SqueezeNet

2017

ALEXNET-LEVEL ACCURACY WITH 50X FEWER PARAMETERS AND <0.5MB MODEL SIZE

code:https://github.com/DeepScale/SqueezeNet

论文出发点或背景

轻量级网络的三个优点：

1.分布式训练期间，使用轻量级网络可以减少各服务器之间的通信；

2.轻量级网络在云端到设备端导出新设备时使用的带宽更少

3.轻量级网络可以很方便部署在FPGA和其他一些内存受限的设备中

我们工作的首要目标是在保持精度的同时，识别出一个参数很少的模型。为了解决这个问题，一个明智的方法是采用一个现有的CNN模型，并以一种有损的方式压缩它。

论文创新思路

三条设计策略：

1.使用1×1卷积核代替3×3卷积核：，大多数卷积核尺寸使用1×1，1x1卷积核的参数比3x3卷积核少9倍

2.减少3×3卷积核的输入通道数

3.在网络的后期进行下采样，从而使卷积层具有较大的激活图

论文方法介绍

一个Fire模块包括: 一个squeeze层 (只有1x1 卷积), 将其放入一个具有1x1 和3x3 卷积组合的expand层中。在Fire模块中随意使用1x1 过滤器是应用策略1。在一个Fire模块

中有三个超参数: s1x1, e1x1和 e3x3。在Fire模块中, s1x1 是squeeze层 (所有 1x1) 中的过滤器数, e1x1是1x1 卷积在expand层的数量, e3x3 3x3卷积在expand层的数量,。当我们使用Fire模块时, 我们设置 s1x1 小于 (e1x1 e + 3x3), 因此, expand层有助于限制3x3卷积中输入通道的数量即策略 2。

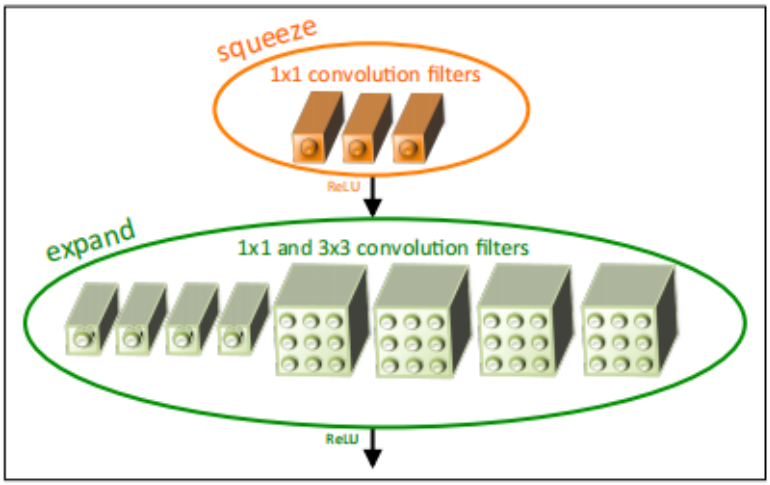


Figure 1: Microarchitectural view: Organization of convolution filters in the **Fire module**. In this example, $s_{1x1} = 3$, $e_{1x1} = 4$, and $e_{3x3} = 4$. We illustrate the convolution filters but not the activations.

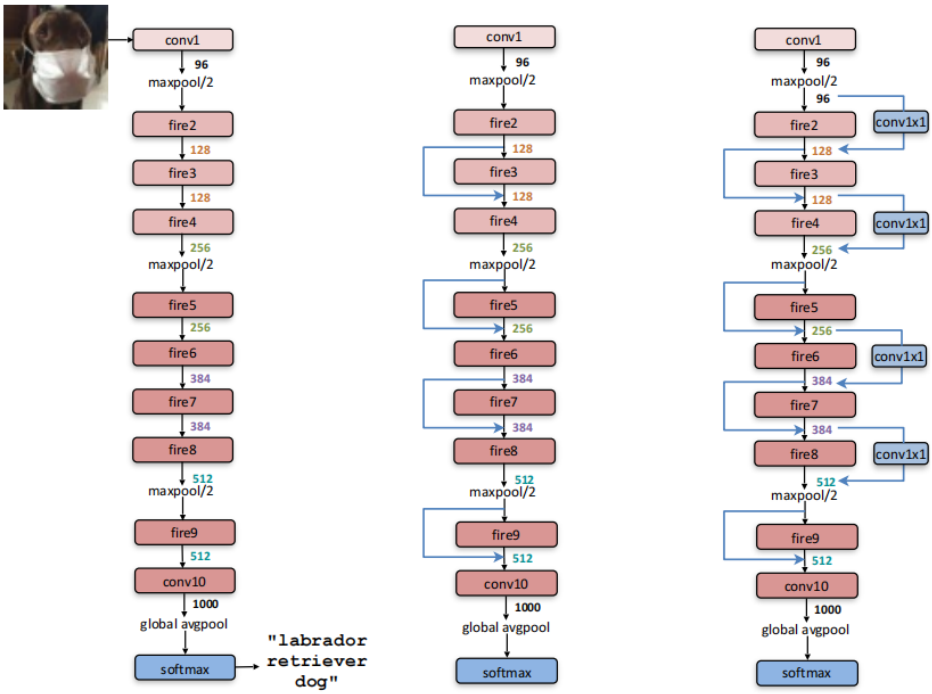


Figure 2: Macroarchitectural view of our SqueezeNet architecture. Left: SqueezeNet (Section 3.3); Middle: SqueezeNet with simple bypass (Section 6); Right: SqueezeNet with complex bypass (Section 6).

