XNOR-NET: IMAGENET CLASSIFICATION USING BINARY CONVOLUTIONAL NEURAL NETWORKS

RYAN MCCAMPBELL

2019 SPRING @ HTTPS://QDATA.GITHUB.IO/DEEP2READ
PROBLEM

• Ordinary neural networks are expensive to store and evaluate
• Not good for embedded and mobile platforms
• How to make them more efficient?
APPROACHES

• More compact networks
• Compressing pre-trained networks
• Quantizing parameters
• Binarizing parameters/activations
BINARY-WEIGHT-NETWORK

• Idea: store weights as binary vectors, $\pm 1$, plus scale

$$W \approx \alpha B, \alpha > 0, B \in \{-1,1\}^n$$

$$I*W \approx (I \oplus B) \alpha$$

• $\oplus$: convolution using only addition/subtraction

• Eliminates most multiplications

• $\sim 32x$ storage reduction
**BINARY-WEIGHT-NETWORK**

- Minimize $J(B, \alpha) = \|W - \alpha B\|_2$

  $B^\star = \text{sign}(W)$

  $\alpha^\star = \sum |W_i| / n = 1 / n \|W\|_1$
TRAINING

• Binarize weights during forward pass and backwards pass
• Use full-parameter weights for update
• Approximate $\frac{\partial}{\partial x} \text{sign}(x) = 1[|r|\leq 1]$
XNOR-NET

• Binarize both weights and inputs

• Convolutions using efficient binary operations: shift, XNOR & bit-count

\[ I \ast W \approx \text{sign}(I) \odot \text{sign}(W) \odot K \alpha \]

• \[ K_{ij} = \beta = 1/n \| \text{subtensor of } X \text{ at } ij \|_1 \]

• \[ K = A \ast k, A = \sum |I_{ij}|, k_{ij} = 1/wh \]
TRAINING XNOR-NET

- Batch normalization first
- Binary activation: compute $K$ and $\text{sign}(l)$
- Binary convolution
- Pool after convolution
RESULTS
## RESULTS

<table>
<thead>
<tr>
<th></th>
<th>Binary-Weight</th>
<th>Binary-Input-Binary-Weight</th>
<th>Full-Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BWN</strong></td>
<td>56.8</td>
<td>44.2</td>
<td>56.6</td>
</tr>
<tr>
<td><strong>BC[11]</strong></td>
<td>79.4</td>
<td>69.2</td>
<td>80.2</td>
</tr>
<tr>
<td><strong>XNOR-Net</strong></td>
<td>35.4</td>
<td>27.9</td>
<td></td>
</tr>
<tr>
<td><strong>BNN[11]</strong></td>
<td>61.0</td>
<td>50.42</td>
<td></td>
</tr>
<tr>
<td><strong>AlexNet[1]</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Top-1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Top-5</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>