Bag of Tricks for Convolutional Neural Networks

mileistone
深度学习（Deep Learning）话题下的优秀答主

+ 关注他

已赞同 735

分享

Pascal、陈天奇、林天威、花花等 735 人赞同了该文章

刚刚看了Bag of Tricks for Image Classification with Convolutional Neural Networks，一篇干货满满的文章，同时也可以认为是GluonCV 0.3: 超越经典的说明书，通过这个说明书，我们也拥有了超越经典的工具箱。

我们都知道trick在CNNs中的重要性，但是很少有文章详细讲解他们使用的trick，更少有文章对比各个trick对最后效果影响，这篇文章把CNNs里几种重要的trick做了详细对比，可以认为是一篇在CNNs中使用trick的cookbook。

这篇文章虽然题目是“for Image Classification”，但是**这里面提到的trick和结论，我认为也适用于其他计算机视觉任务，比如目标检测、语义分割、实例分割等等**，特别地，我专门看了GluonCV里Yolov3的实现，里面有使用label smoothing和mixup。

这篇文章的trick有五个方面：model architecture, data augmentation, loss function, learning rate schedule，optimization。总结一句话就是，**网络input stem和downsample模块、mixup、label smoothing、cosine learning rate decay、lr warmup、zero γ对网络影响都不小**。

model architecture

这一部分主要讨论ResNet-50结构的一些微调，包括input stem和downsample module的细微改变。ResNet-50原始结构，和基于原始结构的一些微调如下图所示。

