计算机视觉课程报告

黄星瑞
21821197
计算机科学与技术学院
765910276@qq.com

摘要

使用ImageNet预训练的网络模型能够帮助目标任务（物体检测、语义分割、细粒度识别等）快速收敛，然而使用预训练模型会带来诸多限制，其中一个问题就是无法改动特征提取网络的结构来适应不同需求的任务。那么，如果不使用预训练模型，进行随机初始化训练，达到较高准确率的某些必要条件是什么？本文推荐了这篇CVPR2019 的论文，作者们从优化的角度出发，通过实验解释了梯度稳定手段之一的BatchNorm是如何帮助随机初始化训练检测器，进而结合了ResNet与VGGNet来加强对小物体的检测。

1. 介绍

这篇文章[1]研究了针对基于随机初始化网络的优化方法因素，发现BatchNorm重新调整了优化问题的参数分布，使其外形更加平滑同时减少了internal covariate shift。基于此，文章作者认为从头训练检测网络收敛性较差的主要原因是缺少BN操作。因此，作者在backbone及检测子网络的头部都添加了BN层，发现添加BN后，从头开始训练的检测网络要比预训练的效果要好，进而可以不依赖于预训练网络对网络结构进行调整。实验发现，第一层卷积的下采样stride也对目标检测的效果产生重要的影响。基于这一点，这篇文章通过引入一个root block来设计检测器的结构。root block可以获得detector feature map中丰富的信息，提高了对小目标的准确率。

这篇文章主要贡献：

1）这篇文章设计了基于scratch训练的single-shot目标检测网络-ScratchDet，该网络结合了BN操作有利于网络的收敛，此方法适用于任意类型的网络结构。

2）引入了新的backbone Root-ResNet，提高了小目标的检测效果。

3） ScratchDet的检测表现效果较为强劲。

2. 动机

现有的检测训练任务存在三个限制：

分类任务与检测任务的Learning bias: 一方面是两者损失函数的不同，一方面是两者对平移不变性的敏感度不同，还有另一方面是数据集的差异：ImageNet数据集是单图单物体，COCO & VOC数据集是单图多物体。

如果想要改动检测模型中的特征提取网络的结构，需要对网络重新预训练再进行检测任务的finetune，而ImageNet预训练实验的代价比较大。这个问题在移动端、CPU实时检测器等设计中尤为突出，比如：Pelee，Tiny-SSD， YOLO-LITE， Fire-SSD， Tiny-YOLO，Tiny-DSOD， MobileNetV2 等等。常用的VGG-16、ResNet的计算量以及参数量对于移动端的负载较大，而设计小网络的每次修改都需要重新在ImageNet上训练，时间代价与计算资源消耗都比较大。再比如像DetNet[2]，想要设计一种专用于检测的网络，用在ImageNet预训练的实验就要花很多的时间。

Domain Transfer问题，比如从ImageNet自然与生活场景图像迁移到医疗图像中（X光图，核磁共振图）的癌症检测（S4ND）、卫星图像检测（You Only Look Twice）是否有用，不同域之间的迁移是否仍然能发挥作用?

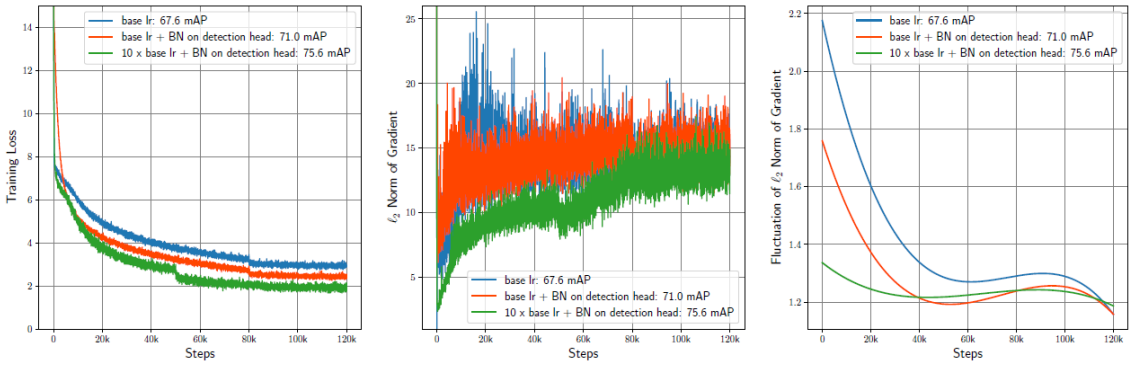
3. 分析

最开始讨论随机初始化训练的DSOD[3]将必要条件归结到一阶段检测器和DenseNet的dense layer-wise connection上，但是这样做很大程度限制了网络结构的设计。作者想找到随机初始化训练检测器的某些通用原因。受到NeurIPS 2018《How Does Batch Normalization Help Optimization?》这篇文章的启发:

