



Model Compression and Acceleration for Deep Neural Networks



CONTENTS



研究背景



研究现状



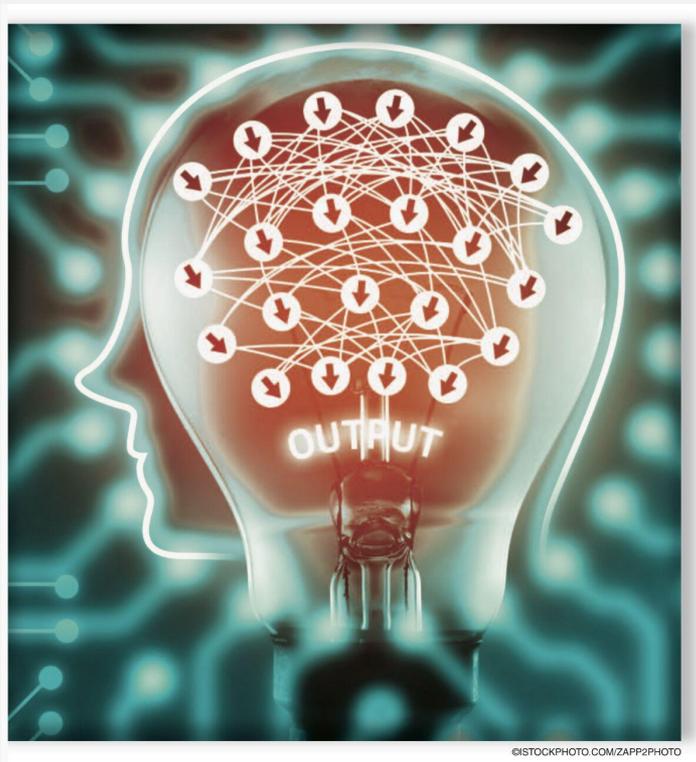
讨论与挑战



神经网络加速

郑板桥在《赠君谋父子》一诗中曾写道，
“删繁就简三秋树，领异标新二月花。”

在人工智能领域，深度神经网络的设计，如同绘制枝蔓繁复的兰竹，需在底层对其删繁就简；而将其拓展至不同场景的应用，则如同面向不同意境的引申，需要创新算法的支撑。



©ISTOCKPHOTO.COM/ZAPP2PHOTO

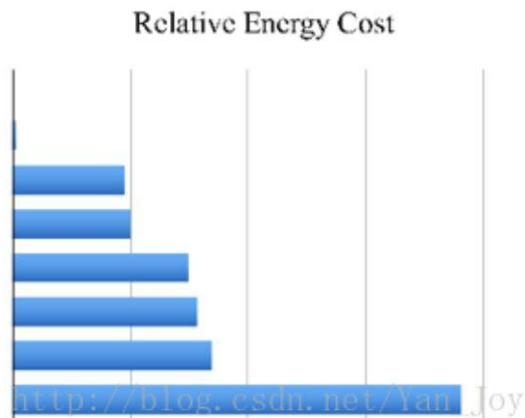


火龙果·整理
uml.org.cn

研究背景和意义



Operation	Energy [pJ]	Relative Cost
32 bit int ADD	0.1	1
32 bit float ADD	0.9	9
32 bit Register File	1	10
32 bit int MULT	3.1	31
32 bit float MULT	3.7	37
32 bit SRAM Cache	5	50
32 bit DRAM Memory	640	6400



研究现状



01

参数修剪和共享

parameter pruning and sharing

02

低秩因子分解

low-rank factorization

03

转移/紧凑卷积滤波器

transferred/compact convolutional filters

04

知识蒸馏

knowledge distillation



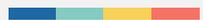
研究现状

Table 1. A summary of different approaches for network compression.

