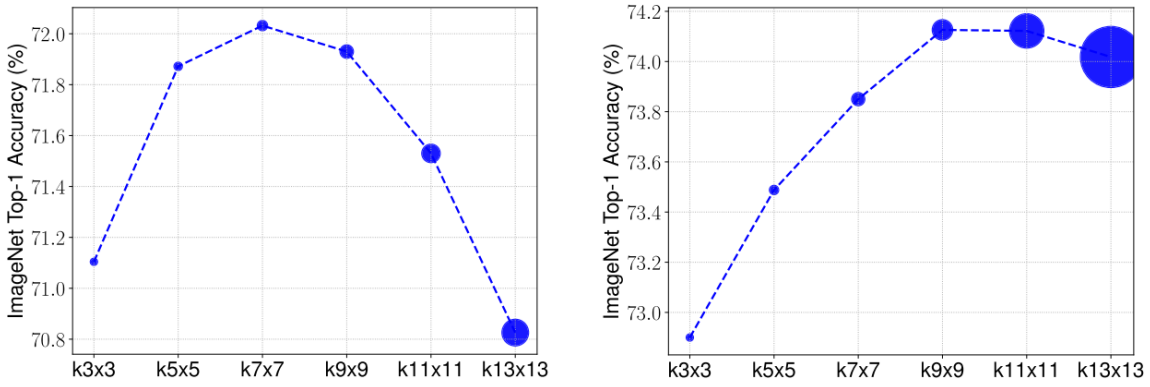
状态

已开源

简介

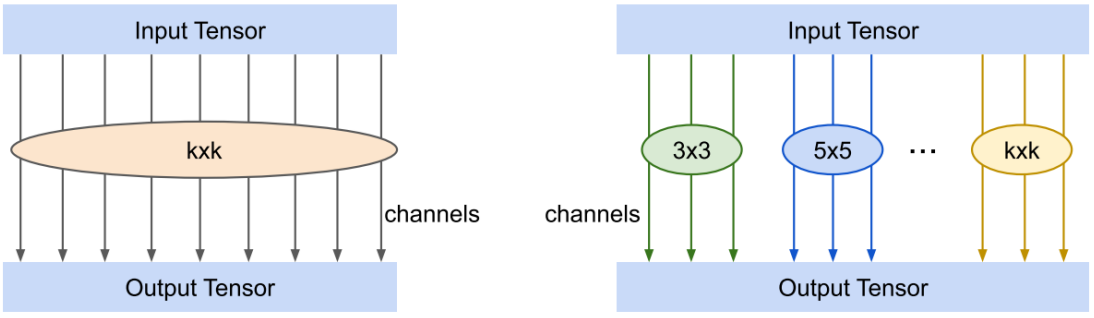
本文提出了一种新的层，或者说是一种新的卷积方式 `MDConv`，其想法基于深度可分离卷积(depthwith convs)，但是与深度可分离卷积不同的是加入了不同尺度的kernel size大小。
首先作者发现单纯的增大kernel size会出现精度先上升后下降的情况，作者认为大kernel有利于学习全局信息，而小kernel有利于学习局部细节，所以最好的方式就是同时结合大kernel和小kernel。



(a) MobileNetV1 (b) MobileNetV2

Figure 1: **Accuracy vs kernel sizes** – Each point represents a model variant of MobileNet V1[5] and V2 [19], where model size is represented by point size. Larger kernels lead to more parameters, but the accuracy actually drops down when kernel size is larger than 9x9.

作者的方法是先将depwith的kernel分组，然后每组kernel都采用不同的kernel size,以3x3,5x5,7x7的方式递增。最后用NAS去搜索一个新的网络架构，达到了非常好的效果（Imagenet SOTA 78.9% FLOPS < 600M）



(a) Vanilla Depthwise Convolution (b) Our proposed MDConv

Figure 2: **Mixed depthwise convolution (MDConv)** – Unlike vanilla depthwise convolution that applies a single kernel to all channels, MDConv partitions channels into groups and apply different kernel size to each group.