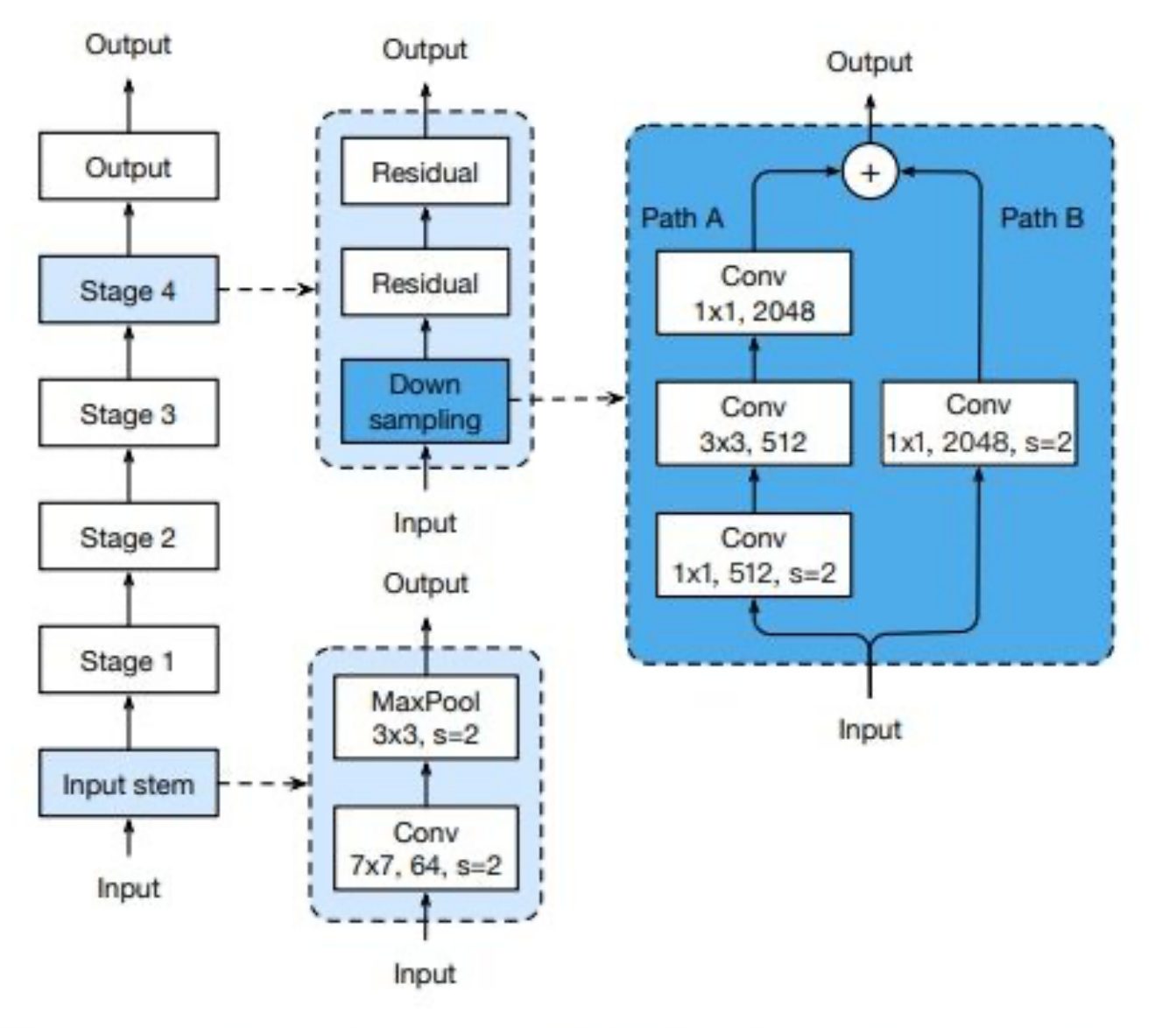
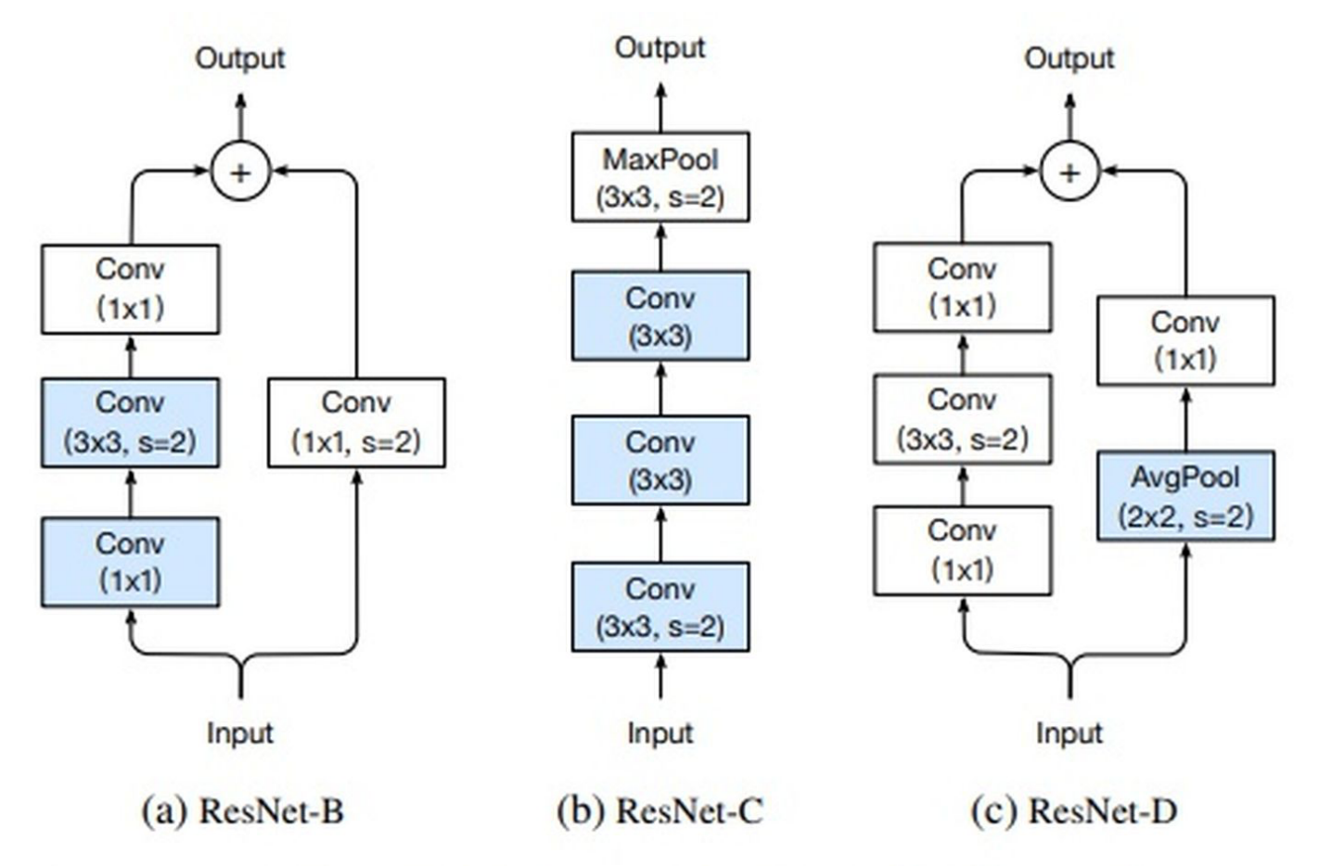


