简单新颖神操作，截断骨干用于检测！YOLO-ReT开源：边缘GPU设备上的高性能检测器

原创 CV开发者都爱看的 极市平台 2021-10-29 22:00:00 手机阅读

作者丨happy
编辑丨极市平台

极市导读

本文提出了一种边缘GPU设备上的高性能检测器YOLO-ReT，它包含两个关键性的改进：(1) 边缘GPU友好的模块RFCR用于多尺度特征交互；(2) 一种基于迁移学习的骨干截断机制。尤其值得称道的是：文中关于骨干截断分析与实验。为下游任务的模型缩放提供了一种更好的骨干网络设计机制。 >>加入极市CV技术交流群，走在计算机视觉的最前沿

YOLO-ReT: Towards High Accuracy Real-time Object Detection on Edge GPUs

Prakhar Ganesh[1*], Yao Chen[1*], Yin Yang[2], Deming Chen[3], Marianne Winslett[3]
[1]Advanced Digital Sciences Center, Singapore
[2]College of Science and Engineering, Hamad Bin Khalifa University, Qatar
[3]University of Illinois at Urbana-Champaign, USA
{prakhar.g, yao.chen}@adsc-create.edu.sg, yyang@hbku.edu.qa, {dchen, winslett}@illinois.edu

论文链接：https://arxiv.org/pdf/2110.13713.pdf

代码链接：https://github.com/prakharg24/yoloret

Abstract

目标检测模型在模型精度与效率两条主线上得到了长足发展。然而，为将DNN检测器部署到边缘设备上，我们需要对齐进行大幅压缩，这就导致了模型性能的牺牲。

本文提出了一种新颖的边缘GPU友好的模块RFCR用于多尺度特征交互；此外，**我们还提出一种新颖的迁移学习骨干调节方案**以补充特征交互模块，并协同提升模型在不同边缘GPU设备上的推理速度以及模型精度。

比如，当硬件平台为Jetson Nano时，**YOLO-ReT+MobileNetV2(x0.75)能够实时推理并取得68.75mAP@VOC、34.91mAP@COCO指标，以3.05mAP@VOC、0.91mAP@COCO以及3.05FPS提升优于对标方案**；当引入了RFCR模块后，YOLOv4-tiny与YOLOv4-tiny(3l)可以取得了1.3与0.9mAP指标提升达到41.5mAP@COCO与48.1mAP@COCO。

Method

本文提出一种新的监测模型YOLO-ReT，它采用了RFCR模块与迁移学习启发的骨干截断机制提升模型性能与边缘GPU设备上的推理速度。下图为本文所提方案的整体架构示意图。

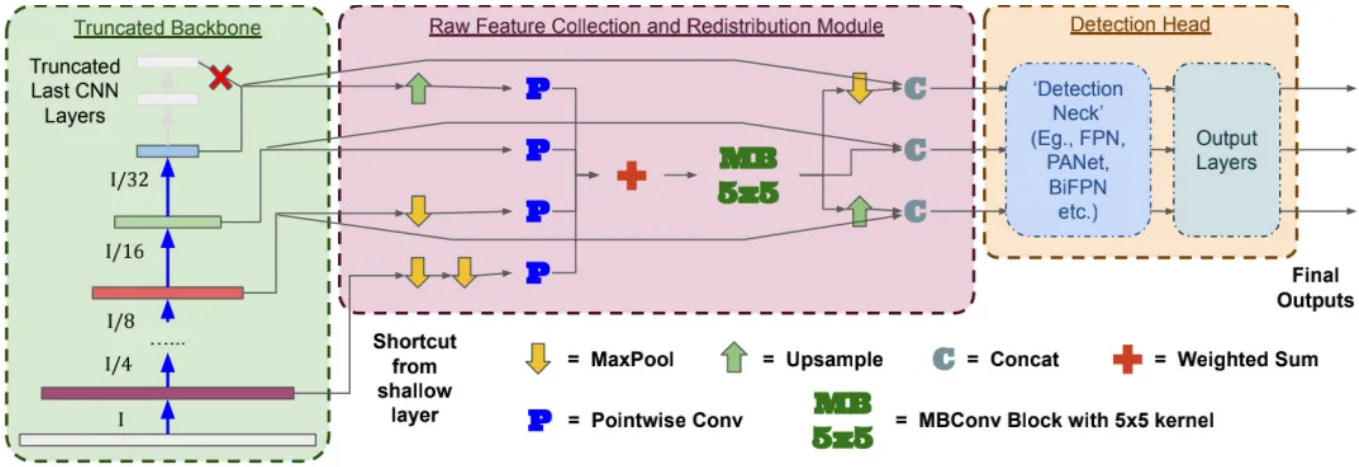


Figure 1. Complete Architecture of YOLO-ReT.