Theme Name	Description	Applications	More Details
Parameter pruning and sharing	Reducing redundant parameters that are not sensitive to the performance	Convolutional layer and fully connected layer	Robust to various settings, can achieve good performance, can support both training from scratch and pretrained model
Low-rank factorization	Using matrix/tensor decomposition to estimate the informative parameters	Convolutional layer and fully connected layer	Standardized pipeline, easily implemented, can support both training from scratch and pretrained model
Transferred/compact convolutional filters	Designing special structural convolutional filters to save parameters	Only for convolutional layer	Algorithms are dependent on applications, usually achieve good performance, only support training from scratch
KD	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 training from scratch



研究现状



方法名称	描述	应用场景	方法细节
剪枝和共享	删除对准确率影响不大的参数	卷积层和全连接层	对不同设置具有鲁棒性，可以达到较好效果，支持从零训练和预训练
低秩分解	使用矩阵对参数进行分解估计	卷积层和全连接层	标准化的途径，很容易实施，支持从零训练和预训练
转移、紧凑卷积核	设计特别的卷积核来保存参数	只有卷积层	算法依赖于应用程序，通常可以取得好的表现，只能从零开始训练
知识蒸馏	训练一个更紧凑的神经网络来从大的模型蒸馏知识	卷积层和全连接层	模型表现对应用程序和网络结构较为敏感，只能从零开始训练



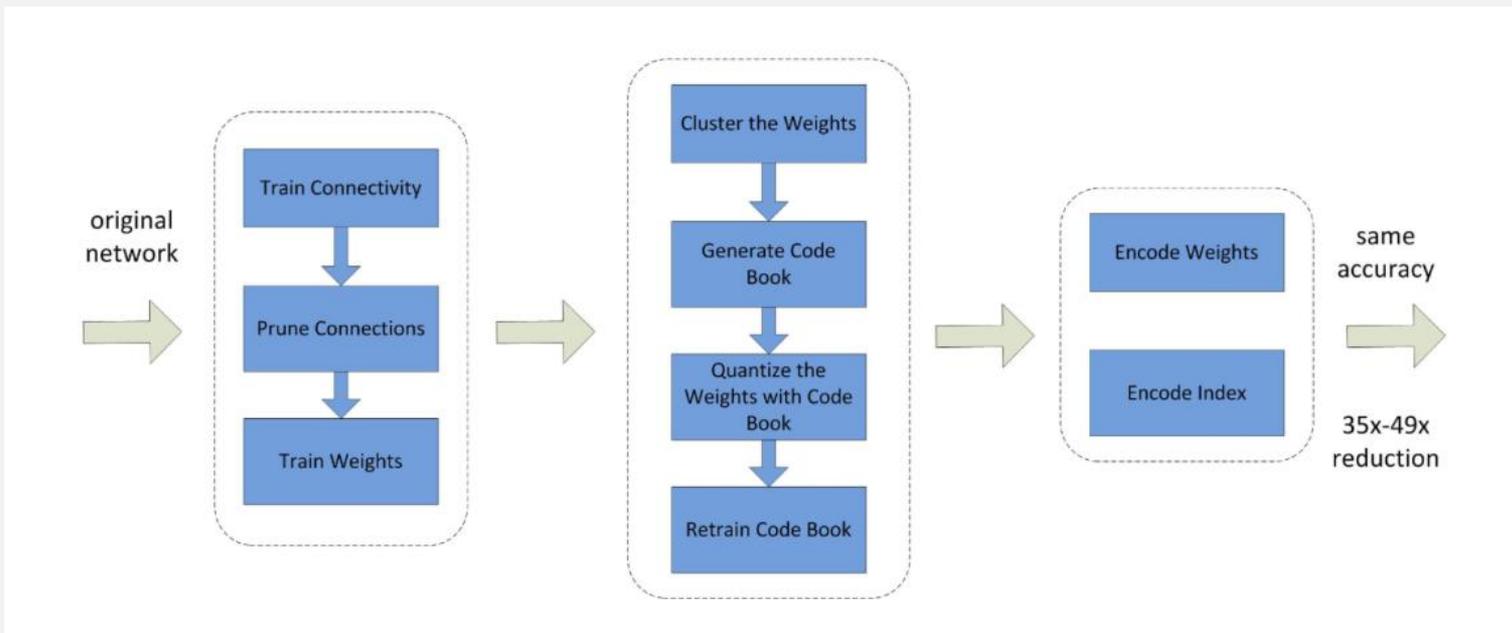
Parameter pruning and sharing



- ✓ 量化和二进制化 (Quantization and Binarization)
- ✓ 剪枝和共享 (Pruning and Sharing)
- ✓ 设计结构化矩阵 (Designing Structural Matrix)

