方法设计了特殊的结构卷积滤波器，以减少参数空间并节省存储/计算。知识蒸馏方法学习蒸馏模型并训练更紧凑的神经网络以再现更大网络的输出。

In Table I, we briefly summarize these four types of methods. Generally, the parameter pruning & sharing, lowrank factorization and knowledge distillation approaches can

在表I中，我们简要总结了这四种方法。通常，参数修剪和共享，低分数分解和知识蒸馏方法可以

TABLE I SUMMARIZATION OF DIFFERENT APPROACHES FOR MODEL COMPRESSION AND ACCELERATION.

表I模型压缩和加速的不同方法的概述。

Theme Name	Description	Applications	More details
Parameter pruning and sharing	Reducing redundant parameters which are not sensitive to the performance	Convolutional layer and fully connected layer	Robust to various settings, can achieve good performance, can support both train from scratch and pre-trained model
Low-rank factorization	Using matrix/tensor decomposition to estimate the informative parameters	Convolutional layer and fully connected layer	Standardized pipeline, easily to be implemented, can support both train from scratch and pre-trained model
Transferred/compact convolutional filters	Designing special structural convolutional filters to save parameters	Convolutional layer only	Algorithms are dependent on applications, usually achieve good performance, only support train from scratch
Knowledge distillation	Training a compact neural network with distilled knowledge of a large model	Convolutional layer and fully connected layer	Model performances are sensitive to applications and network structure only support train from scratch

be used in DNN models with fully connected layers and convolutional layers, achieving comparable performances. On the other hand, methods using transferred/compact filters are designed for models with convolutional layers only. Low-rank factorization and transfered/compact filters based approaches provide an end-to-end pipeline and can be easily implemented in CPU/GPU environment, which is straightforward. while parameter pruning & sharing use different methods such as vector quantization, binary coding and sparse constraints to perform the task. Generally it will take several steps to achieve the goal.
用于具有完全连接层和卷积层的DNN模型，实现相当的性能。另一方面，使用转移/紧凑型过滤器的方法仅适用于具有卷积层的模型。低秩因子分解和基于传输/压缩过滤器的方法提供端到端流水线，并且可以在CPU / GPU环境中轻松实现，这很简单。而参数修剪和共享使用不同的方法，例如矢量量化，二进制编码和稀疏约束来执行任务。通常，实现目标需要几个步骤。

Regarding the training protocols, models based on parameter pruning/sharing low-rank factorization can be extracted from pre-trained ones or trained from scratch. While the transferred/compact filter and knowledge distillation models can only support train from scratch. These methods are independently designed and complement each other. For example, transferred layers and parameter pruning & sharing can be used together, and model quantization & binarization can be used together with low-rank approximations to achieve further speedup. We will describe the details of each theme, their properties, strengths and drawbacks in the following sections.
关于训练协议，可以从预先训练的训练协议中提取基于参数修剪/共享低秩分解的模型，或者从头开始训练。转移/紧凑型过滤器和知识蒸馏模型只能从头开始支持列车。这些方法是独立设计的，相互补充。例如，可以一起使用转移的层和参数修剪和共享，并且可以将模型量化和二值化与低秩近似一起使用以实现进一步的加速。我们将在以下部分中描述每个主题的细节，它们的属性，优点和缺点。

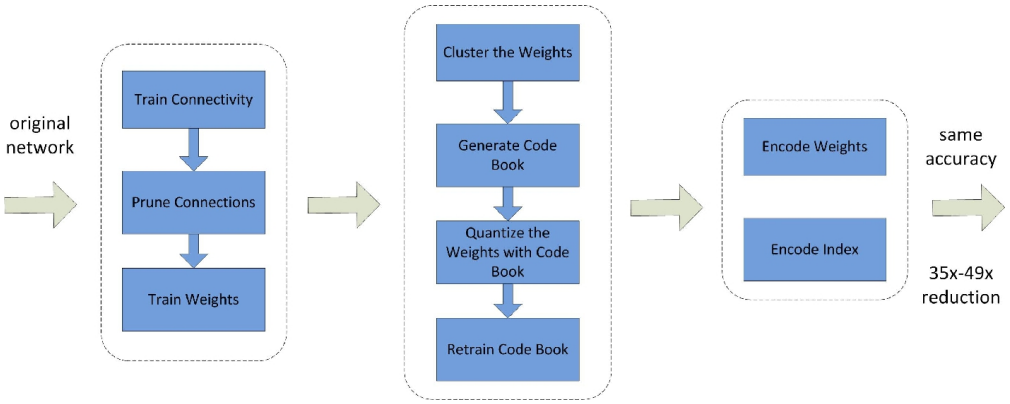
II. PARAMETER PRUNING AND SHARING
II。参数修剪和共享

Early works showed that network pruning is effective in reducing the network complexity and addressing the overfitting problem [6]. After that researcher found pruning originally introduced to reduce the structure in neural networks and hence improve generalization, it has been widely studied to compress DNN models, trying to remove parameters which are not crucial to the model performance. These techniques can be further classified into three sub-categories: quantization and binarization, parameter sharing, and structural matrix.
早期的工作表明，网络修剪可以有效地降低网络复杂性并解决过度配置问题[6]。在此研究人员发现最初引入的修剪以减少神经网络中的结构并因此改进泛化之后，已经广泛研究压缩DNN模型，试图去除对模型性能不重要的参数。这些技术可以进一步分为三个子类：量化和二值化，参数共享和结构矩阵。

A. Quantization and Binarization
A.量化和二值化

Network quantization compresses the original network by reducing the number of bits required to represent each weight. Gong et al. [6] and Wu et al. [7] applied k-means scalar quantization to the parameter values. Vanhoucke et al. [8] showed that 8-bit quantization of the parameters can result in significant speed-up with minimal loss of accuracy. The work in [9] used 16-bit fixed-point representation in stochastic rounding based CNN training, which significantly reduced Fig. 1. The three-stage compression method proposed in [10]: pruning, quantization and encoding. The input is the original model and the output is the compression model.
网络量化通过减少表示每个权重所需的比特数来压缩原始网络。龚等人。 [6]和吴等人。 [7]将k均值标量量化应用于参数值。Vanhoucke等。 [8]表明，参数的8位量化可以导致显着的加速，并且精度损失最小。[9]中的工作在基于随机舍入

的CNN训练中使用了16位固定点表示，这显着地减少了图1。[10]中提出的三阶段压缩方法：修剪，量化和编码。输入是原始模型，输出是压缩模型。



memory usage and float point operations with little loss in classification accuracy.
内存使用和浮点运算，分类精度几乎没有损失。

The method proposed in [10] quantized the link weights using weight sharing and then applied Huffman coding to the quantized weights as well as the codebook to further reduce the rate. As shown in Figure 1, it started by learning the connectivity via normal network training, followed by pruning the small-weight connections. Finally, the network was retrained to learn the final weights for the remaining sparse connections. This work achieved the state-of-art performance among all parameter quantization based methods. It was shown in [11] that Hessian weight could be used to measure the importance of network parameters, and proposed to minimize Hessianweighted quantization errors in average for clustering network parameters.
[10]中提出的方法使用权重共享来量化链路权重，然后将霍夫曼编码应用于量化权重以及码本以进一步降低速率。如图1所示，它首先通过正常的网络培训学习连接，然后修剪小重量连接。最后，对网络进行了重新训练，以了解剩余稀疏连接的最终权重。这项工作在所有基于参数量化的方法中实现了最先进的性能。在[11]中显示，Hessian权重可用于衡量网络参数的重要性，并建议最小化Hessian加权量化误差的平均聚类网络参数。

In the extreme case of the 1-bit representation of each weight, that is binary weight neural networks. There are many works that directly train CNNs with binary weights, for instance, BinaryConnect [12], BinaryNet [13] and XNORNetworks [14]. The main idea is to directly learn binary weights or activation during the model training. The systematic study in
在每个权重的1位表示的极端情况下，即二进制权重神经网络。有许多工作直接用二进制权重训练CNN，例如BinaryConnect [12]，BinaryNet [13]和XNORNetworks [14]。主要思想是在模型训练期间直接学习二进制权重或激活。系统研究

[15] showed that networks trained with back propagation could be resilient to specific weight distortions, including binary weights.
[15]表明，使用反向传播训练的网络可以适应特定的权重失真，包括二进制权重。

Drawbacks: the accuracy of the binary nets is significantly lowered when dealing with large CNNs such as GoogleNet. Another drawback of such binary nets is that existing binarization schemes are based on simple matrix approximations and ignore the effect of binarization on the accuracy loss.
缺点：在处理大型CNN（如GoogleNet）时，二进制网络的准确性会显着降低。这种二进制网络的另一个缺点是现有的二值化方案基于简单的矩阵近似，并忽略二值化对精度损失的影响。

To address this issue, the work in [16] proposed a proximal Newton algorithm with diagonal Hessian approximation that directly minimizes the loss with respect to the binary weights. The work in [17] reduced the time on float point multiplication in the

training stage by stochastically binarizing weights and converting multiplications in the hidden state computation to significant changes.
为了解决这个问题，[16]中的工作提出了一种具有对角线Hessian近似的近端牛顿算法，该算法直接最小化了相对于二元权重的损失。[17]中的工作通过随机二值化权值和将隐藏状态计算中的乘法转换为显著变化，减少了训练阶段浮点乘法的时间。

B. Pruning and Sharing
B.修剪和分享

Network pruning and sharing has been used both to reduce network complexity and to address the over-fitting issue. An early approach to pruning was the Biased Weight Decay [18]. The Optimal Brain Damage [19] and the Optimal Brain Surgeon [20] methods reduced the number of connections based on the Hessian of the loss function, and their work suggested that such pruning gave higher accuracy than magnitudebased pruning such as the weight decay method. The training procedure of those methods followed the way training from scratch manner.
网络修剪和共享已被用于降低网络复杂性和解决过度问题。修剪的早期方法是偏差重量衰减[18]。最佳脑损伤[19]和最佳脑外科医生[20]方法减少了基于损失函数的Hessian的连接数，并且他们的工作表明这种修剪比基于重量的修剪（例如重量衰减方法）提供更高的准确性。这些方法的培训程序采用从头开始的培训方式。

A recent trend in this direction is to prune redundant, non-informative weights in a pre-trained CNN model. For example, Srinivas and Babu [21] explored the redundancy among neurons, and proposed a data-free pruning method to remove redundant neurons. Han et al. [22] proposed to reduce the total number of parameters and operations in the entire network. Chen et al. [23] proposed a HashedNets model that used a low-cost hash function to group weights into hash buckets for parameter sharing. The deep compression method in [10] removed the redundant connections and quantized the weights, and then used Huffman coding to encode the quantized weights. In [24], a simple regularization method based on soft weight-sharing was proposed, which included both quantization and pruning in one simple (re-)training procedure. The above pruning schemes typically produce connections pruning in CNNs.
此方向的最新趋势是在预先训练的CNN模型中修剪冗余的非信息量权重。例如，Srinivas和Babu [21]研究了神经元之间的冗余，并提出了一种无数据修剪方法来去除冗余神经元。韩等人。 [22]建议减少整个网络中的参数和操作总数。陈等人。 [23]提出了一个HashedNets模型，该模型使用低成本散列函数将权重分组为散列桶以进行参数共享。[10]中的深度压缩方法去除了冗余连接并量化了权重，然后使用霍夫曼编码对量化权重进行编码。在[24]中，提出了一种基于软权重共享的简单正则化方法，其中包括在一个简单（重新）训练过程中的量化和修剪。上述修剪方案通常在CNN中产生连接修剪。

There is also growing interest in training compact CNNs with sparsity constraints. Those sparsity constraints are typically introduced in the optimization problem as l_0 or l_1 norm regularizers. The work in [25] imposed group sparsity constraint on the convolutional filters to achieve structured brain Damage, i.e., pruning entries of the convolution kernels in a group-wise fashion. In [26], a group-sparse regularizer on neurons was introduced during the training stage to learn compact CNNs with reduced filters. Wen et al. [27] added a structured sparsity regularizer on each layer to reduce trivial filters, channels or even layers. In the filter-level pruning, all the above works used $l_{2,1}$ -norm regularizers. The work in [28] used l_1 -norm to select and prune unimportant filters.
人们对培养具有稀疏性限制的紧凑型CNN的兴趣也越来越大。这些稀疏性约束通常作为l_0或l_1范数正则化器在优化问题中引入。[25]中的工作对卷积滤波器施加了群体稀疏性约束，以实现结构化脑损伤，即以分组方式修剪卷积核的条目。在[26]中，在训练阶段引入了一组神经元的稀疏正则化器来学习具有减少滤波器的紧凑CNN。温等人。 [27]在每一层上增加了一个结构化的稀疏正则化器，以减少琐碎的过滤器，通道甚至层。在过滤级别的修剪中，所有上述工作都使用了$l_{2,1}$ -norm正规化器。[28]中的工作使用l_1 -norm来选择和修剪不重要的过滤器。