Figure 1: The architecture of ResNet-50. The convolution kernel size, output channel size and stride size (default is 1) are illustrated, similar for pooling layers.

原始ResNet-50

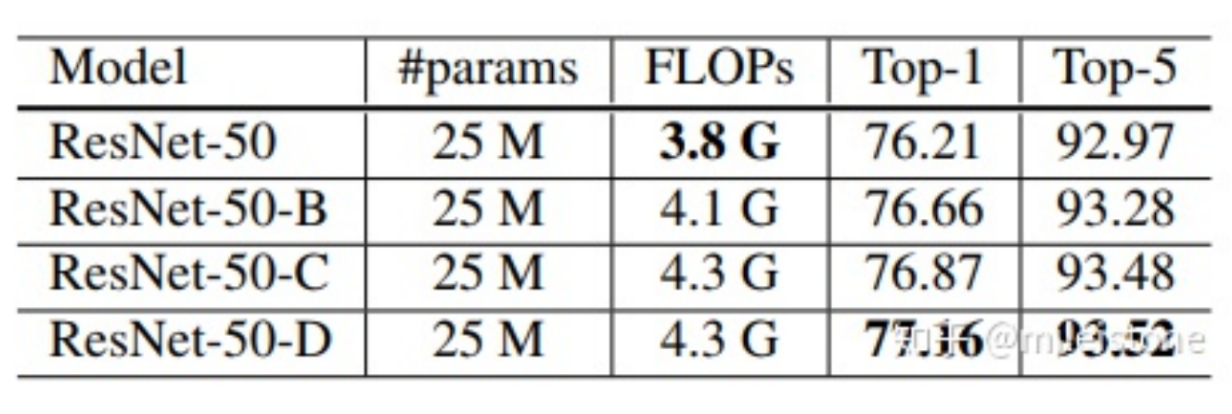


(a) ResNet-B (b) ResNet-C (c) ResNet-D

Figure 2: Three ResNet tweaks. ResNet-B modifies the downsampling block of Resnet. ResNet-C further modifies the input stem. On top of that, ResNet-D again modifies the downsampling block.