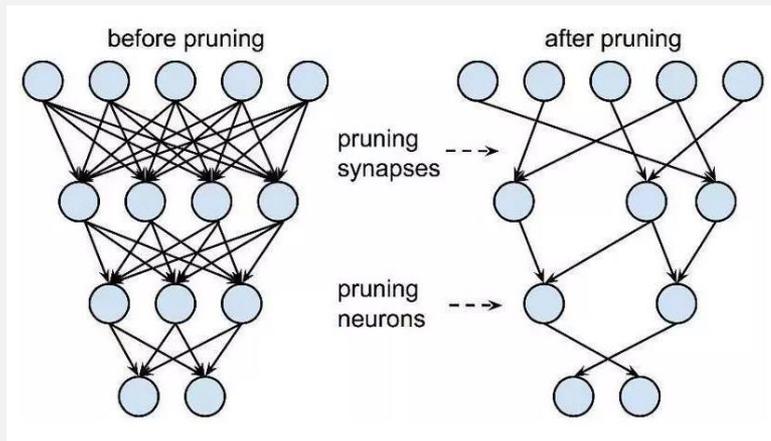
Parameter pruning and sharing

✓ 量化和二进制化 (Quantization and Binarization)



Parameter pruning and sharing

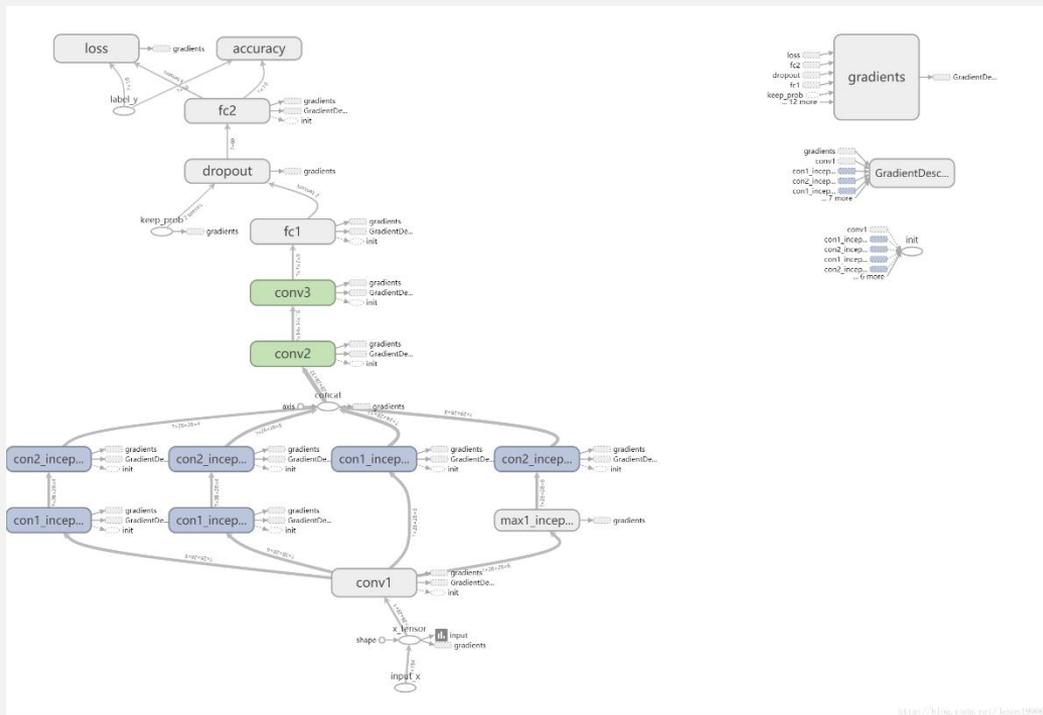
✓ 剪枝和共享 (Pruning and Sharing)