通过理论和实验说明BN在优化过程中发挥的作用：梯度更加稳定，更加可预测；计算梯度时可采用更大的步长，即更大的学习率来加速训练；防止loss函数解空间突变，既不会掉入梯度消失的平坦区域，也不会掉入梯度爆炸的局部最小。

4. 初始化训练

沿着这个思路作者在SSD300 检测框架上给VGG网络与检测子网络分别加上了BN来进行随机初始化训练（PASCAL VOC 07+12 训练，07 测试），调整学习



(a) Loss Value (b) L2 Norm of Gradient (c) Fluctuation of L2 Norm of Gradient

Figure 1. Optimization landscape analysis. (a) The training loss value. (b) L2 Norm of gradient. (c) Fluctuation of L2 Norm of gradient (smoothed). Blue curve is the original SSD, red and green curves represent the SSD trained with BatchNorm in head subnetwork with 1× and 10× base learning rate, respectively. The BatchNorm makes smoother optimization landscape and has more stable gradients (red *v.s* blue). With this advantage, we are able to set larger learning rate (green) to search larger space and converge faster, and thus better solution.

率之后，得到的最好结果 78.7mAP，比直接随机初始化训练SSD的结果（67.6）高 11.6，比原SSD300（77.2）高 1.5，比使用预训练模型VGG-16-BN（78.2）高 0.6。实验细节写在论文里。

从左到右的 3 幅图分别是训练loss，梯度的L2 Norm，梯度的波动程度。通过这三幅图能够从优化角度分析，为什么BN能够帮助随机初始化训练检测器，蓝色曲线代表直接对SSD使用 0.001 的学习率做从 0 训练，红色曲线在蓝色曲线的基础上在VGG网络上加了BN，绿色曲线在红色曲线的基础上使用了 10 倍的学习率。可以看到：从蓝色到红色，给特征提取网络添加了BN之后，梯度的波动程度大幅下降，梯度趋于稳定，优化空间更加平滑，训练loss下降，mAP从 67.6 升高到 72.8。而从红色到绿色，平滑的优化空间允许使用更大的学习率，loss进一步下降，mAP也从 72.8 升高到 77.8。作者们在检测子网络（detection head）也做了一样设置的实验，得出了相似的结论与梯度分析图，具体请参考论文。

作者们在 SSD300 上做了尽可能详细的对比实验（见下表），包括在 3 个不同学习率（0.001, 0.01, 0.05）下给特征提取子网络（VGG）添加 BN，给检测子网络（detection head）添加 BN，给全部网络添加 BN，给全部网络不添加 BN，以上四者的随机初始化训练以及对比预训练 fine-tune 实验。可以看到，在为整个检测网络的不同部分添加 BN 之后会有不同程度的提升，而提升最高的是为整个网络添加 BN，在 VOC2007 测试集上（使用 VOC07+12 trainval 训练）可以达到 78.7mAP。

借着随机初始化训练带来的优势，可以对特征提取网络进行改动。之后作者借鉴了VGGNet和ResNet的优点，最大程度保留原图信息，来对小物体检测（论文中输入图像大小是 300X300，小物体较多）。

5. 模型改进

在SSD的升级版论文DSSD[4]中，作者将SSD的特征提取网络从VGG-16 替换成了ResNet-101，所得实验结果汇总如下表：

Table 6: Experiment results from DSSD

input size	network	training set	testing set	mAP
300×300	VGG-16	PASCAL VOC 07+12	PASCAL VOC 07	77.5
321×321	ResNet-101			77.1
512×512	VGG-16			79.5
513×513	ResNet-101			80.6
300×300	VGG-16	PASCAL VOC 07++12	PASCAL VOC 12	75.8
321×321	ResNet-101			75.4
512×512	VGG-16			78.5
513×513	ResNet-101			79.4
300×300	VGG-16	MS COCO trainval35k	MS COCO test-dev	25.1
321×321	ResNet-101			28.0
512×512	VGG-16			28.8
513×513	ResNet-101			31.2

ResNet-101 在ImageNet的top-5 error上比VGG-16 低了 2.69%，但是在 SSD300-VOC 的结果却低于 VGG-16，为什么？跟VGG-16 相比，ResNet-101 的优点是分类能力强，缺点是对小物体识别能力较差，因为第一个卷积层的stride=2，在初始输入的图片上就进行下采样，会损失某些原图信息，尤其是小物体的信