Drawbacks: there are some potential issues of the pruning and sharing. First, pruning with l_1 or l_2 regularization requires more iterations to converge than general. In addition, all pruning criteria require manual setup of sensitivity for layers, which demands fine-tuning of the parameters and could be cumbersome for some applications.
缺点：修剪和分享存在一些潜在问题。首先，使用l_1或l_2正则化进行修剪需要比一般情况更多的迭代才能收敛。此外，所有修剪标准都需要手动设置层的灵敏度，这需要对参数进行微调，并且对于某些应用来说可能很麻烦。

C. Designing Structural Matrix
C.设计结构矩阵

In architectures that contain fully-connected layers, it is critical to explore this redundancy of parameters in fully connected layers, which is often the bottleneck in terms of memory consumption. These network layers use the nonlinear transforms $f(\mathbf{x}, \mathbf{M}) = \sigma(\mathbf{Mx})$, where $\sigma(\cdot)$ is an element-wise nonlinear operator, x is the input vector, and M is the $m \times n$ matrix of parameters [29]. When M is a large general dense matrix, the cost of storing mn parameters and computing matrix-vector products in $O(mn)$ time. Thus, an intuitive way to prune parameters is to impose x as a parameterized structural matrix. An $m \times n$ matrix that can be described using much fewer parameters than mn is called a structured matrix. Typically, the structure should not only reduce the memory cost, but also dramatically accelerate the inference and training stage via fast matrix-vector multiplication and gradient computations.
在包含完全连接层的架构中，探索完全连接层中的参数冗余至关重要，这通常是内存消耗方面的瓶颈。这些网络层使用非线性变换$f(\mathbf{x}, \mathbf{M}) = \sigma(\mathbf{Mx})$，其中$\sigma(\cdot)$是元素非线性算子，x是输入向量，M是$m \times n$参数矩阵[29]。当M是一个大的一般密集矩阵时，存储mn参数和计算矩阵矢量乘积的成本在$O(mn)$时间内。因此，修剪参数的直观方式是将x强加为参数化结构矩阵。可以使用比mn少得多的参数来描述的$m \times n$矩阵称为结构化矩阵。通常，该结构不仅应该降低存储器成本，而且还应通过快速矩阵向量乘法和梯度计算显着加速推理和训练阶段。

Following this direction, the work in [30], [31] proposed a simple and efficient approach based on circulant projections, while maintaining competitive error rates. Given a vector $\mathbf{r} = (r_0, r_1, \cdots, r_{d-1})$, a circulant matrix $\mathbf{R} \in \mathbf{R}^{d \times d}$ is defined as:
按照这个方向，[30]，[31]中的工作提出了一种基于循环投影的简单而有效的方法，同时保持竞争错误率。给定矢量$\mathbf{r} = (r_0, r_1, \cdots, r_{d-1})$，循环矩阵$\mathbf{R} \in \mathbf{R}^{d \times d}$定义为：

$$\mathbf{R} = \text{circ}(\mathbf{r}) := \begin{bmatrix} r_0 & r_{d-1} & \cdots & r_2 & r_1 \\ r_1 & r_0 & r_{d-1} & & r_2 \\ \vdots & r_1 & r_0 & \ddots & \vdots \\ r_{d-2} & & \ddots & \ddots & r_{d-1} \\ r_{d-1} & r_{d-2} & \cdots & r_1 & r_0 \end{bmatrix}. \quad (1)$$

thus the memory cost becomes $\mathcal{O}(d)$ instead of $\mathcal{O}(d^2)$. This circulant structure also enables the use of Fast Fourier Transform (FFT) to speed up the computation. Given a ddimensional vector r, the above 1-layer circulant neural network in Eq. 1 has time complexity of $\mathcal{O}(d \log d)$.
因此，内存成本变为$\mathcal{O}(d)$而不是$\mathcal{O}(d^2)$。该循环结构还使得能够使用快速傅里叶变换（FFT）来加速计算。给定一个二维向量r，上述1层循环神经网络在方程式中。 1的时间复杂度为$\mathcal{O}(d \log d)$。

In [32], a novel Adaptive Fastfood transform was introduced to reparameterize the matrix-vector multiplication of fully connected layers. The Adaptive Fastfood transform matrix $\mathbf{R} \in \mathbf{R}^{n \times d}$ was defined as:
在[32]中，引入了一种新颖的自适应快速变换，用于重新参数化完全连接层的矩阵向量乘法。自适应快速转换矩阵$\mathbf{R} \in \mathbf{R}^{n \times d}$定义为：

$$\mathbf{R} = \mathbf{SHG\Pi HB} \quad (2)$$

where S, G and B are random diagonal matrices. $\mathbf{\Pi} \in \{0,1\}^{d\times d}$ is a random permutation matrix, and H denotes the Walsh-Hadamard matrix. Reparameterizing a fully connected layer with d inputs and n outputs using the Adaptive Fastfood transform reduces the storage and the computational costs from $\mathcal{O}(nd)$ to $\mathcal{O}(n)$ and from $\mathcal{O}(nd)$ to $\mathcal{O}(n\log d)$, respectively.
其中S，G和B是随机对角矩阵。$\mathbf{\Pi} \in \{0,1\}^{d\times d}$是随机置换矩阵，H表示Walsh-Hadamard矩阵。使用自适应快速转换器对具有d输入和n输出的完全连接层进行重新参数化，可分别将$\mathcal{O}(nd)$和$\mathcal{O}(n)$以及$\mathcal{O}(nd)$和$\mathcal{O}(n\log d)$的存储和计算成本分别降低。

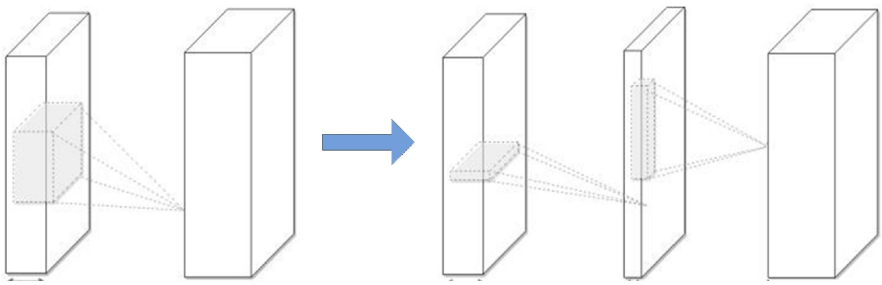
The work in [29] showed the effectiveness of the new notion of parsimony in the theory of structured matrices. Their proposed method can be extended to various other structured matrix classes, including block and multi-level Toeplitz-like
[29]中的工作表明了简约概率在结构矩阵理论中的有效性。他们提出的方法可以扩展到各种其他结构化矩阵类，包括块级和多级Toeplitz类

[33] matrices related to multi-dimensional convolution [34]. Following this idea, [35] proposed a general structured efficient linear layer for CNNs.
[33]与多维卷积有关的矩阵[34]。根据这一想法，[35]提出了一种用于CNN的通用结构化高效线性层。

Drawbacks: one problem of this kind of approaches is that the structural constraint will hurt the performance since the constraint might bring bias to the model. On the other hand, how to find a proper structural matrix is difficult. There is no theoretical way to derive it out.
缺点：这种方法的一个问题是结构约束会损害性能，因为约束可能会给模型带来偏差。另一方面，如何找到合适的结构矩阵是很困难的。没有理论上的方法可以推导出来。

III. LOW-RANK FACTORIZATION AND SPARSITY
III。低秩度因子和稀疏性

Convolution operations contribute the bulk of most computations in deep CNNs, thus reducing the convolution layer Fig. 2. A typical framework of the low-rank regularization method. The left is the original convolutional layer and the right is the low-rank constraint convolutional layer with rank-K.
卷积运算在深CNN中贡献了大部分计算，从而减少了卷积层（图2）。低阶正则化方法的典型框架。左边是原始卷积层，右边是具有秩-K的低秩约束卷积层。



would improve the compression rate as well as the overall speedup. For the convolution kernels, it can be viewed as a 4D tensor. Ideas based on tensor decomposition is derived by the intuition that there is a significant amount of redundancy in the 4D tensor, which is a particularly promising way to remove the redundancy. Regarding the fully-connected layer, it can be view as a 2D matrix and the low-rankness can also help.
将提高压缩率以及整体加速。对于卷积核，它可以被视为4D张量。基于张量分解的思想是通过直觉得出的，即4D张量中存在大量冗余，这是一种特别有前途的去除冗余的方法。关于完全连接的层，它可以被视为2D矩阵，并且低秩也可以

帮助。

It has been a long time for using low-rank filters to accelerate convolution, for example, high dimensional DCT (discrete cosine transform) and wavelet systems using tensor products to be constructed from 1D DCT transform and 1D wavelets respectively. Learning separable 1D filters was introduced by Rigamonti et al. [36], following the dictionary learning idea. Regarding some simple DNN models, a few low-rank approximation and clustering schemes for the convolutional kernels were proposed in [37]. They achieved 2× speedup for a single convolutional layer with 1% drop in classification accuracy. The work in [38] proposed using different tensor decomposition schemes, reporting a 4.5× speedup with 1% drop in accuracy in text recognition.
使用低秩滤波器来加速卷积已经很长时间了，例如，高维DCT（离散余弦变换）和使用张量积的小波系统分别由1D DCT变换和1D小波构成。Rigamonti等人介绍了学习可分离的一维滤波器。 [36]，遵循字典学习的想法。对于一些简单的DNN模型，在[37]中提出了一些卷积核的低秩近似和聚类方案。对于单个卷积层，它们实现了2倍的加速，分类精度降低了1％。[38]中的工作提出使用不同的张量分解方案，报告4.5倍的加速，文本识别的准确度下降1％。

The low-rank approximation was done layer by layer. The parameters of one layer were fixed after it was done, and the layers above were fine-tuned based on a reconstruction error criterion. These are typical low-rank methods for compressing 2D convolutional layers, which is described in Figure 2. Following this direction, Canonical Polyadic (CP) decomposition of was proposed for the kernel tensors in [39]. Their work used nonlinear least squares to compute the CP decomposition. In [40], a new algorithm for computing the low-rank tensor decomposition for training low-rank constrained CNNs from scratch were proposed. It used Batch Normalization (BN) to transform the activation of the internal hidden units. In general, both the CP and the BN decomposition schemes in [40] (BN Low-rank) can be used to train CNNs from scratch. However, there are few differences between them. For example, finding the best low-rank approximation in CP decomposition is an illposed problem, and the best rank-K (K is the rank number) approximation may not exist sometimes. While for the BN scheme, the decomposition always exists. We perform a simple comparison of both methods shown in Table II. The actual speedup and the compression rates are used to measure their performances.
低秩近似是逐层完成的。在完成之后固定一层的参数，并且基于重建误差标准对上述层进行微调。这些是用于压缩2D卷积层的典型低秩方法，如图2所示。按照这个方向，Canonical Polyadic（CP）分解被提出用于[39]中的核张量。他们的工作使用非线性最小二乘法来计算CP分解。在[40]中，提出了一种用于计算低秩张量分解的新算法，用于从头开始训练低秩约束CNN。它使用批量标准化（BN）来转换内部隐藏单元的激活。通常，[40]（BN低级别）中的CP和BN分解方案都可用于从头开始训练CNN。但是，它们之间几乎没有差异。例如，在CP分解中找到最佳低秩近似是一个错误的问题，并且有时可能不存在最佳秩-K（K是秩数）近似。而对于BN方案，分解始终存在。我们对表II中所示的两种方法进行简单比较。实际加速和压缩率用于测量它们的性能。

As we mentioned before, the fully connected layers can be viewed as a 2D matrix and thus the above mentioned methods can also be applied there. There are several classical works on exploiting low-rankness in fully connected layers.
如前所述，完全连接的层可以被视为2D矩阵，因此上述方法也可以应用于那里。有几个关于在完全连接层中利用低秩度的经典着作。

TABLE II COMPARISONS BETWEEN THE LOW-RANK MODELS AND THEIR BASELINES ON ILSVRC-2012.
表II ILSVRC-2012上低级模型与其基线之间的比较。

Model	TOP-5 Accuracy	Speed-up	Compression Rate
AlexNet	80.03%	1.	1.
BN Low-rank	80.56%	1.09	4.94
CP Low-rank	79.66%	1.82	5.
VGG-16	90.60%	1.	1.
BN Low-rank	90.47%	1.53	2.72
CP Low-rank	90.31%	2.05	2.75

CP Low-rank	90.31%	2.05	2.75
GoogleNet	92.21%	1.	1.
BN Low-rank	91.88%	1.08	2.79
CP Low-rank	91.79%	1.20	2.84

For instance, Misha et al. [41] reduced the number of dynamic parameters in deep models using the low-rank method. [42] explored a low-rank matrix factorization of the final weight layer in a DNN for acoustic modeling. In [3], Lu et al. adopted truncated SVD (singular value decomposition) to decompsite the fully connected layer for designing compact multi-task deep learning architectures.