ResNet-50网络结构的几个变体

结果对比如下：



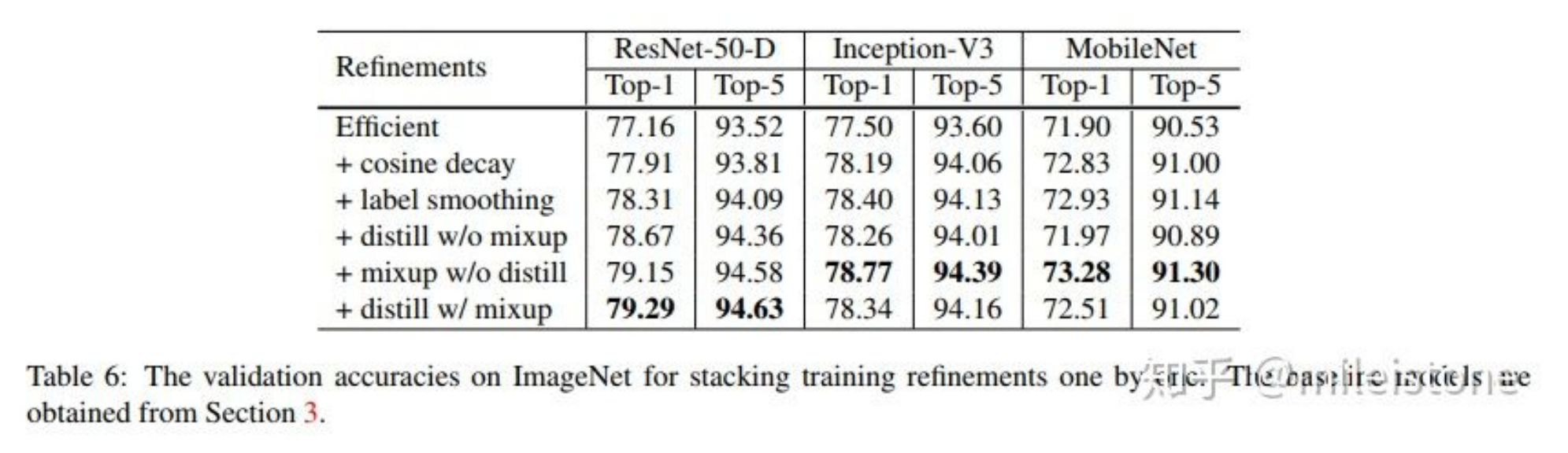
Model	#params	FLOPs	Top-1	Top-5
ResNet-50	25 M	**3.8 G**	76.21	92.97
ResNet-50-B	25 M	4.1 G	76.66	93.28
ResNet-50-C	25 M	4.3 G	76.87	93.48
ResNet-50-D	25 M	4.3 G	**77.16**	**93.52**

网络结构微调的对比

可以看出，这些小修改对计算量的影响很小，但是对最后的accuracy提升效果不小。我在设计目标检测网络的时候，也有类似的结论。多说一句，ResNet-50-C这种修改，虽然对计算量影响不大，不过根据我的经验，对速度的影响应该会比较大。

data augmentation

mixup对模型提升较大，具体对比如下。



Refinements	ResNet-50-D		Inception-V3		MobileNet	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Efficient	77.16	93.52	77.50	93.60	71.90	90.53
+ cosine decay	77.91	93.81	78.19	94.06	72.83	91.00
+ label smoothing	78.31	94.09	78.40	94.13	72.93	91.14
+ distill w/o mixup	78.67	94.36	78.26	94.01	71.97	90.89
+ mixup w/o distill	79.15	94.58	**78.77**	**94.39**	**73.28**	**91.30**
+ distill w/ mixup	**79.29**	**94.63**	78.34	94.16	72.51	91.02

