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

Prakhar Ganesh1*, Yao Chen1*, Yin Yang2, Deming Chen3, Marianne Winslett3

1Advanced Digital Sciences Center, Illinois at Singapore
2College of Science and Engineering, Hamad Bin Khalifa University, Qatar
3University of Illinois at Urbana-Champaign, USA

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.

Diagram: 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.

Results

- **Low-level features** are precise but noisy and inaccurate while **high-level features** are accurate but imprecise.
- RFCR combines them to create **better overall features**.
- 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.