MDconv

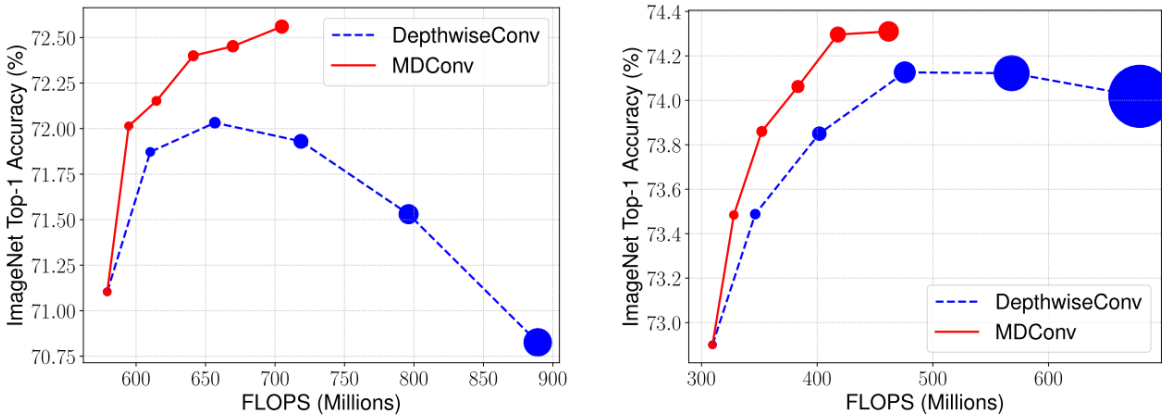
作者提出的MDConv可以作为新的op去使用，其运算量比单独的大Kernel depwith conv要小很多，但是效果反而更好。

设计选择

- group size，作者发现设置4比较通用，NAS的时候设置的是1-5
- kernel size per group: 2i + 1
- channel size per group 两种group channel划分模式，等分或者指数减少
- dilated conv: 扩张卷积，一般代替大卷积核的常用方法，但效果不怎么好

实验

无论是分类还是检测，效果都比单纯的大卷积核效果好很多。



(a) MobileNetV1 (b) MobileNetV2

Figure 4: **MDConv performance on ImageNet** – Each point denotes a model with kernel size from 3x3 to 13x13, same as Figure 1. `MDConv` is smaller, faster, and achieves higher accuracy than vanilla depthwise convolutions.

Network	MobileNetV1 [5]			MobileNetV2 [19]		
	#Params	#FLOPS	*mAP*	#Params	#FLOPS	*mAP*
baseline3x3	5.12M	1.31B	21.7	4.35M	0.79B	21.5
depthwise5x5	5.20M	1.38B	22.3	4.47M	0.87B	22.1
mdconv 35 (ours)	**5.16M**	**1.35B**	**22.2**	**4.41M**	**0.83B**	**22.1**
depthwise7x7	5.32M	1.47B	21.8	4.64M	0.98B	21.2
mdconv 357 (ours)	**5.22M**	**1.39B**	**22.4**	**4.49M**	**0.88B**	**22.3**

Table 1: **Performance comparison on COCO object detection.**

Ablation Study

- MDConv单层替换，作者只替换MobilenetV2的某一层的卷积维MDconv为vanilla DepthConv9x9 or MDconv3579，作者发现s=2的层prefer大kernel

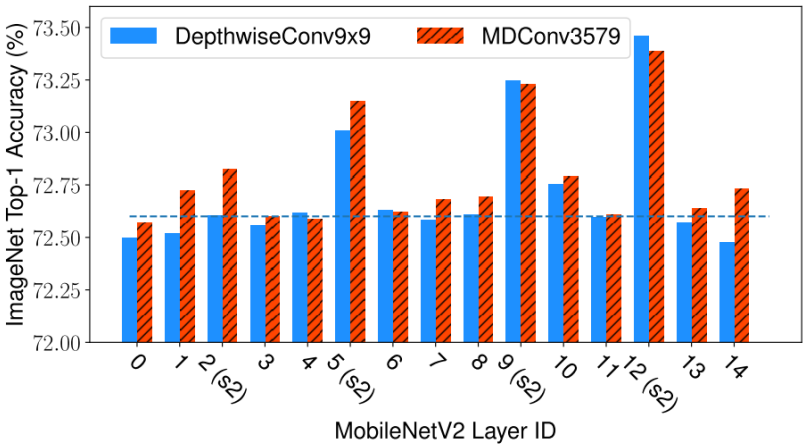


Figure 5: **Per-layer impact of kernel size** – `s2` denotes stride 2, while others have stride 1.