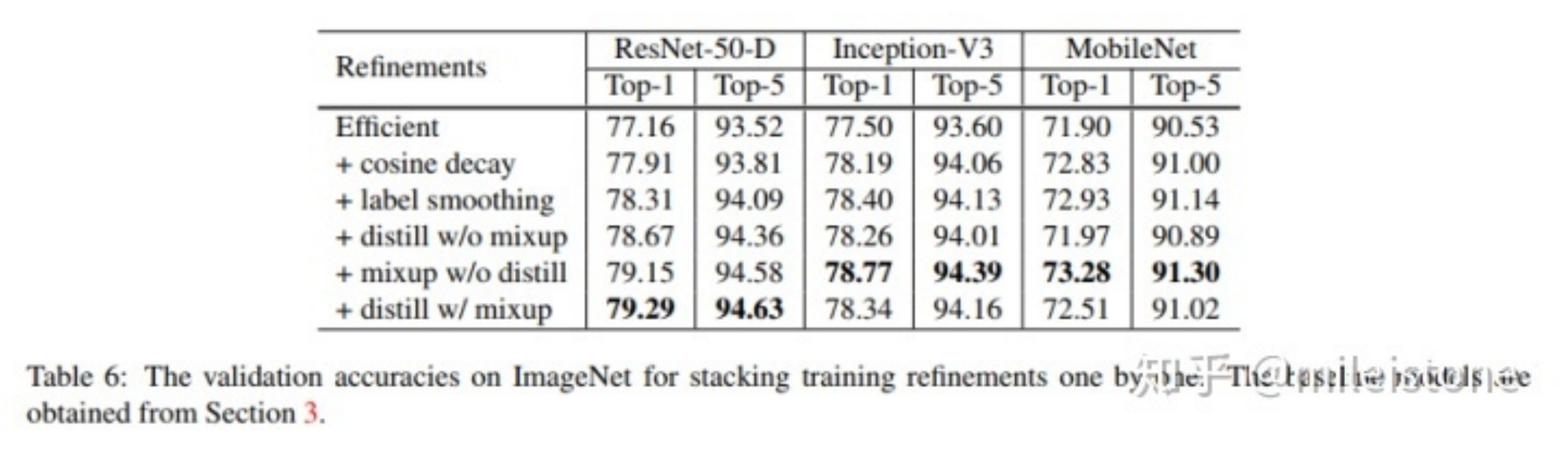
Table 6: The validation accuracies on ImageNet for stacking training refinements one by one. The baseline models are obtained from Section 3.

mixup对模型效果影响

data augmentation对模型效果影响蛮大的，不说mixup，单说resize的范围就能对模型效果有着不小的影响，有时候好好调调data augmentation里的参数，带来的效果提升比对网络结构的改进要还要大。**数据和模型是一个硬币的两面**，虽然改进数据没有改进模型听起来高大上，而且也更脏，但是我认为**对数据的理解才是一个算法工程师的核心竞争力**。

loss function

label smoothing对模型效果影响如下。



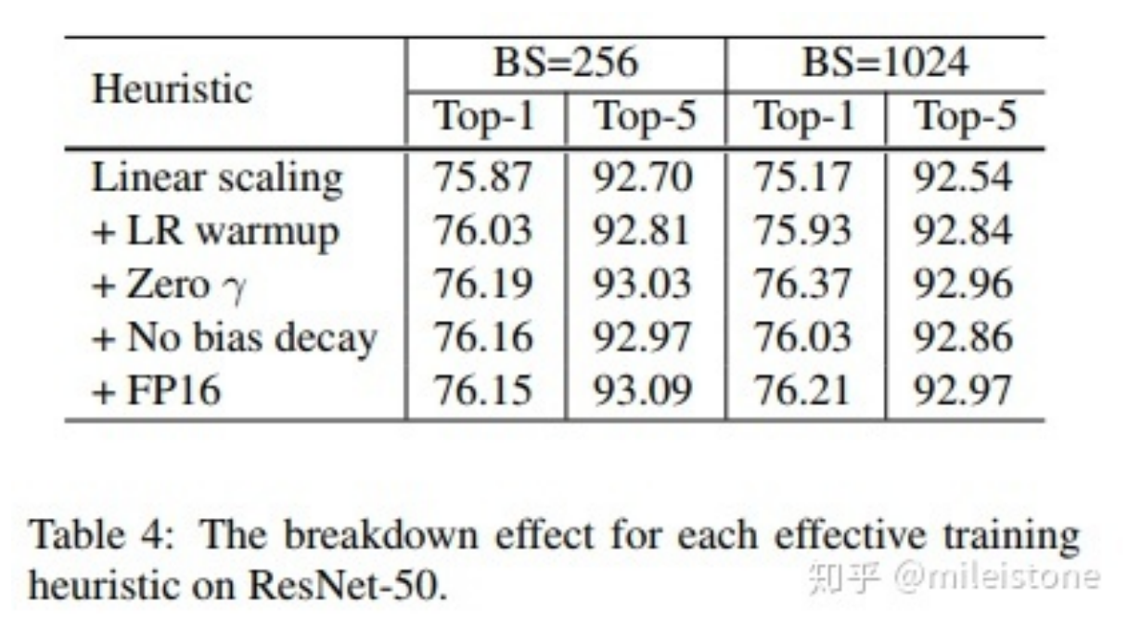
Refinements	ResNet-50-D		Inception-V3		MobileNet	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Efficient	77.16	93.52	77.50	93.60	71.90	90.53
+ cosine decay	77.91	93.81	78.19	94.06	72.83	91.00
+ label smoothing	78.31	94.09	78.40	94.13	72.93	91.14
+ distill w/o mixup	78.67	94.36	78.26	94.01	71.97	90.89
+ mixup w/o distill	79.15	94.58	**78.77**	**94.39**	**73.28**	**91.30**
+ distill w/ mixup	**79.29**	**94.63**	78.34	94.16	72.51	91.02