例如，Misha等人。 [41]使用低秩方法减少深度模型中的动态参数的数量。 [42]探索了用于声学建模的DNN中最终权重层的低秩矩阵分解。在[3]中，Lu等人。采用截断SVD（奇异值分解）对完全连通层进行分解，以设计紧凑的多任务深度学习架构。

Drawbacks: low-rank approaches are straightforward for model compression and acceleration. The idea complements recent advances in deep learning, such as dropout, rectified units and maxout. However, the implementation is not that easy since it involves decomposition operation, which is computationally expensive. Another issue is that current methods perform low-rank approximation layer by layer, and thus cannot perform global parameter compression, which is important as different layers hold different information. Finally, factorization requires extensive model retraining to achieve convergence when compared to the original model.

缺点：低级方法对于模型压缩和加速很简单。这个想法补充了深度学习的最新进展，例如辍学，整修单元和最大化。然而，实现并不那么容易，因为它涉及分解操作，这在计算上是昂贵的。另一个问题是当前方法逐层执行低秩近似，因此不能执行全局参数压缩，这对于不同层保持不同信息是重要的。最后，与原始模型相比，因子分解需要大量的模型再训练以实现收敛。

IV. TRANSFERRED/COMPACT CONVOLUTIONAL FILTERS

IV。转移/紧凑的卷积滤波器

CNNs are parameter efficient due to exploring the translation invariant property of the representations to the input image, which is the key to the success of training very deep models without severe over-fitting. Although a strong theory is currently missing, a large number of empirical evidence support the notion that both the translation invariant property and the convolutional weight sharing are important for good predictive performance. The idea of using transferred convolutional filters to compress CNN models is motivated by recent works in [43], which introduced the equivariant group theory. Let x be an input, $\Phi(\cdot)$ be a network or layer and $\mathcal{T}(\cdot)$ be the transform matrix. The concept of equivalence is defined as:

CNN是参数高效的，因为探索了表示对输入图像的平移不变性，这是在没有严重过度训练的情况下训练非常深的模型成功的关键。虽然目前缺乏强有力的理论，但大量的经验证据支持这样的观点，即翻译不变性和卷积权重分配对良好的预测性能都很重要。使用转移卷积滤波器压缩CNN模型的想法受到[43]中最近的工作的推动，其引入了等变群理论。设x为输入，$\Phi(\cdot)$为网络或层，$\mathcal{T}(\cdot)$为变换矩阵。等价的概念定义为：

$$\mathcal{T}'\Phi(\mathbf{x}) = \Phi(\mathcal{T}\mathbf{x}) \qquad (3)$$

indicating that transforming the input x by the transform $\mathcal{T}(\cdot)$ and then passing it through the network or layer $\Phi(\cdot)$ should give the same result as first mapping x through the network and then transforming the representation. Note that in Eq.

指示通过转换$\mathcal{T}(\cdot)$转换输入x然后将其传递通过网络或层$\Phi(\cdot)$应该提供与通过网络首先映射x然后转换表示相同的结果。请注意，在Eq。

(10), the transforms $\mathcal{T}(\cdot)$ and $\mathcal{T}'(\cdot)$ are not necessarily the same as they operate on different objects. According to this theory, it is reasonable applying transform to layers or filters $\Phi(\cdot)$ to compress the whole network models. From empirical observation, deep CNNs also benefit from using a large set of convolutional filters by applying certain transform $\mathcal{T}(\cdot)$ to a small set of base filters since it acts as a regularizer for the model.

（10），变换$\mathcal{T}(\cdot)$和$\mathcal{T}'(\cdot)$不一定与它们在不同对象上操作相同。根据这个理论，将转换应用于层或过滤器$\Phi(\cdot)$以压缩整个网络模型是合理的。从经验观察来看，深度CNN也可以通过将一些变换$\mathcal{T}(\cdot)$应用于一小组基本滤波器而使用大量卷积滤波器，因为它充当模型的正则化器。

Following this direction, there are many recent reworks proposed to build a convolutional layer from a set of base filters [43]–[46]. What they have in common is that the transform $\mathcal{T}(\cdot)$ lies in the family of functions that only operate in the spatial domain of the convolutional filters. For example, the work in [45] found that the lower convolution layers of CNNs learned redundant filters to extract both positive and negative phase information of an input signal, and defined $\mathcal{T}(\cdot)$ to be the simple negation function:

按照这个方向，最近有许多改造建议用来从一组基础滤波器构建卷积层[43] - [46]。它们的共同点是变换$\mathcal{T}(\cdot)$位于仅在卷积滤波器的空间域中运行的函数族。例如，[45]中的工作发现CNN的较低卷积层学习冗余滤波器以提取输入信号的正相和负相信息，并将$\mathcal{T}(\cdot)$定义为简单的否定函数：

$$\mathcal{T}(\mathbf{W}_x) = \mathbf{W}_x^- \tag{4}$$

where $\mathbf{W}_x$ is the basis convolutional filter and $\mathbf{W}_x^-$ is the filter consisting of the shifts whose activation is opposite to that of $\mathbf{W}_x$ and selected after max-pooling operation. By doing this, the work in [45] can easily achieve 2× compression rate on all the convolutional layers. It is also shown that the negation transform acts as a strong regularizer to improve the classification accuracy. The intuition is that the learning algorithm with pair-wise positive-negative constraint can lead to useful convolutional filters instead of redundant ones.

其中$\mathbf{W}_x$是基础卷积滤波器，$\mathbf{W}_x^-$是由激活与$\mathbf{W}_x$相反并在最大池操作后选择的移位组成的滤波器。通过这样做，[45]中的工作可以容易地在所有卷积层上实现2倍压缩率。还表明，否定变换充当强调节器以提高分类精度。直觉是具有成对正负约束的学习算法可以导致有用的卷积滤波器而不是冗余滤波器。

In [46], it was observed that magnitudes of the responses from convolutional kernels had a wide diversity of pattern representations in the network, and it was not proper to discard weaker signals with a single threshold. Thus a multi-bias nonlinearity activation function was proposed to generates more patterns in the feature space at low computational cost. The transform $\mathcal{T}(\cdot)$ was define as:

在[46]中，观察到来自卷积核的响应的大小在网络中具有多种模式表示，并且用单个阈值丢弃较弱信号是不合适的。因此，提出了一种多偏置非线性激活函数，以低计算成本在特征空间中生成更多模式。变换$\mathcal{T}(\cdot)$定义如下：

$$\mathcal{T}'\Phi(\mathbf{x}) = \mathbf{W}_x + \delta \tag{5}$$

where δ were the multi-bias factors. The work in [47] considered a combination of rotation by a multiple of $90°$ and horizontal/vertical flipping with:

其中δ是多偏差因子。 [47]中的工作考虑了$90°$的倍数和水平/垂直移动的旋转组合：

$$\mathcal{T}'\Phi(\mathbf{x}) = \mathbf{W}^{T_\theta} \tag{6}$$

where $\mathbf{W}^{T_\theta}$ was the transformation matrix which rotated the original filters with angle $\theta \in \{90, 180, 270\}$. In [43], the transform was generalized to any angle learned from data, and θ was directly obtained from data. Both works [47] and [43] can achieve good classification performance.
其中$\mathbf{W}^{T_\theta}$是转换矩阵，它以$\theta \in \{90, 180, 270\}$角度旋转原始滤波器。在[43]中，将变换推广到从数据中学习的任何角度，并且θ直接从数据中获得。两个作品[47]和[43]都可以实现良好的分类性能。

The work in [44] defined $\mathcal{T}(\cdot)$ as the set of translation functions applied to 2D filters:
[44]中的工作将$\mathcal{T}(\cdot)$定义为应用于2D滤波器的一组平移函数：

$$\mathcal{T}^{`}\Phi(\mathbf{x}) = T(\cdot, x, y)_{x,y\in\{-k,\ldots,k\},(x,y)\neq(0,0)} \qquad (7)$$

where $T(\cdot, x, y)$ denoted the translation of the first operand by (x, y) along its spatial dimensions, with proper zero padding at borders to maintain the shape. The proposed framework can be used to 1) improve the classification accuracy as a regularized version of maxout networks, and 2) to achieve parameter efficiency by flexibly varying their architectures to compress networks.
其中$T(\cdot, x, y)$表示(x, y)沿其空间维度对第一个操作数的转换，在边界处具有适当的零填充以保持形状。所提出的框架可用于1）提高分类精度作为maxout网络的正则化版本，以及2）通过灵活地改变其架构来压缩网络来实现参数效率。

Table III briefly compares the performance of different methods with transferred convolutional filters, using VGGNet (16 layers) as the baseline model. The results are reported on CIFAR-10 and CIFAR-100 datasets with Top-5 error. It is observed that they can achieve reduction in parameters with little or no drop in classification accuracy.
表III简要比较了使用VGGNet（16层）作为基线模型的不同方法与转移卷积滤波器的性能。结果报告在具有Top-5错误的CIFAR-10和CIFAR-100数据集上。据观察，它们可以实现参数的减少，而分类精度很少或没有下降。

TABLE III A SIMPLE COMPARISON OF DIFFERENT APPROACHES ON CIFAR-10 AND CIFAR-100.
表III CIFAR-10和CIFAR-100的不同方法的简单比较。

Model CIFAR-100 CIFAR-10 Compression Rate VGG-16 34.26% 9.85% 1. MBA [46] 33.66% 9.76% 2. CRELU [45] 34.57% 9.92% 2. CIRC [43] 35.15% 10.23% 4. DCNN [44] 33.57% 9.65% 1.62 Drawbacks: there are few issues to be addressed for approaches that apply transform constraints to convolutional filters. First, these methods can achieve competitive performance for wide/flat architectures (like VGGNet) but not thin/deep ones (like GoogleNet, Residual Net). Secondly, the transfer assumptions sometimes are too strong to guide the learning, making the results unstable in some cases.
型号CIFAR-100 CIFAR-10压缩率VGG-16 34.26％9.85％1。MBA [46] 33.66％9.76％2。CRELU [45] 34.57％9.92％2。CIRC [43] 35.15％10.23％4。DCNN [44] 33.57％9.65％1.62缺点：对于将变换约束应用于卷积滤波器的方法，几乎没有需要解决的问题。首先，这些方法可以为广泛的架构（如VGGNet）实现竞争性能，但不能实现薄/深层（如GoogleNet，Residual Net）。其次，转移假设有时太强，无法指导学习，在某些情况下结果不稳定。

Using a compact filter for convolution can directly reduce the computation cost. The key idea is to replace the loose and over-parametric filters with compact blocks to improve the speed, which significantly accelerate CNNs on several benchmarks. Decomposing 3×3 convolution into two 1×1 convolutions was used in [48], which achieved significant acceleration on object recognition. SqueezeNet [49] was proposed to replace 3×3 convolution with 1×1 convolution, which created a compact neural network with about 50 fewer parameters and comparable accuracy when compared to AlexNet.
使用紧凑型滤波器进行卷积可以直接降低计算成本。关键的想法是用紧凑的模块替换松散和过度参数化的滤波器，以提高速度，从而在几个基准测试中显着加速CNN。在[48]中使用了3×3卷积分解为两个1×1卷积，它在物体识别上实

