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

Prakhar Ganesh\textsuperscript{1*}, Yao Chen\textsuperscript{1*}, Yin Yang\textsuperscript{2}, Deming Chen\textsuperscript{1,3}, Marianne Winslett\textsuperscript{1,3}

\textsuperscript{1} Advanced Digital Sciences Center, Illinois at Singapore
\textsuperscript{2} College of Science and Engineering, Hamad Bin Khalifa University, Qatar
\textsuperscript{3} University of Illinois at Urbana-Champaign, USA
Background & Motivation

- Transfer Learning in Object Detection
- Multi-Scale Feature Interaction
Transfer Learning in Object Detection

ImageNet

Feature Extraction Backbone

Rapidly Increasing channel number in the last few layers

Classification Head (Fully Connected Layers)

COCO/VOC

Feature Extraction Backbone

Detection Head
Transfer Learning in Object Detection

What if we had randomly initialised the **last layer**?
Transfer Learning in Object Detection

What if we had randomly initialised the last three layers?
What if we had used **no transfer learning at all?**
Transfer Learning in Object Detection

- Transfer learning plays an important role in model training, specially in a low data setting.
- Not every layer of a pre-trained model is equally useful. Initial layers are known to be task-agnostic, and last layers can be task-specific.
- Despite the existence of active research in transfer learning, most SOTA models in vision have not adapted to this behavior.

Multi-Scale Feature Interaction

Feature Extraction Backbone

Multi-Scale Feature Interaction

Detection Head

Segmentation Head

Tracking Head

... and many more
Multi-Scale Feature Interaction

FPN  PANet  NAS-FPN  BiFPN

repeated blocks

Multi-Scale Feature Interaction

- Existing work focuses on some combination of top-down and/or bottom-up approaches.
- With the increasing complexity of these modules, the tradeoff between accuracy and efficiency has started saturating.
- NAS-based architectures have revealed the importance of direct connections between non-adjacent feature scales.

YOLO-ReT

- Importance of Individual Layers
- Backbone Truncation
- Raw Feature Collection and Redistribution
Complete Model

Truncated Backbone

Raw Feature Collection and Redistribution Module

Detection Head

I

I/4

I/8

I/16

I/32

Truncated Last CNN Layers

Raw Feature Collection and Redistribution Module

Output Layers

Detection Neck’ (Eg., FPN, PANet, BiFPN etc.)

Final Outputs

\[ \text{I} \]

\[ \text{I/4} \]

\[ \text{I/8} \]

\[ \text{I/16} \]

\[ \text{I/32} \]

\[ \text{P} \]

\[ \text{MB} \]

\[ \text{5x5} \]

\[ \text{C} \]

\[ \text{P} \]

\[ \text{C} \]

\[ \text{P} \]

\[ \text{C} \]

\[ \text{P} \]

\[ \text{C} \]

\[ \text{P} \]

\[ \text{C} \]

\[ \text{P} \]

\[ \text{C} \]

\[ \text{MaxPool} \]

\[ \text{Upsample} \]

\[ \text{Concat} \]

\[ \text{Weighted Sum} \]

\[ \text{Pointwise Conv} \]

\[ \text{MBConv Block with 5x5 kernel} \]
Complete Model

Truncated Backbone
- Truncated Last CNN Layers
  - I/32
  - I/16
  - I/8
  - I/4

Raw Feature Collection and Redistribution Module
- Raw Feature Collection
- Redistribution Module
  - P
  - MB
  - 5x5
  - Concat
  - Weighted Sum
  - Pointwise Conv

Detection Head
- Detection Neck (Eg., FPN, PANet, BiFPN etc.)
- Output Layers

Final Outputs
- MaxPool
- Upsample
- Concat
- Weighted Sum
- Pointwise Conv

= MBConv Block with 5x5 kernel
Importance of Individual Layers

![Graph showing the mAP (0.5) on Pascal VOC for different networks as a function of the percentage of backbone initialized with pre-trained weights. The graph compares MobileNetV2 x 0.75, MobileNetV2 x 1.4, and EfficientNet-B3, with peaks indicating the percentage of layers that significantly impact the mAP.]
Backbone Truncation

- Initializing the last layers of the feature extraction backbone with transfer learning weights actually ‘hurts’ the performance.
- Since these last layers hold no transfer learning importance, they can be analysed purely from an architecture viewpoint.
- We propose that a truncated version of the feature extraction backbone is a better alternative to width reduction.
Raw Feature Collection and Redistribution
Raw Feature Collection and Redistribution

Simplistic Design

Minimal Network Fragmentation (Each Collection and Redistribution Path Can be Executed in Parallel)
Raw Feature Collection and Redistribution

Direct Connection Paths Even Between Non-Adjacent Scales
Raw Feature Collection and Redistribution

Independent of the number of Output Scales
Raw Feature Collection and Redistribution

Cannot replace the meticulousness provided by other Multi-Scale Feature Interaction methods.