Raw Feature Collection and Redistribution

RFCR(见上图红色区域)的目标在于：对骨干网络提供的原生特征进行进一步增强，进而提升检测性能，同时不会大幅影响推理速度。注：尽管本文聚焦于检测，但RFCR可以扩展到其他相似任务。

现有的多尺度特征交互方案可以划分为top-down与bottom-up两种，两者均聚焦于相邻尺度特征交互，而忽视非相邻尺度特征交互。受NAS-FPN中的非相邻尺度特征连接启发，我们提出了一种轻量型RFCR模块：它对骨干网络的多尺度特征进行融合并重分布回每个特征尺度。因此，每个尺度包含了直接包含其他所有尺度的特征。

需要注意的是：**RFCR并不是用于替换其他特征聚合方案，而是提供一个超轻量特征处理模块对送入FPN等特征聚合方案的特征进行增强，并进一步提升整体性能** 。

此外值得注意的是，**RFCR独立于检测头的输出尺度数量，即输入输出特征数量之间无约束** 。比如，尽管YOLOV3检测头有三个输出尺度，我们在特征收集阶段采用骨干网络提供的四个不同特征以利用更细粒度底层特征提升模型性能。

如所知，特征融合方式与聚合路径同等重要。**为尽可能降低额外的延迟负担，我们将原生特征融入1×1卷积并采用简单的加权方式融合，将融合后特征融入MBConv模块处理后重分布回不同尺度** 。

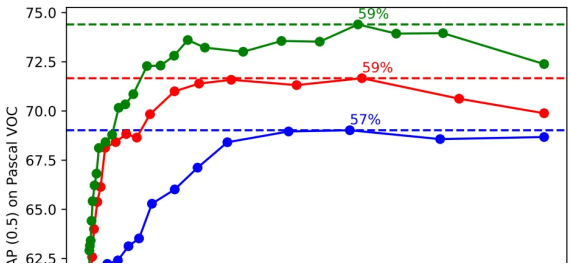
当对不同尺度特征进行融合时，常规上采样与下采样会导致语义不一致、局部位置不匹配问题。因此，**我们在MBConv中采用5×5卷积以提升检测性能** 。注：卷积核尺寸提升到7×7并不会带来进一步的性能提升。

Backbone Truncation

从分类模型预训练向下游任务迁移的重要性已受到过质疑，恺明大大也曾针对此写了《Rethinking ImageNet Pre-training》一文；此外某些文章甚至专门针对检测设计骨干模型。

基于这样的直觉：**通过连续CNN层处理的信息流会随任务而变化** 。比如，分类模型不需要尺度空域信息；而检测模型则需要保持空域信息完整性以输出细粒度检测结果。我们认为：**骨干网络的初始层的迁移学习能力非常重要，而尾部层并不会给检测/识别提供重要信息** 。

> The transfer learning capabilities of the initial layers of the feature extraction backbone are quite vital, and it iis th e last layers that do not provide critical information for the detection/recognition.



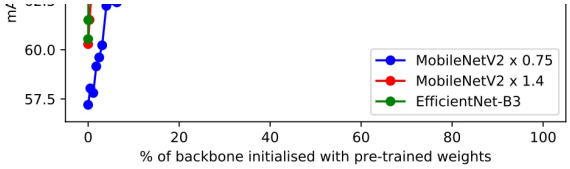


Figure 2. Transfer Learning Curve.