Table 6: The validation accuracies on ImageNet for stacking training refinements one by one. The baseline models are obtained from Section 3.

label smoothing对模型效果的影响

optimization

optimization涉及到lr warmup、zero γ、no bias decay、cosine decay。前三者对效果影响如下图所示，可以看出lr warmup和zero γ比较重要。

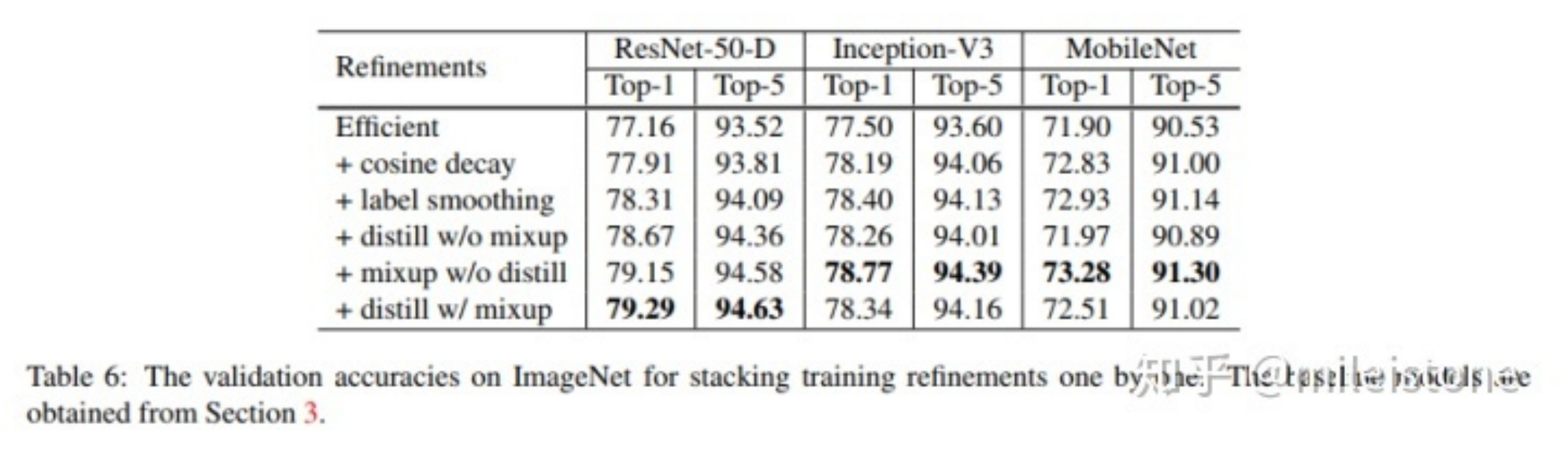


Heuristic	BS=256		BS=1024	
	Top-1	Top-5	Top-1	Top-5
Linear scaling	75.87	92.70	75.17	92.54
+ LR warmup	76.03	92.81	75.93	92.84
+ Zero γ	76.19	93.03	76.37	92.96
+ No bias decay	76.16	92.97	76.03	92.86
+ FP16	76.15	93.09	76.21	92.97