But can be easily integrated as an additional feature processing.
Evaluation

- Experiment Setup
- Component Ablation
- State of the art Models
Experiment Setup

- We tested with 3 lightweight feature extraction backbones (MobileNetV2x0.75, x1.4, and EfficientNet-B3) and various feature interaction methods (FPN, PANet and BiFPN).
- We evaluated our methods on Pascal VOC and COCO datasets.
- We tested our models with on-device performance latencies, on Jetson Nano, Jetson Xavier NX and Jetson AGX Xavier.
Qualitative Heatmap Study

(a) Multi-Scale Raw Features

(b) Multi-Scale Raw Features passed through our RFCR Module

(c) Multi-Scale Raw Features passed through our RFCR module and PANet

(d) Multi-Scale Raw Features passed directly through PANet without the RFCR Module
Qualitative Heatmap Study

(a) Multi-Scale Raw Features
(b) Multi-Scale Raw Features passed through our RFCR Module
(c) Multi-Scale Raw Features passed through our RFCR module and PANet
(d) Multi-Scale Raw Features passed directly through PANet without the RFCR Module
Qualitative Heatmap Study

(a) Multi-Scale Raw Features

(b) Multi-Scale Raw Features passed through our RFCR Module

(c) Multi-Scale Raw Features passed through our RFCR module and PANet

(d) Multi-Scale Raw Features passed directly through PANet without the RFCR Module
Qualitative Heatmap Study

(a) Multi-Scale Raw Features

(b) Multi-Scale Raw Features passed through our RFCR Module

(c) Multi-Scale Raw Features passed through our RFCR module and PANet

(d) Multi-Scale Raw Features passed directly through PANet without the RFCR Module
## State of the art Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Input Resolution</th>
<th>FPS Nano</th>
<th>FPS NX</th>
<th>FPS AGX</th>
<th>FPS VOC</th>
<th>FPS COCO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tiny-YOLOv3</td>
<td>416</td>
<td>27.36</td>
<td>66.55</td>
<td>91.71</td>
<td>61.30</td>
<td>33.10</td>
</tr>
<tr>
<td>Tinier-YOLO</td>
<td>416</td>
<td>30.14</td>
<td>68.73</td>
<td>92.09</td>
<td>65.70</td>
<td>34.00</td>
</tr>
<tr>
<td>YOLO-ReT-MobileNetV2 x 0.75</td>
<td>320</td>
<td>33.19</td>
<td>71.64</td>
<td>95.97</td>
<td>68.75</td>
<td>34.91</td>
</tr>
<tr>
<td>YOLO Nano</td>
<td>416</td>
<td>13.62</td>
<td>54.03</td>
<td>85.81</td>
<td>69.10</td>
<td>--</td>
</tr>
<tr>
<td>YOLO-ReT-MobileNetV2 x 1.4</td>
<td>320</td>
<td>23.01</td>
<td>65.37</td>
<td>93.49</td>
<td>70.35</td>
<td>35.77</td>
</tr>
<tr>
<td>YOLO Fastest</td>
<td>320</td>
<td>42.41</td>
<td>76.13</td>
<td>126.82</td>
<td>61.02</td>
<td>--</td>
</tr>
<tr>
<td>YOLO-ReT-MobileNetV2 x 1.4</td>
<td>224</td>
<td>43.16</td>
<td>84.32</td>
<td>113.94</td>
<td>62.91</td>
<td>31.63</td>
</tr>
</tbody>
</table>
## State of the art Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Input Resolution</th>
<th>FPS Nano</th>
<th>FPS NX</th>
<th>FPS AGX</th>
<th>AP50 (Detailed Results in Paper)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tiny-YOLOv3</td>
<td>416</td>
<td>27.36</td>
<td>66.55</td>
<td>91.71</td>
<td>61.30 (Detailed Results in Paper)</td>
</tr>
<tr>
<td>Tinier-YOLO</td>
<td>416</td>
<td>30.14</td>
<td>68.73</td>
<td>92.09</td>
<td>65.70 (Detailed Results in Paper)</td>
</tr>
<tr>
<td>YOLO-ReT-MobileNetV2 x 0.75</td>
<td>320</td>
<td>33.19</td>
<td>71.64</td>
<td>95.97</td>
<td>68.75 (Detailed Results in Paper)</td>
</tr>
<tr>
<td>YOLO Nano</td>
<td>416</td>
<td>13.62</td>
<td>54.03</td>
<td>85.81</td>
<td>69.10 (Detailed Results in Paper)</td>
</tr>
<tr>
<td>YOLO-ReT-MobileNetV2 x 1.4</td>
<td>320</td>
<td>23.01</td>
<td>65.37</td>
<td>93.49</td>
<td>70.35 (Detailed Results in Paper)</td>
</tr>
<tr>
<td>YOLO Fastest</td>
<td>320</td>
<td>42.41</td>
<td>76.13</td>
<td>126.82</td>
<td>61.02 (Detailed Results in Paper)</td>
</tr>
<tr>
<td>YOLO-ReT-MobileNetV2 x 1.4</td>
<td>224</td>
<td>43.16</td>
<td>84.32</td>
<td>113.94</td>
<td>62.91 (Detailed Results in Paper)</td>
</tr>
</tbody>
</table>

Code available at: [github.com/prakharg24/yoloret](https://github.com/prakharg24/yoloret)
Thank You