RNNPool: Efficient Non-linear Pooling for RAM Constrained Inference

Oindrila Saha¹, Aditya Kusupati², Harsha Vardhan Simhadri¹, Manik Varma¹ and Prateek Jain¹
¹Microsoft Research ²University of Washington

Our goal: Accurate computer vision models that can be deployed on tiny devices

Barriers:
- CNN models have many layers with large activations
- Large memory footprint, the most constrained resource on microcontrollers
- Standard pooling operators (e.g. max pool) are gross aggregators, thus are only used with a maximum stride of 2
- RNNPool can reduce intermediate feature maps significantly (up to 16x) with small loss in accuracy
  - Ensures heavy convolution blocks run on smaller activation maps

Existing Works

Decreasing Model Size
- Pruning
- Sparsification

Model Compression

Decreasing Compute
- Depthwise Separable Convolutions
- MobileNets, EfficientNets

However, peak RAM requirement still remains high

Evaluation

Task 1: Image Classification
Results on ImageNet-10 dataset formed by subsampling 10 classes from ImageNet1K and MobileNetV2 as the base architecture

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>MAdds</th>
<th>#Params</th>
<th>RAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Network</td>
<td>94.2</td>
<td>0.30GB</td>
<td>5.2GB</td>
<td>2.25MB</td>
</tr>
<tr>
<td>Average Pooling</td>
<td>90.8</td>
<td>0.13GB</td>
<td>7.0GB</td>
<td>1.85MB</td>
</tr>
<tr>
<td>Max Pooling</td>
<td>92.8</td>
<td>0.20GB</td>
<td>7.0GB</td>
<td>2.05MB</td>
</tr>
<tr>
<td>Strided Conv</td>
<td>93.0</td>
<td>0.20GB</td>
<td>7.0GB</td>
<td>2.15MB</td>
</tr>
<tr>
<td>RNNPool</td>
<td>94.4</td>
<td>0.22GB</td>
<td>7.0GB</td>
<td>2.05MB</td>
</tr>
</tbody>
</table>

Task 2: Visual Wakeword
- Predict whether a person is present
- 8x less RAM
- 40% less compute

Task 3: Face Detection
Comparison with SOTA for very low MAdds category on WIDER FACE validation subset
- 188KB peak RAM
- 70M MAdds
- 160KB Model Size

10.45 sec/image on STM32F439-M4 device clocked at 168 MHz


Code: https://github.com/Microsoft/EdgeML