Table 4: The breakdown effect for each effective training heuristic on ResNet-50.

lr warmup、zero γ、no bias decay对模型效果的影响

cosine learning rate decay中对模型效果影响见下图，对比的是step learning rate decay。



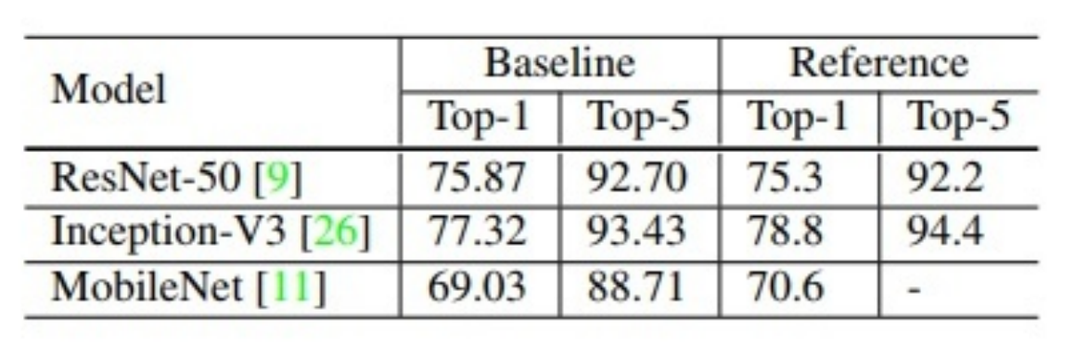
Refinements	ResNet-50-D		Inception-V3		MobileNet	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Efficient	77.16	93.52	77.50	93.60	71.90	90.53
+ cosine decay	77.91	93.81	78.19	94.06	72.83	91.00
+ label smoothing	78.31	94.09	78.40	94.13	72.93	91.14
+ distill w/o mixup	78.67	94.36	78.26	94.01	71.97	90.89
+ mixup w/o distill	79.15	94.58	**78.77**	**94.39**	**73.28**	**91.30**
+ distill w/ mixup	**79.29**	**94.63**	78.34	94.16	72.51	91.02

Table 6: The validation accuracies on ImageNet for stacking training refinements one by one. The baseline models are obtained from Section 3.

cosine learning rate decay对模型效果影响

一个有意思的细节

文章对比了自己复现的baseline和reference模型效果，具体如下。可以看出复现的basline和reference在三个模型结构下各有优劣，差距在0.5%到1%之间。我最近在用Yolov2和Yolov3，也有类似的经历，各个深度学习框架之间本身会有一些细微的差别，自己实现的代码，也可能带来一些细微差别，这些差别可能都细小到我们注意不到，然而最后却能对模型效果带来一个点左右的影响。



Model	Baseline		Reference	
	Top-1	Top-5	Top-1	Top-5
ResNet-50 [9]	75.87	92.70	75.3	92.2
Inception-V3 [26]	77.32	93.43	78.8	94.4
MobileNet [11]	69.03	88.71	70.6	-

Table 2: **Validation accuracy of reference implementations and our baseline.** Note that the numbers for Inception V3 are obtained with 299-by-299 input images.