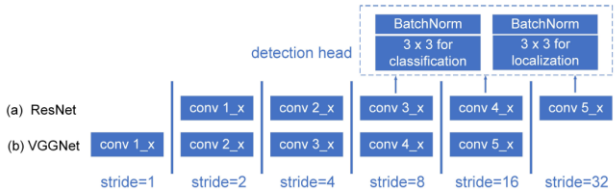
Table 1. Analysis of BatchNorm and learning rate for SSD trained from scratch on VOC 2007 `test` set. All the networks are based on the truncated VGG-16 backbone network. The best performance (78.7% mAP) is achieved when three conditions are satisfied: (1) BatchNorm in backbone and head, (2) non pretraining, (3) larger learning rate. "NAN" indicates that the training is non-convergent.

Component	lr 0.001						lr 0.01						lr 0.05					
pretraining					✓	✓					✓	✓					✓	✓
BN in backbone			✓	✓		✓			✓	✓		✓			✓	✓		✓
BN in head		✓		✓	✓	✓		✓		✓	✓	✓		✓		✓	✓	✓
mAP (%)	67.6	71.0	72.8	71.8	77.1	77.6	NAN	75.6	77.8	77.3	76.9	78.2	NAN	NAN	78.0	78.7	NAN	75.5

息。

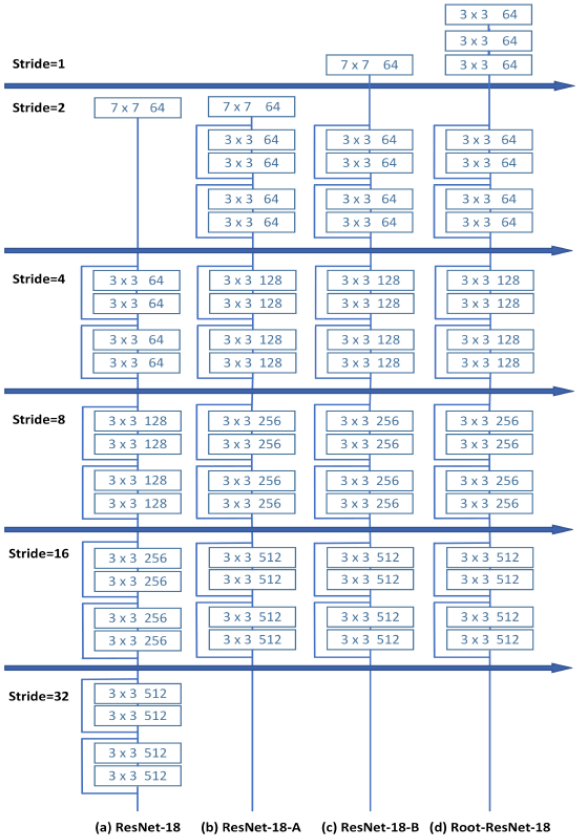
1) 在VOC_300 时，ResNet-101 的缺点＞优点，输入图片较小，图片中小物体数目变多，缺点被放大；且类别只有 20 类，不能发挥ResNet强大的分类能力，在SSD上结果低于VGG-16。
2) 在VOC_512 时，ResNet-101 的缺点<优点，输入图片变大，图片中小物体数目变少，缺点被缩小，在SSD上结果高于VGG-16。
3) 在COCO上时，ResNet-101 的缺点<优点，任务类别有 80 类，是VOC的 4 倍，ResNet-101 能充分发挥分类能力，所以无论输入 300x300 或者 512x512，在SSD上结果均高于VGG-16。

当然这里的结论需要进一步结合数据集本身的特性来验证，比如数据集规模vs模型参数量，数据集每张图片的instance数目，数据集小物体数量等等因素。



a) Blind review

因此作者借鉴了ResNet与VGGNet的优点，首先把ResNet的第一个卷积层的stride从 2 改成 1，也就是取消第一个下采样操作，并且参照了DSOD[3]的方法，替换第一个卷积层为 3 个 3x3 卷积层：这样做的目的是，尽可能保持原图信息不损失，并且充分利用。注意：在将新网络替换到SSD框架上时，仍然最大程度保证实验的公平性。首先，用于检测的特征图在论文中保持 38×38, 19×19, 10×10, 5×5, 3×3, 1×1 的大小，并没有使用大的特征图；其次，保证每个用于检测的特征图的channel数目相同。