上图以PANet+YOLOV3检测头+不同比例层骨干预训练的性能对比，可以看到：

- 随着预训练初始化骨干层比例提升，模型性能逐渐提升。这说明：**迁移学习非常重要**；
- 当预训练初始化层的比例达到60%左右时，模型性能开始恶化。这说明：**相比随机初始化，采用预训练初始化骨干的尾部层会影响模型性能**。可能原因：这些层因与任务相关而陷入局部最优。

从上图可以看到最后2-3层包含约40%参数量。因此，**我们采用截断版骨干用于最终的目标检测模型**，提供了一种比降低缩放因子更优的选择。结合上图，**我们截断MobileNetV2的后两个模块、截断EfficientNet的后三个模块后作为检测器的骨干**。

Experiments

我们采用meanAP指标进行算法性能评估，采用FPS评估其效率。

Evaluation Setups

Backbone	Depth	Acc.	FPS			Size (MB)	Comp.
			Nano	NX	AGX		
ResNet50 [13]	177	74.9%	33.44	102.14	166.17	97.8	3989
MblNetV2 x 1.4 [42]	157	74.7%	47.64	**129.96**	**188.64**	23.5	588
DarkNet19 [39]	62	72.9%	**48.37**	115.07	187.25	79.5	2764
DarkNet53 [39]	187	77.2%	20.81	71.53	142.99	158	7172
CSPDarkNet53 [52]	418	77.2%	20.60	82.49	153.08	109	5038
EfficientNet-B0 [48]	250	77.1%	36.91	112.09	177.97	**20.4**	**396**
EfficientNet-B3 [48]	407	**81.6%**	17.97	74.48	134.94	47.1	1007

Table 1. Backbone accuracy (ImageNet Top-1) and FPS

上表给出了不同骨干网络的深度、精度、参数量、FLOPs以及在不同硬件上的FPS。基于该表汇总信息，**在骨干方面**，我们主要聚焦于三个常用轻量型骨干：MobileNetV2(x0.75)、MobileNetV2(x1.4)以及EfficientNet-B3；**在检测头方面**，我们采用**PANet+YOLOV3检测头**；在数据集方面，我们主要采用了Pascal VOC与COCO；**在度量指标方面**，我们主要采用meanAP与FPS，评价用的硬件平台包含Jetson Nano、Jetson Xavier NX以及AGX Xavier。

在训练方面，我们主要考虑两种类型的数据增广：

- geometric augmentations: 随机裁剪、旋转、镜像、缩放等；
- photometric augmentation：HSV调整、亮度调整等。

此外，我们还采用YOLOV4提到的自对抗训练；对定位方面，我们采用了GIoU损失。训练过程中，我们首先冻结迁移学习部分预训练参数训练100epoch，然后解冻所有参数训练150epoch。

在推理测试方面，我们以TenorRT+FP16进行加速，batch=1，推理1000张图像取平均计算FPS。与此同时，我们还提供了TensorRT+INT8配置的FPS以对标特定平台的整数推理计算（注：由于Jetson Nano不支持INT8加速，故相比FP16并无任何优势）。

Ablation Study

我们先来看一下消融分析并构建最终的检测模型，消融分析包含**骨干截断**与**RFCR模块**。

	x	Complete Backbone				Truncated Backbone			
		AP^{50}	FPS			AP^{50}	FPS		
			Nano	NX	AGX		Nano	NX	AGX
MblNetV2 (width = x)	1.4	69.88	19.96	62.58	90.63	69.67	24.58	67.38	95.19
	1.0	69.40	24.85	67.92	93.20	68.87	29.32	73.87	98.67
	0.75	68.67	28.16	73.25	99.75	66.58	34.02	77.67	103.77
	0.5	63.94	35.18	81.05	117.52	61.27	39.97	86.13	124.27
Efficient--Net-x	B3	72.05	9.69	45.72	73.15	72.28	11.24	49.72	78.61
	B2	71.84	12.44	50.86	80.08	71.92	13.45	55.61	82.47
	B1	71.67	13.34	52.31	82.19	71.59	14.62	57.52	86.22
	B0	71.24	18.27	60.97	95.96	70.98	19.08	64.34	99.95