而且网络后半程的影响比前半程大一些

- 不同的Channel Partion策略和采用Dilated Conv的影响：对于Channel Partion策略，MobilenetV1不敏感，但是V2比较敏感，平均分配精度高一些。然后扩张卷积用的是3x3s1卷积配合不同的dilated rate(比如9x9就是3x3d4s1)，发现扩张卷积效果很很差，主要是大kernel的话dilated rate太大导致卷积会跳过很多信息。

Mixnet

作者又对MDconv做了NAS生成了MixNets家族。但是NAS的时候没有加exp channel partion和dilated conv的选项。下面是跟一些其他网络的对比：

Model	Type	#Parameters	#FLOPS	Top-1 (%)	Top-5 (%)
MobileNetV1 [5]	manual	4.2M	575M	70.6	89.5
MobileNetV2 [19]	manual	3.4M	300M	72.0	91.0
MobileNetV2 (1.4x)	manual	6.9M	585M	74.7	92.5
ShuffleNetV2 [15]	manual	-	299M	72.6	-
ShuffleNetV2 (2x)	manual	-	597M	75.4	-
ResNet-153 [4]	manual	60M	11B	77.0	93.3
NASNet-A [31]	auto	5.3M	564M	74.0	91.3
DARTS [14]	auto	4.9M	595M	73.1	91
MnasNet-A1 [25]	auto	3.9M	312M	75.2	92.5
MnasNet-A2	auto	4.8M	340M	75.6	92.7
FBNet-A [26]	auto	4.3M	249M	73.0	-
FBNet-C	auto	5.5M	375M	74.9	-
ProxylessNAS [2]	auto	4.1M	320M	74.6	92.2
ProxylessNAS (1.4x)	auto	6.9M	581M	76.7	93.3
MixNet-S	**auto**	**4.1M**	**256M**	**75.8**	**92.8**
MixNet-M	**auto**	**5.0M**	**360M**	**77.0**	**93.3**
MixNet-L	**auto**	**7.3M**	**565M**	**78.9**	**94.2**

Table 2: **MixNet performance results on ImageNet 2012 [18].**

- **3x3, 5x5, 7x7, 9x9, 11x11**: MDConv with five groups of filters ($g = 5$) with kernel size {3x3, 5x5, 7x7, 9x9, 11x11}. Each group has roughly the same number of channels.