现了显着的加速。SqueezeNet [49]被提议用1×1卷积代替3×3卷积，它创建了一个紧凑的神经网络，与AlexNet相比，参数减少了50个，精度也相当。

V. KNOWLEDGE DISTILLATION
五，知识精馏

To the best of our knowledge, exploiting knowledge transfer (KT) to compress model was first proposed by Caruana et al. [50]. They trained a compressed/ensemble model of strong classifiers with pseudo-data labeled, and reproduced the output of the original larger network. But the work is limited to shallow models. The idea has been recently adopted in [51] as knowledge distillation (KD) to compress deep and wide networks into shallower ones, where the compressed model mimicked the function learned by the complex model. The main idea of KD based approaches is to shift knowledge from a large teacher model into a small one by learning the class distributions output via softmax.
据我们所知，Caruana等人首先提出利用知识转移（KT）来压缩模型。 [50]。他们训练了一个强类别的压缩/集合模型，标记了伪数据，并重现了原始大网络的输出。但这项工作仅限于浅层模型。最近在[51]中采用了这种思想作为知识蒸馏（KD）将深度和宽度网络压缩成较浅的网络，其中压缩模型模拟了复杂模型所学习的函数。基于KD的方法的主要思想是通过学习通过softmax输出的类分布将知识从大型教师模型转变为小型教师模型。

The work in [52] introduced a KD compression framework, which eased the training of deep networks by following a student-teacher paradigm, in which the student was penalized according to a softened version of the teacher’s output. The framework compressed an ensemble of teacher networks into a student network of similar depth. The student was trained to predict the output and the classification labels. Despite its simplicity, KD demonstrates promising results in various image classification tasks. The work in [53] aimed to address the network compression problem by taking advantage of depth neural networks. It proposed an approach to train thin but deep networks, called FitNets, to compress wide and shallower (but still deep) networks. The method was extended the idea to allow for thinner and deeper student models. In order to learn from the intermediate representations of teacher network, FitNet made the student mimic the full feature maps of the teacher. However, such assumptions are too strict since the capacities of teacher and student may differ greatly.
[52]中的工作引入了KD压缩框架，该框架通过遵循学生 - 教师范式来简化深度网络的训练，其中学生根据教师输出的软化版本受到惩罚。该框架将教师网络集成到一个类似深度的学生网络中。学生接受了预测输出和分类标签的培训。尽管KD非常简单，但它在各种图像分类任务中展示了有希望的结果。[53]中的工作旨在通过利用深度神经网络来解决网络压缩问题。它提出了一种训练瘦而深的网络的方法，称为FitNets，以压缩宽而浅（但仍然很深）的网络。该方法扩展了允许更薄更深的学生模型的想法。为了从教师网络的中间表示中学习，FitNet使学生模仿了教师的完整特征图。然而，由于教师和学生的能力可能差别很大，因此这些假设过于严格。

All the above approaches are validated on MNIST, CIFAR10, CIFAR-100, SVHN and AFLW benchmark datasets, and experimental results show that these methods match or outperform the teacher’s performance, while requiring notably fewer parameters and multiplications.
所有上述方法都在MNIST，CIFAR10，CIFAR-100，SVHN和AFLW基准数据集上得到验证，实验结果表明这些方法匹配或优于教师的表现，同时需要的参数和乘法量明显减少。

There are several extension along this direction of distillation knowledge. The work in [54] trained a parametric student model to approximate a Monte Carlo teacher. The proposed framework used online training, and used deep neural networks for the student model. Different from previous works which represented the knowledge using the soften label probabilities, [55] represented the knowledge by using the neurons in the higher hidden layer, which preserved as much information as the label probabilities, but are more compact. The work in [56] accelerated the experimentation process by instantaneously transferring the knowledge from a previous network to each new deeper or wider network. The techniques are based on the concept of function-preserving

transformations between neural network specifications. Zagoruyko et al. [57] proposed Attention Transfer (AT) to relax the assumption of FitNet. They transferred the attention maps that are summaries of the full activations.
沿着蒸馏知识的这个方向有几个延伸。[54]中的工作训练了一个参数化的学生模型，以近似蒙特卡罗教师。建议的框架使用在线培训，并使用深度神经网络为学生模型。与先前使用软化标签概率表示知识的作品不同，[55]通过使用较高隐藏层中的神经元来表示知识，其中保留了与标签概率一样多的信息，但是更紧凑。[56]中的工作通过即时将知识从先前的网络传输到每个新的更深或更广的网络来加速实验过程。这些技术基于神经网络规范之间保持功能的转换概念。Zagoruyko等。 [57]建议注意转移（AT）放宽FitNet的假设。他们转移了关注图，这些图是完整激活的摘要。

Drawbacks: KD-based approaches can make deeper models thinner and help significantly reduce the computational cost. However, there are a few disadvantages. One of those is that KD can only be applied to classification tasks with softmax loss function, which hinders its usage. Another drawback is the model assumptions sometimes are too strict to make the performance competitive with other type of approaches.
缺点：基于KD的方法可以使更深的模型更薄，并有助于显着降低计算成本。但是，有一些缺点。其中之一是KD只能应用于具有softmax损失功能的分类任务，这会阻碍其使用。另一个缺点是模型假设有时太严格，不能使性能与其他类型的方法竞争。

VI. OTHER TYPES OF APPROACHES
VI。其他类型的方法

We first summarize the works utilizing attention-based methods. Note that attention-based mechanism [58] can reduce computations significantly by learning to selectively focus or “attend” to a few, task-relevant input regions. The work in
我们首先总结了利用基于注意力的方法的作品。请注意，基于注意力的机制[58]可以通过学习选择性地聚焦或“参与”一些与任务相关的输入区域来显着减少计算。工作在

[59] introduced the dynamic capacity network (DCN) that combined two types of modules: the small sub-networks with low capacity, and the large ones with high capacity. The lowcapacity sub-networks were active on the whole input to first find the task-relevant areas, and then the attention mechanism was used to direct the high-capacity sub-networks to focus on the task-relevant regions. By dong this, the size of the CNNs model has been significantly reduced.
[59]引入了动态容量网络（DCN），它结合了两种类型的模块：低容量的小型子网络和高容量的大型子网络。低容量子网络在整个输入上是活跃的，首先发现任务相关区域，然后使用注意机制指导高容量子网络关注任务相关区域。通过这个，CNN模型的大小已经显着减少。

Following this direction, the work in [60] introduced the conditional computation idea, which only computes the gradient for some important neurons. It proposed a sparselygated mixture-of-experts Layer (MoE). The MoE module consisted of a number of experts, each a simple feed-forward neural network, and a trainable gating network that selected a sparse combination of the experts to process each input. In [61], dynamic deep neural networks (D2NN) were introduced, which were a type of feed-forward deep neural network that selected and executed a subset of D2NN neurons based on the input.
按照这个方向，[60]中的工作引入了条件计算思想，它只计算一些重要神经元的梯度。它提出了稀疏混合专家层（MoE）。教育部模块由许多专家组成，每个专家都是一个简单的前馈神经网络，以及一个可训练的选通网络，它选择了专家的稀疏组合来处理每个输入。在[61]中，引入了动态深度神经网络（D2NN），它是一种前馈深度神经网络，它根据输入选择并执行D2NN神经元的子集。

There have been other attempts to reduce the number of parameters of neural networks by replacing the fully connected layer with global average pooling [44], [62]. Network architecture such as GoogleNet or Network in Network, can achieve state-of-the-art results on several benchmarks by adopting this idea. However, these architectures have not been fully optimized the utilization of

the computing resources inside the network. This problem was noted by Szegedy et al. [62] and motivated them to increase the depth and width of the network while keeping the computational budget constant.
已经有其他尝试通过用全局平均池替换完全连接的层来减少神经网络的参数数量[44]，[62]。GoogleNet或Network in Network等网络架构可以通过采用这一理念在几个基准测试中实现最先进的结果。但是，这些体系结构还没有完全优化网络内部计算资源的利用率。Szegedy等人指出了这个问题。 [62]并激励他们增加网络的深度和宽度，同时保持计算预算不变。

The work in [63] targeted the Residual Network based model with a spatially varying computation time, called stochastic depth, which enabled the seemingly contradictory setup to train short networks and used deep networks at test time. It started with very deep networks, while during training, for each mini-batch, randomly dropped a subset of layers and bypassed them with the identity function. Following this direction, thew work in [64] proposed a pyramidal residual networks with stochastic depth. In [65], Wu et al. proposed an approach that learns to dynamically choose which layers of a deep network to execute during inference so as to best reduce total computation. Veit et al. exploited convolutional networks with adaptive inference graphs to adaptively define their network topology conditioned on the input image [66].
[63]中的工作针对基于剩余网络的模型，其空间变化的计算时间称为随机深度，这使得看似矛盾的设置能够训练短网络并在测试时使用深度网络。它从非常深的网络开始，而在训练期间，对于每个小批量，随机丢弃一部分层并使用身份功能绕过它们。按照这个方向，[64]中的工作提出了具有随机深度的金字塔残差网络。在[65]，吴等人。提出了一种方法，该方法学习动态选择在推理期间执行深层网络的哪些层，以便最好地减少总计算。Veit等人。利用自适应推理图的卷积网络自适应地定义其输入图像条件下的网络拓扑[66]。

Other approaches to reduce the convolutional overheads include using FFT based convolutions [67] and fast convolution using the Winograd algorithm [68]. Zhai et al. [69] proposed a strategy call stochastic spatial sampling pooling, which speedup the pooling operations by a more general stochastic version. Saeedan et al. presented a novel pooling layer for convolutional neural networks termed detail-preserving pooling (DPP), based on the idea of inverse bilateral filters [70]. Those works only aim to speed up the computation but not reduce the memory storage.
减少卷积开销的其他方法包括使用基于FFT的卷积[67]和使用Winograd算法的快速卷积[68]。翟等人。 [69]提出了一种策略调用随机空间采样池，它通过更一般的随机版本加速池化操作。Saeedan等人。基于逆双边滤波器的思想，提出了一种新的汇集层，用于卷积神经网络，称为细节保留池（DPP）[70]。这些工作仅旨在加速计算，但不会减少内存存储。

VII. BENCHMARKS, EVALUATION AND DATABASES
七。基准，评估和数据库

In the past five years the deep learning community had made great efforts in benchmark models. One of the most well-known model used in compression and acceleration for CNNs is Alexnet [1], which has been occasionally used for assessing the performance of compression. Other popular standard models include LeNets [71], All-CNN-nets [72] and many others. LeNet-300-100 is a fully connected network with two hidden layers, with 300 and 100 neurons each. LeNet-5 is a convolutional network that has two convolutional layers and two fully connected layers. Recently, more and more state-of-the-art architectures are used as baseline models in many works, including network in networks (NIN) [73], VGG nets [74] and residual networks (ResNet) [75]. Table IV summarizes the baseline models commonly used in several typical compression methods.
在过去的五年中，深度学习社区在基准模型方面做出了巨大努力。用于CNN压缩和加速的最着名的模型之一是Alexnet [1]，它偶尔用于评估压缩性能。其他流行的标准模型包括LeNets [71]，All-CNN-net [72]等等。LeNet-300-100是一个完全连接的网络，有两个隐藏层，每个层有300个和100个神经元。LeNet-5是一个卷积网络，有两个卷积层和两个完全连接的层。最近，越来越多的最先进的架构被用作许多工作的基线模型，包括网络中的网络（NIN）[73]，VGG网[74]和剩余网络（ResNet）[75]。表IV总结了几种典型压缩方法中常用的基线模型。

The standard criteria to measure the quality of model compression and acceleration are the compression and the speedup rates. Assume that a is the number of the parameters in the original model M and a^* is that of the compressed model M^*, then the compression rate $\alpha(M, M^*)$ of M^* over M is

$$\alpha(M, M^*) = \frac{a}{a^*}. \tag{8}$$

TABLE IV SUMMARIZATION OF BASELINE MODELS USED IN DIFFERENT REPRESENTATIVE WORKS OF NETWORK COMPRESSION.

