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, Illinois at Singapore [2]College of Science and Engineering, Hamad Bin Khalifa University, Qatar

[3]University of Illinois at Urbana-Champaign, USA

WACV 2021
WAIKOLOA, HAWAII

Summary

➢ Objective: Map Deep Neural Network (DNN) based object detection models to resource limited edge platforms with improved detection accuracy and runtime efficiency.

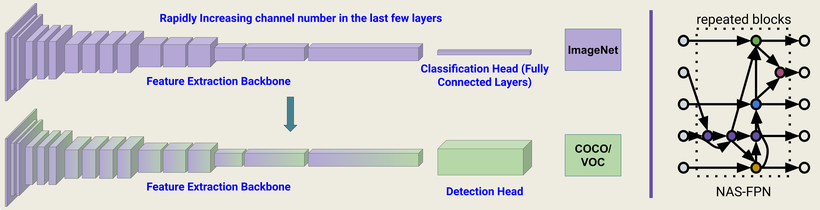
➢ Approaches:

- **Backbone Truncation**: Efficient adaptation of transfer learning weights from the pre-trained backbone by truncating the task-specific layers.
- **Raw Feature Collection and Redistribution (RFCR)**: Lightweight multi-scale feature interaction that exploits the absence of connections between non-adjacent feature scales in existing state-of-the-art methods.

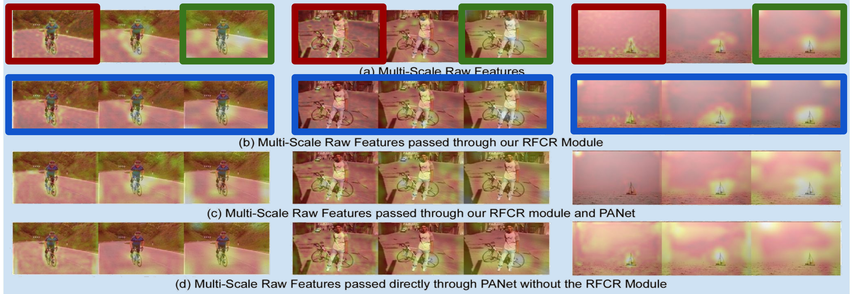
Motivation

➢ Later layers of the feature extraction backbone are extremely heavy and contains task-specific features.

➢ Neural architecture search in multi-scale feature interaction has explored connections between non-adjacent scales, but are not yet adapted by SOTA methods.



Results



➢ **Low-level features** are precise but noisy and inaccurate while **high-level features** are accurate but imprecise.

➢ RFCR combines them together to create **better overall features**.

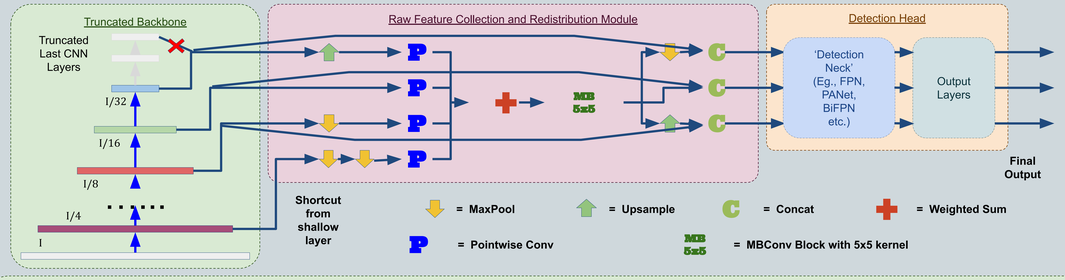
Model	Input Res.	Size (MB)	FPS Nano	FPS NX	FPS AGX	AP^{50} VOC	AP^{50} COCO
Tiny-YOLOv3	416	34.9	27.36	66.55	91.71	61.30	33.10
Tinier-YOLO	416	8.9	30.14	68.73	92.09	65.70	**34.00**
YOLO Nano	416	4.0	13.62	54.03‡	85.81‡	69.10	–
YOLO-Fastest	320	1.3	**42.41**	**76.13**	**126.82**	61.02	–
YOLO-Fastest XL	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

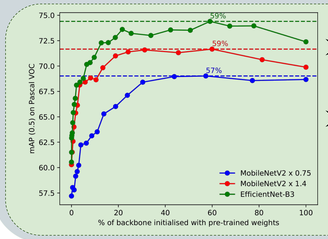
Model	Input Res.	FPS Nano	COCO AP	COCO AP^{50}	COCO AP^{75}
YOLOv4-tiny	416	29.55	21.7	40.2	22.5
YOLOv4-tiny+RFCR	416	27.81	**22.9**	**41.5**	**23.3**
YOLOv4-tiny (3l)	608	24.87	28.7	47.2	29.7
YOLOv4-tiny (3l)+ RFCR	608	21.40	**29.3**	**48.1**	**30.5**

➢ Improvements in both latency and accuracy across a number of feature extraction backbones and detection heads.

➢ Our RFCR module in YOLOv4-tiny improves performance to **41.5 mAP on COCO**, outperforming the original version by 1.3 mAP.

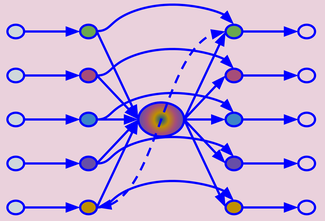
Complete Model Pipeline with Backbone Truncation and RFCR





➢ Initializing the last layers of the feature extraction backbone with transfer learning weights actually **hurts** the performance.

➢ Truncating these later but expensive layers of the model provides a better alternative to width reduction towards designing a lightweight object detection model.



➢ Minimum network fragmentation.

➢ Direct connections even for non-adjacent feature scales.

➢ Flexible settings for the number of input and output scales.

➢ Lightweight and compatible with various detection heads.