Imagenet测试和模型结构

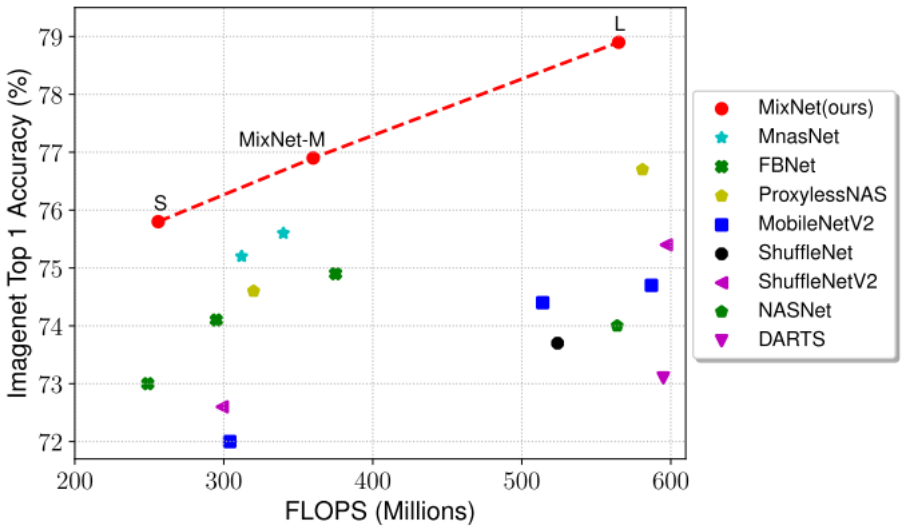


Figure 7: **ImageNet performance comparison.**

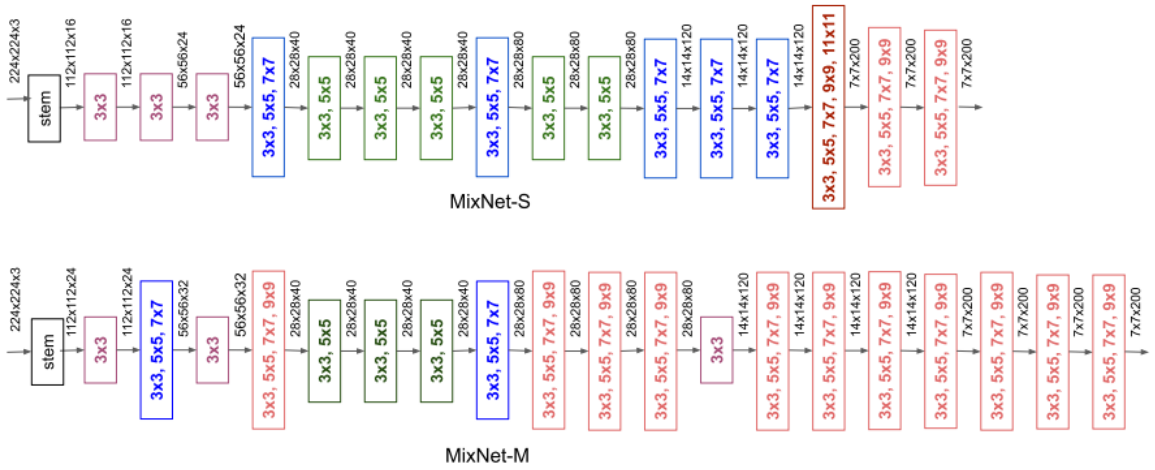


Figure 8: **MixNet architectures** – `MixNet-S` and `MixNet-M` are from Table 2. We mainly highlight MDConv kernel size (e.g. {3x3, 5x5}) and input/output tensor shape.

可以发现NAS的结果，在比较深的层偏好用大kernel，在比较浅的层偏好小kernel

迁移学习效果

Dataset	TrainSize	TestSize	Classes
CIFAR-10 [10]	50,000	10,000	10
CIFAR-100 [10]	50,000	10,000	100
Oxford-IIIT Pets [16]	3,680	3,369	37
Food-101 [1]	75,750	25,250	101

Table 3: **Transfer learning datasets.**

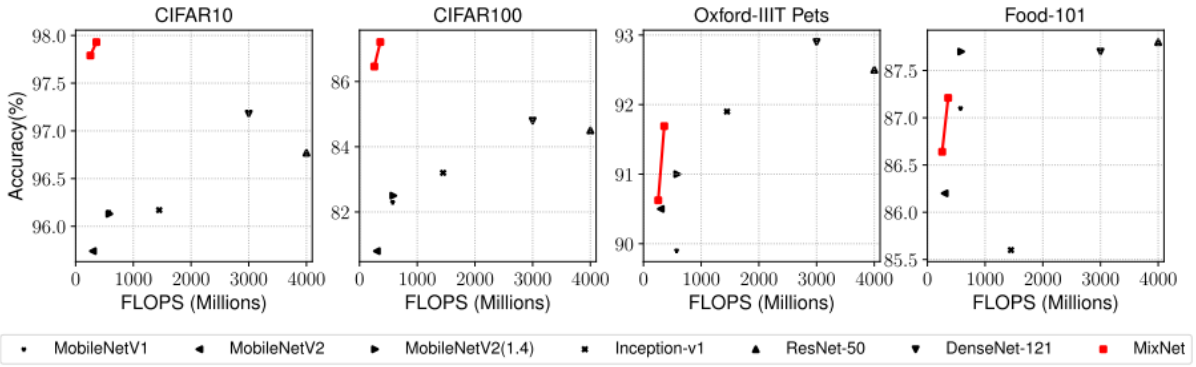


Figure 9: **Transfer learning performance** – `MixNet-S/M` are from Table 2.

在cifar上甚至超过了densenet121这样的模型

总结

这篇文章有一个很有意义的结论，浅层小Kernel比较有用，深层大kernel比较有用，而且再一次证明了NAS在模型结构构造上的重要性。
同时对模型的dilated conv kernel size和partion模式再进行一些细粒度的微调和搜索，相信可以得到更快效果更好的模型，因为毕竟这个改进是在mobilenet的基础上加入了大kernel，速度会有所下降。
同时文章BlazeFace当中提到了一个结论，mobile GPU上增大kernel size比增大channel更经济，那么大kernel较小的channel是不是可以进一步进行trade off，也是工程上值得思考的问题。