表IV在网络压缩的不同表示工作中使用的基线模型的概述。

Baseline Models	Representative Works
Alexnet [1]	structural matrix [29], [30], [32] low-rank factorization [40]
Network in network [73]	low-rank factorization [40]
VGG nets [74]	transferred filters [44] low-rank factorization [40]
Residual networks [75]	compact filters [49], stochastic depth [63] parameter sharing [24]
All-CNN-nets [72]	transferred filters [45]
LeNets [71]	parameter sharing [24] parameter pruning [20], [22]

Another widely used measurement is the index space saving defined in several papers [30], [35] as

另一种广泛使用的测量是在几篇论文[30]，[35]中定义的索引空间节省

$$\beta(M, M^*) = \frac{a - a^*}{a^*}, \tag{9}$$

where a and a are the number of the dimension of the index space in the original model and that of the compressed model, respectively.

其中a和a分别是原始模型中索引空间和压缩模型的维数的维数。

Similarly, given the running time s of M and s^* of M^*, the speedup rate $\delta(M, M^*)$ is defined as:

同样，给定M^*的M和s^*的运行时间s，加速率$\delta(M, M^*)$定义为：

$$\delta(M, M^*) = \frac{s}{s^*}. \tag{10}$$

Most work used the average training time per epoch to measure the running time, while in [30], [35], the average testing time was used. Generally, the compression rate and speedup rate are highly correlated, as smaller models often results in faster computation for both the training and the testing stages.

大多数工作使用每个时期的平均训练时间来测量运行时间，而在[30]，[35]中，使用了平均测试时间。通常，压缩率和加速率是高度相关的，因为较小的模型通常导致训练和测试阶段的更快计算。

Good compression methods are expected to achieve almost the same performance as the original model with much smaller parameters and less computational time. However, for different applications with different CNN designs, the relation between parameter size and computational time may be different. For example, it is observed that for deep CNNs with fully connected layers, most of the parameters are in the fully connected layers; while for image classification tasks, float point operations are mainly in the first few convolutional layers since each filter is convolved with the whole image, which is usually very large at the beginning. Thus compression and acceleration of the network should focus on different type of layers for different applications.
预期良好的压缩方法可以获得与原始模型几乎相同的性能，具有更小的参数和更少的计算时间。然而，对于具有不同CNN设计的不同应用，参数大小和计算时间之间的关系可能不同。例如，观察到对于具有完全连接层的深CNN，大多数参数在完全连接的层中;而对于图像分类任务，浮点运算主要在最初的几个卷积层中，因为每个滤波器与整个图像卷积，这通常在开始时非常大。因此，网络的压缩和加速应该集中在不同类型的层上以用于不同的应用。

VIII. DISCUSSION AND CHALLENGES
八。讨论和挑战

In this paper, we summarized recent efforts on compressing and accelerating deep neural networks (DNNs). Here we discuss more details about how to choose different compression approaches, and possible challenges/solutions on this area.
在本文中，我们总结了最近在压缩和加速深度神经网络（DNN）方面的努力。在这里，我们将讨论有关如何选择不同压缩方法的更多细节，以及该领域可能的挑战/解决方案。

A. General Suggestions
A.一般建议

There is no golden rule to measure which approach is the best. How to choose the proper method is really depending on the applications and requirements. Here are some general guidance we can provide:
没有黄金法则可以衡量哪种方法最好。如何选择合适的方法实际上取决于应用和要求。以下是我们可以提供的一般指导：

• If the applications need compacted models from pretrained models, you can choose either pruning & sharing
•如果应用程序需要预训练模型的压缩模型，您可以选择修剪和共享

or low rank factorization based methods. If you need end-to-end solutions for your problem, the low rank and transferred convolutional filters approaches could be considered.
或基于低秩分解的方法。如果您需要针对您的问题的端到端解决方案，可以考虑低级别和传输的卷积滤波器方法。

• For applications in some specific domains, methods with human prior (like the transferred convolutional filters, structural matrix) sometimes have benefits. For example, when doing medical images classification, transferred convolutional filters could work well as medical images (like organ) do have the rotation transformation property. • Usually the approaches of pruning & sharing could give reasonable compression rate while not hurt the accuracy. Thus for applications which requires stable model accuracy, it is better to utilize pruning & sharing.
•对于某些特定领域的应用，具有人类先验的方法（如转移的卷积滤波器，结构矩阵）有时会带来好处。例如，当进行医学图像分类时，传输的卷积滤波器可以很好地工作，因为医学图像（如器官）确实具有旋转变换特性。 •通常修剪和共享的方法可以提供合理的压缩率，同时不会损害准确性。因此，对于需要稳定模型精度的应用，最好利用修剪和共享。

• If your problem involves small/medium size datasets, you can try the knowledge distillation approaches. The compressed student model can take the benefit of transferring knowledge from teacher model, making it robust datasets which are not large.

•如果您的问题涉及中小型数据集，您可以尝试知识蒸馏方法。压缩的学生模型可以利用从教师模型转移知识的好处，使其成为不大的强大数据集。

• As we mentioned before, techniques of the four groups are orthogonal. It is reasonable to combine two or three of them to maximize the performance. For some specific applications, like object detection, which requires both convolutional and fully connected layers, you can compress the convolutional layers with low rank based method and the fully connected layers with a pruning technique.

•正如我们之前提到的，这四组技术是正交的。将两个或三个组合起来以最大化性能是合理的。对于某些特定应用，例如需要卷积和完全连接层的物体检测，您可以使用基于低秩的方法压缩卷积层，使用修剪技术压缩完全连接的层。

B. Technique Challenges

B.技术挑战

Techniques for deep model compression and acceleration are still in the early stage and the following challenges still need to be addressed.

深模型压缩和加速的技术仍处于早期阶段，仍然需要解决以下挑战。

• Most of the current state-of-the-art approaches are built on well-designed CNN models, which have limited freedom to change the configuration (e.g., network structural, hyper-parameters). To handle more complicated tasks, it should provide more plausible ways to configure the compressed models.

•大多数当前最先进的方法都建立在精心设计的CNN模型上，这些模型在改变配置方面具有有限的自由度（例如，网络结构，超参数）。为了处理更复杂的任务，它应该提供更合理的方法来配置压缩模型。

• Pruning is an effective way to compress and accelerate CNNs. The current pruning techniques are mostly designed to eliminate connections between neurons. On the other hand, pruning channel can directly reduce the feature map width and shrink the model into a thinner one. It is efficient but also challenging because removing channels might dramatically change the input of the following layer.

•修剪是压缩和加速CNN的有效方法。目前的修剪技术主要用于消除神经元之间的连接。另一方面，修剪通道可以直接减少特征图宽度并将模型缩小为更薄的模型。它是有效的，但也具有挑战性，因为删除通道可能会显着改变下一层的输入。

• As we mentioned before, methods of structural matrix and transferred convolutional filters impose prior human knowledge to the model, which could significantly affect the performance and stability. It is critical to investigate how to control the impact of those prior knowledge.

•正如我们之前提到的，结构矩阵和转移卷积滤波器的方法将先前的人类知识强加给模型，这可能对性能和稳定性产生重大影响。研究如何控制这些先验知识的影响至关重要。

• The methods of knowledge distillation provide many benefits such as directly accelerating model without special hardware or implementations. It is still worthy developing KD-based approaches and exploring how to improve their performances.

•知识蒸馏的方法提供了许多好处，例如在没有特殊硬件或实现的情况下直接加速模型。仍然值得开发基于KD的方法并探索如何提高他们的表现。

• Hardware constraints in various of small platforms (e.g., mobile, robotic, self-driving car) are still a major problem

•各种小型平台（例如，移动，机器人，自动驾驶汽车）中的硬件限制仍然是主要问题

to hinder the extension of deep CNNs. How to make full use of the limited computational source and how to design special compression methods for such platforms are still challenges that need to be addressed.

阻碍CNN深度扩展。如何充分利用有限的计算源以及如何为这样的平台设计特殊的压缩方法仍然是需要解决的挑战。

• Despite the great achievements of these compression approaches, the black box mechanism is still the key barrier to the adoption. Exploring the knowledge interpret-ability is still an important problem.
•尽管这些压缩方法取得了巨大成就，但黑盒机制仍然是采用这种方法的关键障碍。探索知识解释能力仍然是一个重要问题。

C. Possible Solutions
C.可能的解决方案

To solve the hyper-parameters configuration problem, we can rely on the recent learning-to-learn strategies [76], [77]. This framework provides a mechanism allowing the algorithm to automatically learn how to exploit structure in the problem of interest. Very recently, leveraging reinforcement learning to efficiently sample the design space and improve the model compression has also been tried [78].
为了解决超参数配置问题，我们可以依靠最近的学习 - 学习策略[76]，[77]。该框架提供了一种机制，允许算法自动学习如何在感兴趣的问题中利用结构。最近，还尝试利用强化学习来有效地对设计空间进行采样并改进模型压缩[78]。

Channel pruning provides the efficiency benefit on both CPU and GPU because no special implementation is required. But it is also challenging to handle the input configuration. One possible solution is to use the training-based channel pruning methods [79], which focus on imposing sparse constraints on weights during training. However, training from scratch for such method is costly for very deep CNNs. In [80], the authors provided an iterative two-step algorithm to effectively prune channels in each layer.
通道修剪可在CPU和GPU上提供高效优势，因为无需特殊实施。但处理输入配置也很具挑战性。一种可能的解决方案是使用基于训练的频道修剪方法[79]，其侧重于在训练期间对权重施加稀疏约束。然而，对于非常深的CNN，从头开始训练这种方法是昂贵的。在[80]中，作者提供了一种迭代的两步算法来有效地修剪每层中的信道。

Exploring new types of knowledge in the teacher models and transferring it to the student models is useful for the knowledge distillation (KD) approaches. Instead of directly reducing and transferring parameters, passing selectivity knowledge of neurons could be helpful. One can derive a way to select essential neurons related to the task [81], [82]. The intuition is that if a neuron is activated in certain regions or samples, that implies these regions or samples share some common properties that may relate to the task.
在教师模型中探索新类型的知识并将其转移到学生模型对于知识蒸馏（KD）方法是有用的。而不是直接减少和传递参数，传递神经元的选择性知识可能是有帮助的。人们可以推导出一种选择与任务相关的基本神经元的方法[81]，[82]。直觉是如果在某些区域或样本中激活神经元，则意味着这些区域或样本共享可能与任务相关的一些共同属性。

For methods with the convolutional filters and the structural matrix, we can conclude that the transformation lies in the family of functions that only operations on the spatial dimensions. Hence to address the imposed prior issue, one solution is to provide a generalization of the aforementioned approaches in two aspects: 1) instead of limiting the transformation to belong to a set of predefined transformations, let it be the whole family of spatial transformations applied on 2D filters or matrix, and 2) learn the transformation jointly with all the model parameters.
对于卷积滤波器和结构矩阵的方法，我们可以得出这样的结论:变换属于只对空间维度进行运算的函数族。因此，为了解决强加的先前问题，一种解决方案是在两个方面提供上述方法的概括：1）不是将变换限制为属于一组预定变换，而是将其应用于整个空间变换族。 2D过滤器或矩阵，以及2）与所有模型参数一起学习变换。

Regarding the use of CNNs in small platforms, proposing some general/unified approaches is one direction. Wang et al.
关于在小型平台中使用CNN，提出一些通用/统一方法是一个方向。王等人。

[83] presented a feature map dimensionality reduction method by excavating and removing redundancy in feature maps generated from different filters, which could also preserve intrinsic information of the original network. The idea can be applied to make CNNs more applicable for different platforms. The work in [84] proposed a one-shot whole network compression scheme consisting of three components: rank selection, lowrank tensor decomposition, and fine-tuning to make deep CNNs work in mobile devices.
[83]通过挖掘和去除从不同过滤器生成的特征图中的冗余来提出特征图降维方法，这也可以保留原始网络的内在信息。该想法可用于使CNN更适用于不同平台。[84]中的工作提出了一次性全网络压缩方案，包括三个组件：秩选择，低冲张量分解和微调，以使深CNN在移动设备中工作。

Despite the classification task, people are also adapting the compacted models in other tasks [85]–[87]. We would like to see more work for applications with larger deep nets (e.g., video and image frames [88], [89]).
尽管有分类任务，但人们也在其他任务中调整压缩模型[85] - [87]。我们希望看到更多深层网络应用的工作（例如视频和图像帧[88]，[89]）。

IX. ACKNOWLEDGMENTS
IX。致谢

The authors would like to thank the reviewers and broader community for their feedback on this survey. In particular, we would like to thank Hong Zhao from the Department of Automation of Tsinghua University for her help on modifying the paper. This research is supported by National Science Foundation of China with Grant number 61401169.
作者要感谢审稿人和更广泛的社群对本次调查的反馈。特别是，我们要感谢清华大学自动化系的赵红，感谢她对修改论文的帮助。该研究得到了国家自然科学基金的资助，编号为61401169。