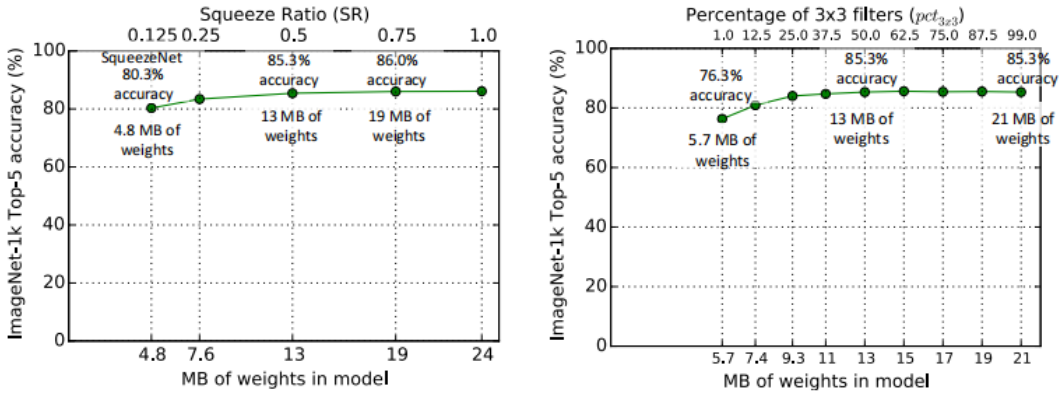
Table 1: SqueezeNet architectural dimensions. (The formatting of this table was inspired by the Inception2 paper (Ioffe & Szegedy, 2015).)

layer name/type	output size	filter size / stride (if not a fire layer)	depth	s_{1x1} (#1x1 squeeze)	e_{1x1} (#1x1 expand)	e_{3x3} (#3x3 expand)	s_{1x1} sparsity	e_{1x1} sparsity	e_{3x3} sparsity	# bits	#parameter before pruning	#parameter after pruning
input image	224x224x3										-	-
conv1	111x111x96	7x7/2 (x96)	1				100% (7x7)			6bit	14,208	14,208
maxpool1	55x55x96	3x3/2	0									
fire2	55x55x128		2	16	64	64	100%	100%	**33%**	6bit	11,920	5,746
fire3	55x55x128		2	16	64	64	100%	100%	**33%**	6bit	12,432	6,258
fire4	55x55x256		2	32	128	128	100%	100%	**33%**	6bit	45,344	20,646
maxpool4	27x27x256	3x3/2	0									
fire5	27x27x256		2	32	128	128	100%	100%	**33%**	6bit	49,440	24,742
fire6	27x27x384		2	48	192	192	100%	**50%**	**33%**	6bit	104,880	44,700
fire7	27x27x384		2	48	192	192	**50%**	100%	**33%**	6bit	111,024	46,236
fire8	27x27x512		2	64	256	256	100%	**50%**	**33%**	6bit	188,992	77,581
maxpool8	13x12x512	3x3/2	0									
fire9	13x13x512		2	64	256	256	**50%**	100%	**30%**	6bit	197,184	77,581
conv10	13x13x1000	1x1/1 (x1000)	1				**20%** (3x3)			6bit	513,000	103,400
avgpool10	1x1x1000	13x13/1	0									
	activations		parameters				compression info				1,248,424 (total)	**421,098** (total)

实际效果

Table 2: Comparing SqueezeNet to model compression approaches. By *model size*, we mean the number of bytes required to store all of the parameters in the trained model.

CNN architecture	Compression Approach	Data Type	Original → Compressed Model Size	Reduction in Model Size vs. AlexNet	Top-1 ImageNet Accuracy	Top-5 ImageNet Accuracy
AlexNet	None (baseline)	32 bit	240MB	1x	57.2%	80.3%
AlexNet	SVD (Denton et al. 2014)	32 bit	240MB → 48MB	5x	56.0%	79.4%
AlexNet	Network Pruning (Han et al. 2015b)	32 bit	240MB → 27MB	9x	57.2%	80.3%
AlexNet	Deep Compression (Han et al. 2015a)	5-8 bit	240MB → 6.9MB	35x	57.2%	80.3%
SqueezeNet (ours)	None	32 bit	4.8MB	**50x**	57.5%	80.3%
SqueezeNet (ours)	Deep Compression	8 bit	4.8MB → 0.66MB	**363x**	57.5%	80.3%
SqueezeNet (ours)	Deep Compression	6 bit	4.8MB → 0.47MB	**510x**	57.5%	80.3%



(a) Exploring the impact of the squeeze ratio (SR) on model size and accuracy.

(b) Exploring the impact of the ratio of 3x3 filters in expand layers (pct_{3x3}) on model size and accuracy.

我们将挤压比（SR）定义为挤压层中滤波器的数量与膨胀层中滤波器的数量之间的比率

Table 3: SqueezeNet accuracy and model size using different macroarchitecture configurations

Architecture	Top-1 Accuracy	Top-5 Accuracy	Model Size
Vanilla SqueezeNet	57.5%	80.3%	4.8MB
SqueezeNet + Simple Bypass	**60.4%**	**82.5%**	4.8MB
SqueezeNet + Complex Bypass	58.8%	82.0%	7.7MB

个人理解

文章中说是参数少了50倍，但是实际上alexnet中很大一部分参数都是在全连接层中，然后squeezenet仅仅是减少了参数量，并未讨论在嵌入式系统中的实时性，主要提出的fire 模块，通过组合卷积的形式，减少了原来的3×3卷积，实际上的核心感觉和inception差不多，但是inception用的是不同尺寸的卷积核，文章最成功的一点就是保持精度不变的同时减少了参数量。