网络剪枝和共享起初是解决过拟合问题的，现在更多得被用于降低网络复杂度。

Parameter pruning and sharing

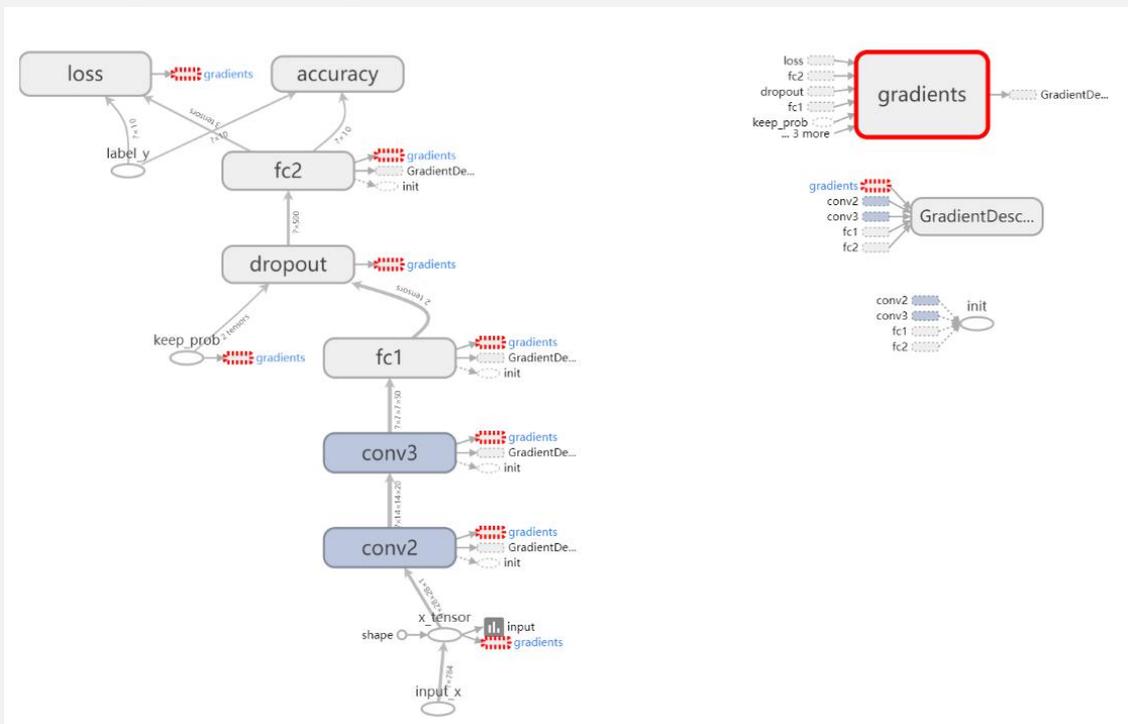
✓ 剪枝和共享 (Pruning and Sharing)



原始神经网络的架构

Parameter pruning and sharing

✓ 剪枝和共享 (Pruning and Sharing)



更改神经网络的架构

Parameter pruning and sharing

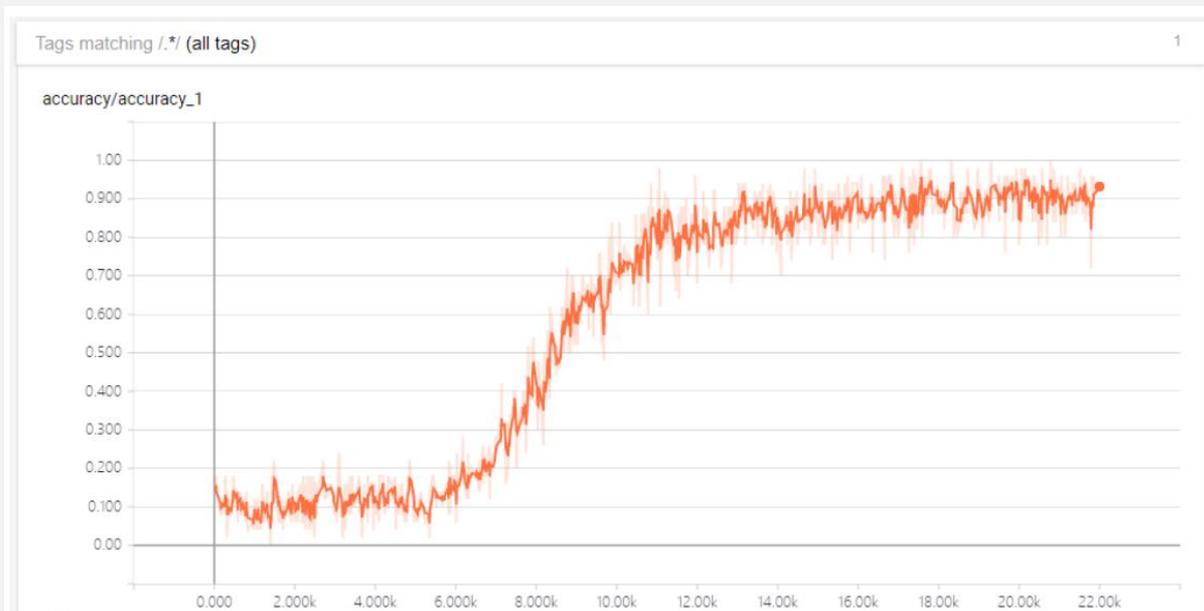
✓ 剪枝和共享 (Pruning and Sharing)