PyTorch党的福利

这个链接里有支持pytorch的预训练模型权重。

编辑于 2018-12-11

深度学习（Deep Learning） 卷积神经网络（CNN）

文章被以下专栏收录

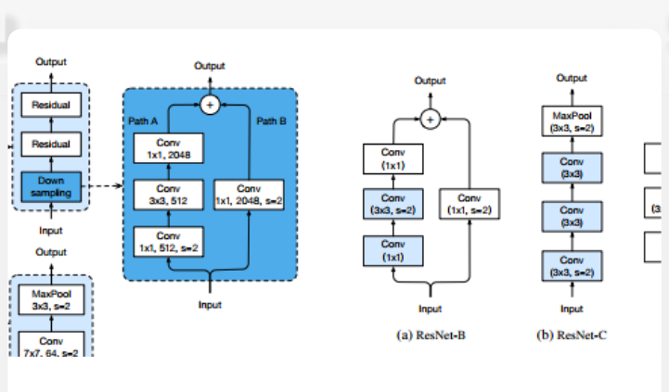
藏经阁
对一些有意思论文的理解

关注专栏

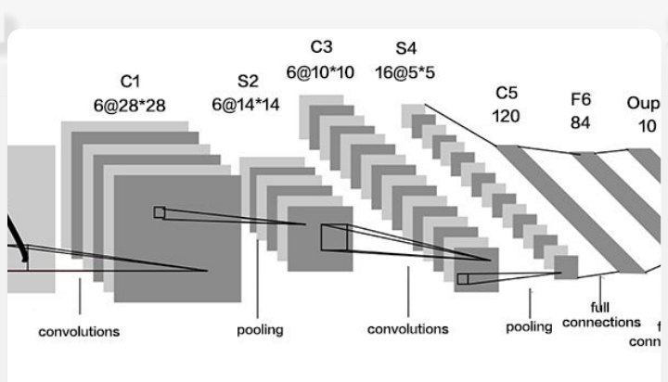
推荐阅读


mage Classification with Convolutional
ang Hang Zhang Zhongyue Zhang Juny
Amazon Web Services
ong,zhiz,hzaws,zhongyue,junyuanx,mli}@amazon

深度学习 cnn trick合集
sticky



《Bag of Tricks for Image Classification with CNN》
千佛山彭于... 发表于机器学习、...



12 个常见 CNN 模型论文集锦与 PyTorch 实现
红色石头 发表于AI有道

ResNeSt: Split-Attention Networks
这篇文章是在ResNet基础上的工作，融合了GoogleNet的Multi-path和SENet、SKNet中的attention思想，将ResNeSt用作分类、分割、目标检测的backbone，大大提升了任务的性能。...
dilligencer

12 条评论

切换为时间排序

写下你的评论...

董力 2018-12-09
[赞][赞]
赞

圈圈虫 2018-12-09
其实讨论resnet50这里，算力需要多13%了，也不少了
2

Gohomeeatlaunch 2018-12-12
文章里5.2. Label Smoothing 里面的p、q是不是写混了

赞

mileistone (作者) 回复 **Gohomeeatlaunch** 2018-12-12
是有点问题
赞

千佛山彭于晏 回复 **Gohomeeatlaunch** 2018-12-26
是写混了

赞

郑华滨 2018-12-21
话说没人注意到no bias decay反而掉点了么……是论文的typo吗？但是我们同事也试过no bias decay，据说时灵时不灵
赞

千佛山彭于晏 回复 **郑华滨** 2018-12-26
既然是tricks而不是大家都会用的通用方法，感觉是不是有效还得看具体的任务。
赞

RogerYu 回复 **郑华滨** 2019-01-07
那不是掉点吧，这四个trick不是一个个叠加的，而都是并行单独加到baseline上的，所以只用和baseline比就行了
1

展开其他 1 条回复

千佛山彭于晏 2018-12-24
“对数据的理解才是算法工程师的核心竞争力”这句话好评！
3

zzzz 2019-03-01
作者试过mixup对faster rcnn的效果吗，貌似掉点还掉不少
赞

kaka 回复 **zzzz** 2019-06-17
我试过在one-stage， 也是不涨反跌
赞