REFERENCES
参考

[1] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks,” in NIPS, 2012.
[1] A. Krizhevsky，I。Sutskever和G. Hinton，“深度卷积神经网络的Imagenet分类”，NIPS，2012年。

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in CVPR, 2014.
[2] Y. Taigman，M。Yang，M。Ranzato和L. Wolf，“Deepface：在人脸检验中弥合人类绩效的差距”，CVPR，2014年。

[3] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, and R. S. Feris, “Fullyadaptive feature sharing in multi-task networks with applications in person attribute classification,” CoRR, vol. abs/1611.05377, 2016.
[3] Y. Lu，A。Kumar，S。Zhai，Y。Cheng，T。Javidi和R. S. Feris，“具有应用程序的多任务网络中的完全自适应特征共享属性分类”，CoRR，第一卷。 abs / 1611.05377,2016。

[4] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. Le, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Ng, “Large scale distributed deep networks,” in NIPS, 2012.
[4] J. Dean，G。Corrado，R。Monga，K。Chen，M。Devin，Q。Le，M。Mao，M。Ranzato，A。Senior，P。Tucker，K。Yang，and A. Ng，“大规模分布式深度网络”，NIPS，2012年。

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” CoRR, vol. abs/1512.03385, 2015.
[5] K. He，X。Zhang，S。Ren和J. Sun，“用于图像识别的深度残留学习”，CoRR，第一卷。 abs / 1512.03385,2015。

[6] Y. Gong, L. Liu, M. Yang, and L. D. Bourdev, “Compressing deep convolutional networks using vector quantization,” CoRR, vol. abs/1412.6115, 2014.
[6] Y. Gong，L。Liu，M。Yang和L. D. Bourdev，“使用矢量量化压缩深度卷积网络”，CoRR，vol。 abs / 1412.6115,2014。

[7] Y. W. Q. H. Jiaxiang Wu, Cong Leng and J. Cheng, “Quantized convolutional neural networks for mobile devices,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[7] Y. W. Q. H. Jiaxiang Wu，Cong Leng和J. Cheng，“用于移动设备的量化卷积神经网络”，IEEE计算机视觉和模式识别会议（CVPR），2016年。

[8] V. Vanhoucke, A. Senior, and M. Z. Mao, “Improving the speed of neural networks on cpus,” in Deep Learning and Unsupervised Feature Learning Workshop, NIPS 2011, 2011.
[8] V. Vanhoucke，A。Senior和M. Z. Mao，“提高神经网络在cpus上的速度”，在深度学习和无监督特征学习研讨会上，NIPS 2011,2011。

[9] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, “Deep learning with limited numerical precision,” in Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, ser. ICML’15, 2015, pp. 1737–1746.
[9] S. Gupta，A。Agrawal，K。Gopalakrishnan和P. Narayanan，“深度学习，数学精度有限”，载于第32届国际机器学习会议论文集 - 第37卷，ser。 ICML'15,2015，pp.1737-1746。

[10] S. Han, H. Mao, and W. J. Dally, “Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding,” International Conference on Learning Representations (ICLR), 2016.
[10] S. Han，H。Mao和W. J. Dally，“深度压缩：压缩深度神经网络与修剪，训练量化和霍夫曼编码”，国际学习表示会议（ICLR），2016年。

[11] Y. Choi, M. El-Khamy, and J. Lee, “Towards the limit of network quantization,” CoRR, vol. abs/1612.01543, 2016.
[11] Y. Choi，M。El-Khamy和J. Lee，“迈向网络量化的极限”，CoRR，第一卷。 abs / 1612.01543,2016。

[12] M. Courbariaux, Y. Bengio, and J. David, “Binaryconnect: Training deep neural networks with binary weights during propagations,” in Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, 2015, pp. 3123–3131.
[12] M. Courbariaux，Y。Bengio和J. David，“Binaryconnect：在传播期间训练具有二进制权重的深度神经网络”，神经信息处理系统进展28：神经信息处理系统年会2015，12月7日-12,2015，Montreal，Quebec，Canada，2015，pp.3123-3131。

[13] M. Courbariaux and Y. Bengio, “Binarynet: Training deep neural networks with weights and activations constrained to +1 or -1,” CoRR, vol. abs/1602.02830, 2016.
[13] M. Courbariaux和Y. Bengio，“Binarynet：训练深度神经网络，权重和激活约束为+1或-1，”CoRR，vol。 abs / 1602.02830,2016。

[14] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, “Xnor-net: Imagenet classification using binary convolutional neural networks,” in ECCV, 2016.
[14] M. Rastegari，V。Ordonez，J。Redmon和A. Farhadi，“Xnor-net：使用二进制卷积神经网络的Imagenet分类”，在ECCV，2016年。

[15] P. Merolla, R. Appuswamy, J. V. Arthur, S. K. Esser, and D. S. Modha, “Deep neural networks are robust to weight binarization and other nonlinear distortions,” CoRR, vol. abs/1606.01981, 2016.

[15] P. Merolla，R。Appuswamy，J。V.Arthur，S.K. Esser和D. S. Modha，“深度神经网络对于重量二值化和其他非线性失真具有鲁棒性”，CoRR，vol。 abs / 1606.01981,2016。

[16] L. Hou, Q. Yao, and J. T. Kwok, “Loss-aware binarization of deep networks,” CoRR, vol. abs/1611.01600, 2016.
[16] L. Hou，Q。Yao和J. T. Kwok，“深度网络的损失感知二值化”，CoRR，vol。 abs / 1611.01600,2016。

[17] Z. Lin, M. Courbariaux, R. Memisevic, and Y. Bengio, “Neural networks with few multiplications,” CoRR, vol. abs/1510.03009, 2015.
[17] Z. Lin，M。Courbariaux，R。Memisevic和Y. Bengio，“几乎没有乘法的神经网络”，CoRR，vol。 abs / 1510.03009,2015。

[18] S. J. Hanson and L. Y. Pratt, “Comparing biases for minimal network construction with back-propagation,” in Advances in Neural Information Processing Systems 1, D. S. Touretzky, Ed., 1989, pp. 177–185.
[18] S.J.Hanson和L.Y.Pratt，“将最小网络构建的偏差与反向传播相比较”，在Advances in Neural Information Processing Systems 1，D. S. Touretzky，Ed。，1989，pp.177-185中。

[19] Y. L. Cun, J. S. Denker, and S. A. Solla, “Advances in neural information processing systems 2,” D. S. Touretzky, Ed., 1990, ch. Optimal Brain Damage, pp. 598–605.
[19] Y. L. Cun，J。S. Denker和S. A. Solla，“Advances in neural information processing systems 2”，D。S. Touretzky，Ed。，1990，ch。最佳脑损伤，第598-605页。

[20] B. Hassibi, D. G. Stork, and S. C. R. Com, “Second order derivatives for network pruning: Optimal brain surgeon,” in Advances in Neural Information Processing Systems 5. Morgan Kaufmann, 1993, pp. 164– 171.
[20] B. Hassibi，D.G。Stork和S. C. R. Com，“网络修剪的二阶导数：最佳脑外科医生”，神经信息处理系统的进展5. Morgan Kaufmann，1993，pp.164-171。

[21] S. Srinivas and R. V. Babu, “Data-free parameter pruning for deep neural networks,” in Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015, 2015, pp. 31.1–31.12.
[21] S. Srinivas和RV Babu，“深度神经网络的无数据参数修剪”，2015年英国机器视觉会议论文集，BMVC 2015，英国斯旺西，2015年9月7日至10日，2015年。 31.1-31.12。

[22] S. Han, J. Pool, J. Tran, and W. J. Dally, “Learning both weights and connections for efficient neural networks,” in Proceedings of the 28th International Conference on Neural Information Processing Systems, ser. NIPS’15, 2015.
[22] S. Han，J。Pool，J。Tran和W. J. Dally，“学习有效神经网络的权重和连接”，载于第28届神经信息处理系统国际会议论文集，ser。 NIPS'15,2015。

[23] W. Chen, J. Wilson, S. Tyree, K. Q. Weinberger, and Y. Chen, “Compressing neural networks with the hashing trick.” JMLR Workshop and Conference Proceedings, 2015.
[23] W. Chen，J。Wilson，S。Tyree，K。Q. Weinberger和Y. Chen，“用散列技巧压缩神经网络。”JMLR研讨会和会议论文集，2015年。

[24] K. Ullrich, E. Meeds, and M. Welling, “Soft weight-sharing for neural network compression,” CoRR, vol. abs/1702.04008, 2017.
[24] K. Ullrich，E。Meeds和M. Welling，“用于神经网络压缩的软重量共享”，CoRR，第一卷。 abs / 1702.04008,2017。

[25] V. Lebedev and V. S. Lempitsky, “Fast convnets using group-wise brain damage,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 2016, pp. 2554–2564.

[25] V. Lebedev和VS Lempitsky，“使用群体性脑损伤的快速网络”，2016年IEEE计算机视觉和模式识别会议，CVPR 2016，美国内华达州拉斯维加斯，2016年6月27日至30日，2016年，第2554-2564页。

[26] H. Zhou, J. M. Alvarez, and F. Porikli, “Less is more: Towards compact cnns,” in European Conference on Computer Vision, Amsterdam, the Netherlands, October 2016, pp. 662–677.
[26] H. Zhou，J。M. Alvarez和F. Porikli，“少即是多：走向紧凑型”，参加欧洲计算机视觉会议，荷兰阿姆斯特丹，2016年10月，第662-677页。

[27] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li, “Learning structured sparsity in deep neural networks,” in Advances in Neural Information Processing Systems 29, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 2074–2082.
[27] W. Wen，C。Wu，Y。Wang，Y。Chen和H. Li，“学习深层神经网络中的结构化稀疏性”，神经信息处理系统进展29，DD Lee，M。Sugiyama， UV Luxburg，I。Guyon和R. Garnett，Eds。，2016，pp.2074-2082。

[28] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, “Pruning filters for efficient convnets,” CoRR, vol. abs/1608.08710, 2016.
[28] H. Li，A。Kadav，I。Durdanovic，H。Samet和H. P. Graf，“修剪有效网络的过滤器”，CoRR，第一卷。 abs / 1608.08710,2016。

[29] V. Sindhwani, T. Sainath, and S. Kumar, “Structured transforms for small-footprint deep learning,” in Advances in Neural Information Processing Systems 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., 2015, pp. 3088–3096.
[29] V. Sindhwani，T。Sainath和S. Kumar，“小型深度学习的结构转换”，神经信息处理系统进展28，C。Cortes，ND Lawrence，DD Lee，M。Sugiyama，和R. Garnett，Eds。，2015，pp.3088-3096。

[30] Y. Cheng, F. X. Yu, R. Feris, S. Kumar, A. Choudhary, and S.-F. Chang, “An exploration of parameter redundancy in deep networks with circulant projections,” in International Conference on Computer Vision (ICCV), 2015.
[30] Y. Cheng，F。X. Yu，R。Feris，S。Kumar，A。Choudhary和S.-F. Chang，“在具有循环预测的深度网络中参数冗余的探索”，在2015年计算机视觉国际会议（ICCV）上。

[31] Y. Cheng, F. X. Yu, R. S. Feris, S. Kumar, A. N. Choudhary, and S. Chang, “Fast neural networks with circulant projections,” CoRR, vol. abs/1502.03436, 2015.
[31] Y. Cheng，F。X. Yu，R。S. Feris，S。Kumar，A。N. Choudhary和S. Chang，“Fast neural networks with circulant projections，”CoRR，vol。 abs / 1502.03436,2015。

[32] Z. Yang, M. Moczulski, M. Denil, N. de Freitas, A. Smola, L. Song, and Z. Wang, “Deep fried convnets,” in International Conference on Computer Vision (ICCV), 2015.
[32] Z. Yang，M。Moczulski，M。Denil，N。de Freitas，A。Smola，L。Song和Z. Wang，“Deep fried convnets”，在国际计算机视觉会议（ICCV），2015年。

[33] J. Chun and T. Kailath, Generalized Displacement Structure for BlockToeplitz, Toeplitz-block, and Toeplitz-derived Matrices. Berlin, Heidelberg: Springer Berlin Heidelberg, 1991, pp. 215–236.
[33] J. Chun和T. Kailath，BlockToeplitz，Toeplitz-block和Toeplitz-derived Matrices的广义位移结构。柏林，海德堡：施普林格柏林海德堡，1991年，页。 215-236。

[34] M. V. Rakhuba and I. V. Oseledets, “Fast multidimensional convolution in low-rank tensor formats via cross approximation,” SIAM J. Scientific Computing, vol. 37, no. 2, 2015.