原始网络的准确度曲线

Parameter pruning and sharing

✓ 剪枝和共享 (Pruning and Sharing)



mobile net 的准确度曲线

Parameter pruning and sharing

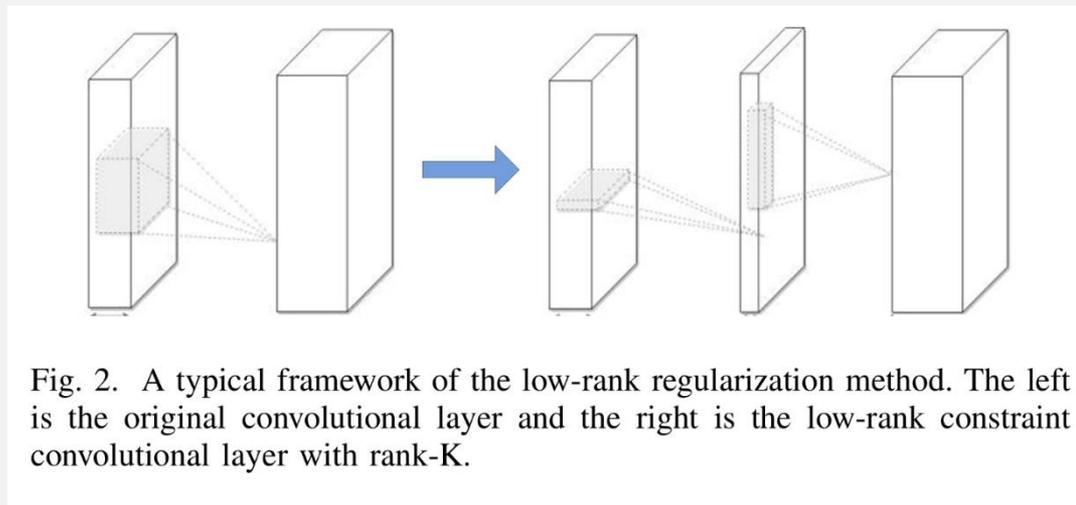
✓ 设计结构化矩阵 (Designing Structural Matrix)

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 $\mathbf{r} = (r_0, r_1, \dots, r_{d-1})$, a circulant matrix $\mathbf{R} \in \mathbb{R}^{d \times d}$ is defined as:

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

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 d -dimensional vector \mathbf{r} , the above 1-layer circulant neural network in Eq. 1 has time complexity of $\mathcal{O}(d \log d)$.

Low-rank factorization and Sparsity



—

Low-rank factorization and Sparsity



TABLE II
COMPARISONS BETWEEN THE LOW-RANK MODELS AND THEIR BASELINES
ON ILSVRC-2012.

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
GoogLeNet	92.21%	1.	1.
BN Low-rank	91.88%	1.08	2.79
CP Low-rank	91.79%	1.20	2.84

低

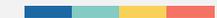
Transferred/compact convolutional filters