Table 2. Comparing Width reduction vs Depth reduction

上表对比了完整骨干与截断骨干的性能对比，从中可以看到：

- 对于**EfficientNet**，**截断版本能够取得更佳的精度与FPS**；
- 对于**MobileNetV2**，**当具有相似FPS时，截断版本具有更高的性能**。比如MobileNetV2(x1.4)的截断版本比MobileNetV2(x1.0)完整版本具有相似FPS，但截断版本的性能高0.27AP；MobileNetV2(x0.75)的截断版本比MobileNetV2(x0.5)完整版本具有相似FPS，但截断版本的性能高2.64AP。
- 总而言之，相比完整骨干，**骨干截断能提供更精确、更快速的特征提取网络**。

Backbone	Feature Aggr. Path	Without RFCR module				With RFCR module			
		AP^{50}	FPS			AP^{50}	FPS		
			Nano	NX	AGX		Nano	NX	AGX
MblNetV2 x 0.75	None	59.92	38.24	80.17	126.42	63.97	36.93	78.38	119.41
	FPN	64.15	37.82	78.42	111.02	66.11	36.33	76.95	101.94
	PANet	66.58	34.02	77.67	103.77	68.75	33.19	71.64	95.97
	BiFPNx1	66.29	34.81	78.04	103.47	66.65	33.78	72.17	95.70
	BiFPNx2	66.69	32.98	76.10	101.78	66.83	31.43	75.61	99.40
	BiFPNx3	66.78	31.54	74.45	100.37	66.90	30.72	73.68	98.27
MblNetV2 x 1.4	None	68.15	28.21	72.97	110.24	69.26	26.50	71.81	106.43
	FPN	69.02	26.06	69.69	103.18	69.95	24.19	66.23	98.73
	PANet	69.67	24.58	67.38	95.19	70.35	23.01	65.37	93.49
	BiFPNx1	69.50	25.26	67.42	95.55	70.14	23.60	65.79	93.71
	BiFPNx2	69.84	23.25	65.77	92.96	70.53	22.18	64.43	92.03
	BiFPNx3	69.87	20.99	62.81	91.50	70.61	20.33	62.17	90.93
Efficient-Net-B3	None	71.24	12.21	62.31	87.60	72.37	11.94	59.84	84.58
	FPN	71.60	11.75	56.04	85.84	72.63	11.34	53.28	82.54
	PANet	72.28	11.24	49.72	78.61	72.96	10.96	47.07	75.61
	BiFPNx1	72.07	12.12	47.70	79.28	72.80	11.71	45.02	75.39
	BiFPNx2	72.39	11.15	43.53	76.22	73.01	10.72	42.24	72.89
	BiFPNx3	72.51	9.92	39.91	72.55	73.08	9.38	38.25	67.99

Table 3. Effectiveness of RFCR module

上图对比了 RFCR模块添加前后的性能与速度对比，从中可以看到：

- 无论何种骨干与特征聚合方案，**所提RFCR均能取得一致的性能提升**，当然FPS也会轻微下降；
- 当不存在特征聚合时，所提RFCR可以带来更多的性能提升；对于BiFPNx3，所提RFCR仍可进一步提升其性能，说明了**非近邻层连接的重要性**。

Baseline	+Truncate	+RFCR	+Shortcut	AP^{50}	FPS
MobileNetV2 x 0.75 on Jetson Nano	✗	✗	✗	68.67	28.16
	✓	✗	✗	66.58	34.02
	✗	✓	✗	69.50	26.97
	✓	✓	✗	68.40	33.35
	✓	✓	✓	68.75	33.19
MobileNetV2 x 1.4 on Jetson NX	✗	✗	✗	69.88	62.58
	✓	✗	✗	69.67	67.38
	✗	✓	✗	70.56	62.11
	✓	✓	✗	70.21	65.91
	✓	✓	✓	70.35	65.37
EfficientNet-B3 on Jetson AGX	✗	✗	✗	72.05	73.15
	✓	✗	✗	72.28	78.61
	✗	✓	✗	72.37	72.90
	✓	✓	✗	72.58	75.72
	✓	✓	✓	72.96	75.61

Table 4. Ablation study

上图给出了复合消融分析，可以看到：

- 对于骨干截断与完整骨干，RFCR均表现良好；而完整版骨干方案FPS下降明显；
- RFCR+短连接组合可以进一步提升性能，FPS几乎不变；
- 总而言之，**通过组合骨干截断与RFCR，我们可以提升推理速度与模型性能** 。