[34] M. V. Rakhuba和I. V. Oseledets，“通过交叉近似在低秩张量格式中快速多维卷积”，SIAM J. Scienti fi c Computing，vol。 37，不。 2015年2月2日。

[35] M. Moczulski, M. Denil, J. Appleyard, and N. de Freitas, “Acdc: A structured efficient linear layer,” in International Conference on Learning Representations (ICLR), 2016.
[35] M. Moczulski，M。Denil，J。Appleyard和N. de Freitas，“Acdc：一个结构化的高效线性层”，参加2016年国际学习代表大会（ICLR）。

[36] R. Rigamonti, A. Sironi, V. Lepetit, and P. Fua, “Learning separable filters,” in 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013, 2013, pp. 2754– 2761.
[36] R. Rigamonti，A。Sironi，V。Lepetit和P. Fua，“学习可分离滤波器”，2013年IEEE计算机视觉和模式识别会议，美国俄勒冈州波特兰，2013年6月23日至28日， 2013年，第2754-276页。

[37] 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 27, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., 2014, pp. 1269–1277.
[37] EL Denton，W。Zaremba，J。Bruna，Y。LeCun和R. Fergus，“利用卷积网络中的线性结构进行有效评估”，神经信息处理系统进展27，Z。Ghahramani，M。 Welling，C。Cortes，ND Lawrence和KQ Weinberger，Eds。，2014，pp.1269-1277。

[38] M. Jaderberg, A. Vedaldi, and A. Zisserman, “Speeding up convolutional neural networks with low rank expansions,” in Proceedings of the British Machine Vision Conference. BMVA Press, 2014.
[38] M. Jaderberg，A。Vedaldi和A. Zisserman，“加速低级扩展的卷积神经网络”，在英国机器视觉会议论文集中。BMVA出版社，2014年。

[39] V. Lebedev, Y. Ganin, M. Rakhuba, I. V. Oseledets, and V. S. Lempitsky, “Speeding-up convolutional neural networks using fine-tuned cpdecomposition,” CoRR, vol. abs/1412.6553, 2014.
[39] V. Lebedev，Y。Ganin，M。Rakhuba，I。V. Oseledets和V. S. Lempitsky，“使用微调cpdecomposition的加速卷积神经网络”，CoRR，vol。 abs / 1412.6553,2014。

[40] C. Tai, T. Xiao, X. Wang, and W. E, “Convolutional neural networks with low-rank regularization,” vol. abs/1511.06067, 2015.
[40] C. Tai，T。Xiao，X。Wang和W. E，“具有低秩正则化的卷积神经网络”，第一卷。 abs / 1511.06067,2015。

[41] M. Denil, B. Shakibi, L. Dinh, M. Ranzato, and N. D. Freitas, “Predicting parameters in deep learning,” in Advances in Neural Information Processing Systems 26, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., 2013, pp. 2148–2156. [Online]. Available: http://media.nips.cc/nipsbooks/nipspapers/paper files/nips26/1053.pdf [42] T. N. Sainath, B. Kingsbury, V. Sindhwani, E. Arisoy, and B. Ramabhadran, “Low-rank matrix factorization for deep neural network training with high-dimensional output targets,” in in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2013.
[41] M. Denil，B。Shakibi，L。Dinh，M。Ranzato和ND Freitas，“预测深度学习中的参数”，神经信息处理系统的进展26，C。Burges，L。Bottou，M。 Welling，Z。Ghahramani和K. Weinberger，Eds。，2013，pp.2148-2156。 [线上]。可用：http：//media.nips.cc/nipsbooks/nipspapers/paper fi les / nips26 / 1053.pdf [42] TN Sainath，B。Kingsbury，V。Sindhwani，E。Arisoy和B. Ramabhadran，“Low-具有高维输出目标的深度神经网络训练的秩矩阵分解，“in Proc。 IEEE Int。 CONF。关于声学，语音和信号处理，2013年。

[43] T. S. Cohen and M. Welling, “Group equivariant convolutional networks,” arXiv preprint arXiv:1602.07576, 2016.
[43] T. S. Cohen和M. Welling，“群体等变卷积网络”，arXiv preprint arXiv：1602.07576,2016。

[44] S. Zhai, Y. Cheng, and Z. M. Zhang, “Doubly convolutional neural networks,” in Advances In Neural Information Processing Systems, 2016, pp. 1082–1090.
[44] S. Zhai，Y。Cheng和Z. M. Zhang，“双重卷积神经网络”，“神经信息处理系统进展”，2016年，第1082-1090页。

[45] W. Shang, K. Sohn, D. Almeida, and H. Lee, “Understanding and improving convolutional neural networks via concatenated rectified linear units,” arXiv preprint arXiv:1603.05201, 2016.
[45] W. Shang，K。Sohn，D。Almeida和H. Lee，“通过连接的整形线性单元理解和改进卷积神经网络”，arXiv preprint arXiv：1603.05201,2016。

[46] H. Li, W. Ouyang, and X. Wang, “Multi-bias non-linear activation in deep neural networks,” arXiv preprint arXiv:1604.00676, 2016.
[46] H. Li，W。Ouyang和X.Wang，“深度神经网络中的多偏置非线性激活”，arXiv preprint arXiv：1604.00676,2016。

[47] S. Dieleman, J. De Fauw, and K. Kavukcuoglu, “Exploiting cyclic symmetry in convolutional neural networks,” in Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ser. ICML’16, 2016.
[47] S. Dieleman，J。De Fauw和K. Kavukcuoglu，“在卷积神经网络中利用循环对称性”，载于第33届国际机器学习会议论文集 - 第48卷，ser。 ICML'16,2016。

[48] C. Szegedy, S. Ioffe, and V. Vanhoucke, “Inception-v4, inceptionresnet and the impact of residual connections on learning.” CoRR, vol. abs/1602.07261, 2016.
陈建新，“网络学习与网络学习的关系”，国立台湾师范大学资讯工程学研究所硕士论文。“CoRR, vol. abs/1602.07261, 2016。”

[49] B. Wu, F. N. Iandola, P. H. Jin, and K. Keutzer, “Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving,” CoRR, vol. abs/1612.01051, 2016.
[49] B. Wu，F。N. Iandola，P。H. Jin和K. Keutzer，“Squeezedet：Uni fi ed，小型，低功率完全卷积神经网络，用于自动驾驶的实时物体检测，”CoRR，vol。 abs / 1612.01051,2016。

[50] C. Buciluˇa, R. Caruana, and A. Niculescu-Mizil, “Model compression,” in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD ’06, 2006, pp. 535– 541.
[50] C. Bucilua，R。Caruana和A. Niculescu-Mizil，“模型压缩”，在第12届ACM SIGKDD知识发现和数据挖掘国际会议论文集中。 KDD '06，2006，pp.535-541。

[51] J. Ba and R. Caruana, “Do deep nets really need to be deep?” in Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, 2014, pp. 2654–2662.
[51] J. Ba和R. Caruana，“深网真的需要深入吗？”神经信息处理系统进展27：神经信息处理系统年会2014，2014年12月8日至13日，魁北克蒙特利尔，加拿大，2014年，第2654-2662页。

[52] G. E. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” CoRR, vol. abs/1503.02531, 2015.
[52] G. E. Hinton，O。Vinyals和J. Dean，“在神经网络中提炼知识”，CoRR，第一卷。 abs / 1503.02531,2015。

[53] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, “Fitnets: Hints for thin deep nets,” CoRR, vol. abs/1412.6550, 2014.
[53] A. Romero，N。Ballas，S。E. Kahou，A。Chassang，C。Gatta和Y. Bengio，“Fitnets：Teints for thin deep nets”，CoRR，vol。 abs / 1412.6550,2014。

[54] A. Korattikara Balan, V. Rathod, K. P. Murphy, and M. Welling, “Bayesian dark knowledge,” in Advances in Neural Information Processing Systems 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., 2015, pp. 3420–3428.
[54] A. Korattikara Balan，V。Rathod，KP Murphy和M. Welling，“贝叶斯黑暗知识”，神经信息处理系统进展28，C。Cortes，ND Lawrence，DD Lee，M。Sugiyama，and R. Garnett，Eds。，2015，pp.3420-3428。

[55] P. Luo, Z. Zhu, Z. Liu, X. Wang, and X. Tang, “Face model compression by distilling knowledge from neurons,” in Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA., 2016, pp. 3560–3566.
[55] P. Luo，Z。Zhu，Z。Liu，X。Wang和X. Tang，“通过从神经元中提取知识来压缩面部模型”，载于第三十届AAAI人工智能会议论文集，2月12日至17日，2016，Phoenix，Arizona，USA。，2016，pp.3560-3566。

[56] T. Chen, I. J. Goodfellow, and J. Shlens, “Net2net: Accelerating learning via knowledge transfer,” CoRR, vol. abs/1511.05641, 2015.
[56] T. Chen，I。J. Goodfellow和J. Shlens，“Net2net：通过知识转移加速学习”，CoRR，第一卷。 abs / 1511.05641,2015。

[57] S. Zagoruyko and N. Komodakis, “Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer,” CoRR, vol. abs/1612.03928, 2016.
[57] S. Zagoruyko和N. Komodakis，“更加关注注意力：通过注意力转移改善卷积神经网络的性能”，CoRR，第一卷。 abs / 1612.03928,2016。

[58] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” CoRR, vol. abs/1409.0473, 2014.
[58] D. Bahdanau，K。Cho和Y. Bengio，“神经机器翻译，共同学习调整和翻译”，CoRR，第一卷。 abs / 1409.0473,2014。

[59] A. Almahairi, N. Ballas, T. Cooijmans, Y. Zheng, H. Larochelle, and A. C. Courville, “Dynamic capacity networks,” in Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016, 2016, pp. 2549–2558.
[59] A. Almahairi，N。Ballas，T。Cooijmans，Y。Zheng，H。Larochelle和AC Courville，“动态能力网络”，第33届国际机器学习会议论文集，ICML 2016，纽约市，NY，USA，2016年6月19日至24日，2016年，第2549-2558页。

[60] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, “Outrageously large neural networks: The sparsely-gated mixture-of-experts layer,” 2017.
[60] N. Shazeer，A。Mirhoseini，K。Maziarz，A。Davis，Q。Le，G。Hinton和J. Dean，“Outrageously large neural networks：the sparsely-gated mixture-of-experts layer”， 2017年。

[61] D. Wu, L. Pigou, P. Kindermans, N. D. Le, L. Shao, J. Dambre, and J. Odobez, “Deep dynamic neural networks for multimodal gesture segmentation and recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 8, pp. 1583–1597, 2016.
[61] D. Wu，L。Pigou，P。Kindermans，N。D. Le，L。Shao，J。Dambre和J. Odobez，“Deep dynamic neural networks for multimodal gesture segmentation and recognition，”IEEE Trans。模式肛门。马赫。 Intell。，vol。 38，不。 8，pp.1583-1597,2016。

[62] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Computer Vision and Pattern Recognition (CVPR), 2015.