(a) 原ResNet-18: 结果为 73.1 mAP。

(b) ResNet-18-A: 去掉了ResNet-18 的max-pooling层，即取消第二个下采样操作，结果为 75.3 mAP。

(c) ResNet-18-B: 将ResNet-18 的第一个卷积层的stride=2 改为 1，即取消第一个下采样操作，结果为 77.6 mAP 。

(d) Root-ResNet-18: 将ResNet-18-B的第一个 7x7 卷积核替换成 3 个 3x3 卷积，结果为 78.5mAP。

分析：在 300x300 大小的输入图像上（小物体较多）：

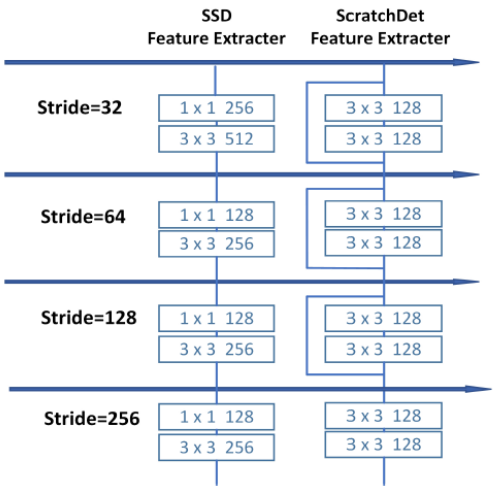
对比(a)与(c): 取消第一个下采样操作，提升了 4.5mAP。

对比(a)与(b): 取消第二个，保留第一个下采样操作，提升 2.2mAP。

对比(b)与(c): 是否对原图进行下采样，会有 2.3mAP 的影响。

对比(c)与(d): 替换 7x7 为 3 个 3x3，使用更加冗余的特征会提升结果。

之后，作者将SSD在特征提取网络后面添加的多个卷积层替换为残差模块，减少了参数量，计算量，提升了FPS（SSD300-Root-Res34:20FPS->25FPS），而且检测准确率没有下降（在VOC07 上，80.4mAP)：



最后，作者使用了Root-ResNet-34 来做随机初始化训练，得到较好的检测结果：

(07+12 训练，07 测试）：80.4 mAP ,

(07++12 训练，12 测试）：78.5 mAP,

(COCOtrainval35k，COCO测试）：32.7 AP,

值得注意的是，AP@S 13.0 ，在小物体检测结果相对较好。

作者们还对比了训练时间，使用mmdetection检测框架（使用了repeat dataset加速训练trick），在输入为300x300 的时候，随机初始化训练大约需要 84.6 小时，而使用预训练模型fine-tune需要 29.7 小时。

但是相比起ImageNet数以百万计的图片数目与几周的训练时间来说，随机初始化训练检测器使用的时间相对更少的，可以被人们所接受。

6. 总结

关于随机初始化训练检测器有三个必要条件：需要稳定梯度的优化手段(clip_gradient、BN、GN、SN等)；训练足够多的epoch与合适的学习率；数据训练集要相对较大。而对于小数据集，在训练时需要一定的数据增广。其中，对于需要能够稳定梯度的优化手段：BatchNorm 、 GroupNorm 、 SwitchableNorm 以及clip_gradient都是能够稳定梯度的优化手段（注意BN需要在单张显卡中达到一个较大的batch以获得足够的统计量，sync_BN解决了这个问题）。

对于这篇文章，我认为现有可以继续深入研究的地方：为什么使用ImageNet或其他超大规模数据集预训练的模型能够加速模型收敛；更加有效的稳定梯度的优化手段；探索图像分类任务与检测，分割等其他任务的learning bias与gap；借助随机初始化训练的优势，大量尝试移动端的小型检测网络；如果在该领域能够成功随机初始化训练，那么可以在该领域针对数据集的特点设计专用的特征提取器，而不是一味采取针对ImageNet设计的分类网络。

目前，在JD AI Research 已经部分团队成功把随机初始化训练用在其他任务上，比如：JD AI Reserach在WIDER FACE人脸检测竞赛排名第一的ISRN，使用GN与两倍epoch随机初始化训练来重新设计用于检测小人脸并且相对节省显存的网络

ImageNet数据集深深地影响了计算机视觉的发展，相信未来会有许多好的工作解释清楚迁移学习中的奥秘、提出更加高效的训练策略、打开深度学习的黑箱。

References

[1] Zhu R, Zhang S, Wang X, et al. Scratchdet: Exploring to train single-shot object detectors from scratch[J]. arXiv preprint arXiv:1810.08425, 2018.

[2] Li Z, Peng C, Yu G, et al. Detnet: A backbone network for object detection[J]. arXiv preprint arXiv:1804.06215, 2018.

[3] Shen Z, Liu Z, Li J, et al. Dsod: Learning deeply supervised object detectors from scratch[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 1919-1927.

[4] Fu C Y, Liu W, Ranga A, et al. Dssd: Deconvolutional single shot detector[J]. arXiv preprint arXiv:1701.06659, 2017.