该

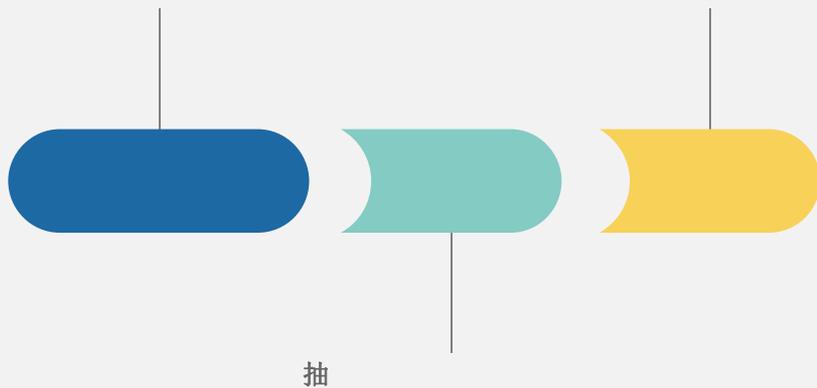
TABLE III
A SIMPLE COMPARISON OF DIFFERENT APPROACHES ON CIFAR-10 AND CIFAR-100.

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

Knowledge distillation



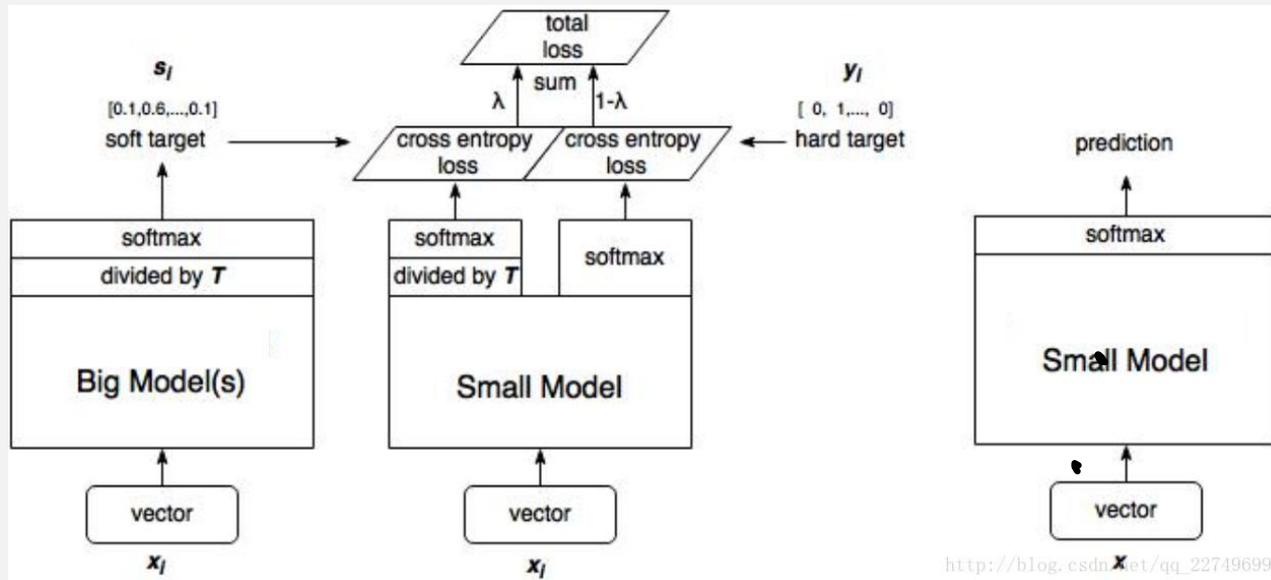
蝴



蝴

学

Knowledge distillation



讨论与挑战



- ✓ 目前大多数的顶尖方法都建立在设计完善的 **CNN** 模型的基础上，这限制了改变配置的自由度（例如，网络结构和超参数）。为了处理更加复杂的任务，还需要更加可靠的模型压缩方法
- ✓ 剪枝是一种压缩和加速 **CNN** 的有效方式。目前大多数的剪枝技术都是以减少神经元之间的连接设计的。另一方面，对通道进行剪枝可以直接减小特征映射的宽度并压缩模型。这很有效，但也存在挑战，因为减少通道会显著地改变下一层的输入。确定这类问题的解决方式同样很重要
- ✓ 多种小型平台（例如，移动设备、机器人、自动驾驶汽车）的硬件限制仍然是阻碍深层 **CNN** 扩展的主要问题。如何全面利用有限的可用计算资源以及如何为这些平台设计特定的压缩方法仍然是个挑战结构化矩阵（ **Designing Structural Matrix** ）





Thank you for watch