[62] C. Szegedy，W。Liu，Y。Jia，P。Sermanet，S。Reed，D。Anguelov，D。Erhan，V。Vanhoucke和A. Rabinovich，“在计算机视觉中更深入地研究”和模式识别（CVPR），2015年。

[63] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, Deep Networks with Stochastic Depth, 2016.
[63] G. Huang，Y。Sun，Z。Liu，D。Sedra和K. Q. Weinberger，Deep Networks with Stochastic Depth，2016。

[64] Y. Yamada, M. Iwamura, and K. Kise, “Deep pyramidal residual networks with separated stochastic depth,” CoRR, vol. abs/1612.01230, 2016.
[64] Y. Yamada，M。Iwamura和K. Kise，“具有分离随机深度的深锥体残余网络”，CoRR，vol。 abs / 1612.01230,2016。

[65] Z. Wu, T. Nagarajan, A. Kumar, S. Rennie, L. S. Davis, K. Grauman, and R. Feris, “Blockdrop: Dynamic inference paths in residual networks,” in CVPR, 2018.
[65]吴振中，T. Nagarajan, A. Kumar, S. Rennie, L. S. Davis, K. Grauman, and R. Feris，《区块链:剩余网络中的动态推理路径》，CVPR, 2018。

[66] A. Veit and S. Belongie, “Convolutional networks with adaptive inference graphs,” 2018.
[66] A. Veit和S. Belongie，“具有自适应推理图的卷积网络”，2018年。

[67] M. Mathieu, M. Henaff, and Y. Lecun, Fast training of convolutional networks through FFTs, 2014.
[67] M. Mathieu，M。Henaff和Y. Lecun，通过FFT快速训练卷积网络，2014年。

[68] A. Lavin and S. Gray, “Fast algorithms for convolutional neural networks,” in 2016 IEEE Conference on Computer Vision and Pattern
[68] A. Lavin和S. Gray，“卷积神经网络的快速算法”，2016年IEEE计算机视觉与模式会议

Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 2016, pp. 4013–4021.
认可，CVPR 2016，拉斯维加斯，内华达州，美国，2016年6月27日至30日，2016年，第4013-4021页。

[69] S. Zhai, H. Wu, A. Kumar, Y. Cheng, Y. Lu, Z. Zhang, and R. S. Feris, “S3pool: Pooling with stochastic spatial sampling,” CoRR, vol. abs/1611.05138, 2016.
[69] S. Zhai，H。Wu，A。Kumar，Y。Cheng，Y。Lu，Z。Zhang和R. S. Feris，“S3pool：Pooling with randomstic spatial sampling，”CoRR，vol。 abs / 1611.05138,2016。

[70] F. Saeedan, N. Weber, M. Goesele, and S. Roth, “Detail-preserving pooling in deep networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[70] F. Saeedan，N。Weber，M。Goesele和S. Roth，“深度网络中的细节保留池”，参见IEEE计算机视觉和模式识别会议论文集，2018年。

[71] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” in Proceedings of the IEEE, 1998, pp. 2278–2324.
[71] Y. Lecun，L。Bottou，Y。Bengio和P. Haffner，“基于梯度的学习应用于文档识别”，参见IEEE，1998，pp.2278-2324。

[72] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller, “Striving for simplicity: The all convolutional net,” CoRR, vol. abs/1412.6806, 2014.
[72] J. T. Springenberg，A。Dosovitskiy，T。Brox和M. A. Riedmiller，“努力实现简单：全部卷积网”，CoRR，第一卷。abs / 1412.6806,2014。

[73] M. Lin, Q. Chen, and S. Yan, “Network in network,” in ICLR, 2014.

[73] M. Lin，Q。Chen和S. Yan，“网络中的网络”，ICLR，2014年。

[74] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” CoRR, vol. abs/1409.1556, 2014.
[74] K. Simonyan和A. Zisserman，“用于大规模图像识别的非常深的卷积网络”，CoRR，第一卷。 abs / 1409.1556,2014。

[75] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” arXiv preprint arXiv:1512.03385, 2015.
[75] K. He，X。Zhang，S。Ren和J. Sun，“图像识别的深度残留学习”，arXiv preprint arXiv：1512.03385,2015。

[76] M. Andrychowicz, M. Denil, S. G. Colmenarejo, M. W. Hoffman, D. Pfau, T. Schaul, and N. de Freitas, “Learning to learn by gradient descent by gradient descent,” in Neural Information Processing Systems (NIPS), 2016.
[76] M. Andrychowicz，M。Denil，SG Colmenarejo，MW Hoffman，D。Pfau，T。Schaul和N. de Freitas，“通过梯度下降学习通过梯度下降学习”，神经信息处理系统（NIPS）），2016。

[77] D. Ha, A. Dai, and Q. Le, “Hypernetworks,” in International Conference on Learning Representations 2016, 2016.
[77] D. Ha，A。Dai和Q. Le，“Hypernetworks”，2016年国际学习代表会议。

[78] Y. He, J. Lin, Z. Liu, H. Wang, L.-J. Li, and S. Han, “Amc: Automl for model compression and acceleration on mobile devices,” in The European Conference on Computer Vision (ECCV), September 2018.
[78] Y.他，J。Lin，Z。Liu，H。Wang，L.-J。 Li和S. Han，“Amc：Automl用于移动设备上的模型压缩和加速”，参见欧洲计算机视觉会议（ECCV），2018年9月。

[79] J. M. Alvarez and M. Salzmann, “Learning the number of neurons in deep networks,” pp. 2270–2278, 2016.
[79] J. M. Alvarez和M. Salzmann，“学习深层网络中神经元的数量”，第2270-2278页，2016年。

[80] Y. He, X. Zhang, and J. Sun, “Channel pruning for accelerating very deep neural networks,” in The IEEE International Conference on Computer Vision (ICCV), Oct 2017.
[80] Y. He，X。Zhang和J. Sun，“加速非常深的神经网络的频道修剪”，在IEEE国际计算机视觉会议（ICCV），2017年10月。

[81] Z. Huang and N. Wang, “Data-driven sparse structure selection for deep neural networks,” ECCV, 2018.
[81] Z. Huang和N.Wang，“深度神经网络的数据驱动稀疏结构选择”，ECCV，2018。

[82] Y. Chen, N. Wang, and Z. Zhang, “Darkrank: Accelerating deep metric learning via cross sample similarities transfer,” in Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, 2018, pp. 2852– 2859.
[82] Y. Chen，N。Wang和Z. Zhang，“Darkrank：通过交叉样本相似性转移加速深度度量学习”，载于第三十二届AAAI人工智能会议论文集（AAAI-18），新版美国路易斯安那州奥尔良市，2018年2月2日至7日，2018年，第2852-2859页。

[83] Y. Wang, C. Xu, C. Xu, and D. Tao, “Beyond filters: Compact feature map for portable deep model,” in Proceedings of the 34th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 06–11 Aug 2017, pp. 3703–3711.
[83] Y. Wang，C。Xu，C。Xu和D. Tao，“超越过滤器：便携式深层模型的紧凑特征图”，在第34届国际机器学习会议论文集中。机器学习研究论文集，D。Precup和Y. W. Teh，Eds。，vol。 70.澳大利亚悉尼国际会议中心：PMLR，2017年8月6日至11日，第3703-3711页。

[84] 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,” CoRR, vol. abs/1511.06530, 2015.

[84] Y.-D. Kim，E。Park，S。Yoo，T。Choi，L。Yang和D. Shin，“压缩深度卷积神经网络用于快速和低功率移动应用”，CoRR，vol。 abs / 1511.06530,2015。

[85] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, “Learning efficient object detection models with knowledge distillation,” in Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., 2017, pp. 742–751.
[85] G. Chen，W。Choi，X。Yu，T。Han和M. Chandraker，“用知识蒸馏学习有效物体检测模型”，神经信息处理系统进展30，I。Guyon，UV Luxburg，S。Bengio，H.Wallach，R.Fergus，S.Vishwanathan，and R. Garnett，Eds。，2017，pp.742-751。

[86] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.
[86] M. Sandler，A。Howard，M。Zhu，A。Zhmoginov和L.-C.陈，“Mobilenetv2：倒置残差和线性瓶颈”，参见IEEE计算机视觉和模式识别会议（CVPR），2018年6月。

[87] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy, “Speed/accuracy trade-offs for modern convolutional object detectors,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, 2017, pp. 3296–3297.
[87] J. Huang，V。Rathod，C。Sun，M。Zhu，A。Korattikara，A。Fathi，I。Fischer，Z。Wojna，Y。Song，S。Guadarrama和K. Murphy，“Speed /现代卷积物体探测器的准确性权衡，“2017年IEEE计算机视觉和模式识别会议，CVPR 2017，檀香山，HI，USA，2017年7月21日至26日，2017年，第3296-3297页。

[88] Y. Cheng, Q. Fan, S. Pankanti, and A. Choudhary, “Temporal sequence modeling for video event detection,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2014.
[88] Y. Cheng，Q。Fan，S。Pankanti和A. Choudhary，“用于视频事件检测的时间序列建模”，参见IEEE计算机视觉和模式识别会议（CVPR），2014年6月。

[89] L. Cao, S.-F. Chang, N. Codella, C. V. Cotton, D. Ellis, L. Gong, M. Hill, G. Hua, J. Kender, M. Merler, Y. Mu, J. R. Smith, and F. X. Yu, “Ibm research and columbia university trecvid-2012 multimedia event detection (med), multimedia event recounting (mer), and semantic indexing (sin) systems,” 2012.
[89] L. Cao，S.-F。 Chang，N。Codella，CV Cotton，D。Ellis，L。Gong，M。Hill，G。Hua，J。Kender，M。Merler，Y。Mu，JR Smith，和FX Yu，“Ibm research and columbia university trecvid-2012多媒体事件检测（med），多媒体事件重述（mer）和语义索引（sin）系统，“2012。

Yu Cheng (yu.cheng@microsoft.com) currently is a Researcher at Microsoft. Before that, he was a Research Staff Member at IBM T.J. Watson Research Center. Yu got his Ph.D. from Northwestern University in 2015 and bachelor from Tsinghua University in 2010. His research is about deep learning in general, with specific interests in the deep generative model, model compression, and transfer learning. He regularly serves on the program committees of top-tier AI conferences such as NIPS, ICML, ICLR, CVPR and ACL.
Yu Cheng（yu.cheng@microsoft.com）目前是微软的研究员。在此之前，他是IBM T.J的研究员。沃森研究中心。 Yu获得博士学位。 2015年获得西北大学学士学位，2010年获得清华大学学士学位。他的研究主要涉及深度学习，特别是深层生成模型，模型压缩和转移学习。他经常在顶级AI会议的计划委员会任职，如NIPS，ICML，ICLR，CVPR和ACL。

Duo Wang (d-wang15@mail.tsinghua.edu.cn) received the B.S. degree in automation from the Harbin Institute of Technology, China, in 2015. Currently he is purchasing his Ph.D. degree at the Department of Automation, Tsinghua University, Beijing, P.R. China. Currently his research interests are about deep learning, particularly in few-shot learning and deep generative models. He also works on a lot of applications in computer vision and robotics vision.

Duo Wang（d-wang15@mail.tsinghua.edu.cn）获得了B.S. 2015年获得中国哈尔滨工业大学自动化学位。目前他正在购买他的博士学位。清华大学自动化系，北京，中国。目前，他的研究兴趣是深度学习，特别是在少数学习和深度生成模型中。他还致力于计算机视觉和机器人视觉领域的许多应用。

Pan Zhou (panzhou@hust.edu.cn) is currently an associate professor with School of Electronic Information and Communications, Wuhan, China. He received his Ph.D. in the School of Electrical and Computer Engineering at the Georgia Institute of Technology in 2011. Before that, he received his B.S. degree in the Advanced Class of HUST, and a M.S. degree in the Department of Electronics and Information Engineering from HUST, Wuhan, China, in 2006 and 2008, respectively. His current research interest includes big data analytics and machine learning, security and privacy, and information networks.

潘周（panzhou@hust.edu.cn）目前是中国武汉电子信息与通信学院的副教授。他获得了博士学位。 2011年在佐治亚理工学院电气与计算机工程学院学习。在此之前，他获得了他的学士学位。华中科技大学高级班的学位和硕士学位分别于2006年和2008年在中国武汉华中科技大学电子与信息工程系获得学位。他目前的研究兴趣包括大数据分析和机器学习，安全和隐私以及信息网络。

Tao Zhang (taozhang@mail.tsinghua.edu.cn) obtained his B.S., M.S., and Ph.D. degrees from Tsinghua University, Beijing, China, in 1993, 1995, and 1999, respectively, and another Ph.D. degree from Saga University, Saga, Japan, in 2002, all in control engineering. He is currently a Professor with the Department of Automation, Tsinghua University. He serves the Associate Dean,

School of Information Science and Technology and Head of the Department of Automation. His current research interests include artificial intelligence, robotics, image processing, control theory, and control of spacecraft.

张涛（taozhang@mail.tsinghua.edu.cn）获得了博士学位，硕士学位和博士学位。 1993年，1995年和1999年分别获得中国北京清华大学的学位和另一位博士学位。 2002年毕业于日本佐贺县佐贺大学，全部从事控制工程。他目前是清华大学自动化系教授。他是信息科学与技术学院副院长兼自动化系主任。他目前的研究兴趣包括人工智能，机器人技术，图像处理，控制理论和航天器控制。