Comparison with SOTA Models

Model	Input Res.	Size (MB)	FPS Nano	FPS NX	FPS AGX	AP^{50} VOC	AP^{50} COCO
Tiny-YOLOv3 [10]	416	34.9	27.36	66.55	91.71	61.30	33.10
Tinier-YOLO [9]	416	8.9	30.14	68.73	92.09	65.70	**34.00**
YOLO Nano [55]	416	4.0	13.62	54.03‡	85.81‡	69.10	–
YOLO-Fastest [37]	320	1.3	**42.41**	**76.13**	**126.82**	61.02	–
YOLO-Fastest XL [37]	320	3.5	27.93	61.33	108.76	**69.43**	32.45
YOLO-ReT-M0.75	416	5.2	19.87	58.24	71.16	**72.39**	**36.44**
	320	5.2	33.19	71.64	95.97	68.75	34.91
	224	5.2	**55.16**	**84.10**	**134.87**	60.77	30.76
YOLO-ReT-M1.4	416	12.3	13.17	46.07	66.23	**73.32**	**36.52**
	320	12.3	23.01	65.37	93.49	70.35	35.77
	224	12.3	**43.16**	**84.32**	**113.94**	62.91	31.63
YOLO-ReT-EB3	416	28.3	6.35	28.83	49.07	**76.49**	**39.12**
	320	28.3	10.96	44.59	75.61	72.96	36.51
	224	28.3	**18.57**	**54.87**	**93.55**	65.52	33.11

‡ Calculated with INT8 optimization

Table 5. Comparison with other state-of-the-art models

上表给出了不同实时检测器的性能对比，从中可以看到：

- 当**输入分辨率变小时，模型具有更快的推理速度、更低的模型精度** 。
- 当硬件平台为Jetson Nano时，YOLO-ReT-M0.75@320x320具有比Tinier-YOLO高3.05mAP@VOC、0.91mAP@COCO的性能，同时推理速度快3.05FPS；
- 当硬件平台为Jetson Xavier NX时，YOLO-ReT-M1.4@320x320具有比YOLO-Fastest-XL高0.92mAP@VOC、3.34mAP@COCO的性能，同时推理速度快4.02FPS；
- 当平键平台为AGX Xavier时，YOLO-ReT-EB3@416x416取得了最佳性能，同时仍具有实时推理速度；
- 需要注意的是：在具有相似FPS时，基于MobileNetV2的检测器性能要优于骨干为EfficientNet的检测器。

Model	Input Res.	FPS Nano	COCO AP	COCO AP^{50}	COCO AP^{75}
YOLOv4-tiny	416	29.55	21.7	40.2	22.5
YOLOv4-tiny+RFCR	416	27.81	**22.9**	**41.5**	**23.3**
YOLOv4-tiny (3l)	608	24.87	28.7	47.2	29.7
YOLOv4-tiny (3l)+ RFCR	608	21.40	**29.3**	**48.1**	**30.5**

Table 7. Comparison with YOLOv4-tiny [51]

上表对比了YOLOv4与RFCR组合前后的性能对比，从中可以看到：**RFCR可以显著提升基线模型性能；RFCR对YOLOv4-tiny的提升更大** 。

本文亮点总结

1.本文提出了一种新颖的边缘GPU友好的模块RFCR用于多尺度特征交互 ； 此外，我们还提出一种新颖的迁移学习骨干调节方案 以补充特征交互模块，并协同提升模型在不同边缘GPU设备上的推理速度以及模型精度。

2.FCR并不是用于替换其他特征聚合方案，而是提供一个超轻量特征处理模块对送入FPN等特征聚合方案的特征进行增强，并进一步提升整体性能 。

极市干货

项目/比赛：珠港澳人工智能算法大赛｜算法打榜

算法trick：目标检测比赛中的tricks集锦｜从39个kaggle竞赛中总结出来的图像分割的Tips和Tricks

技术综述：一文弄懂各种loss function｜工业图像异常检测最新研究总结（2019-2020）