A Data and Compute Efficient Design for Limited-Resources Deep Learning

Mirgahney Mohamed*
African Institute for Mathematical Sciences
Intern at Qualcomm AI Research

Gabriele Cesa*
Qualcomm AI Research

Taco S. Cohen
Qualcomm AI Research

Max Welling
Qualcomm AI Research
Machine Learning for Developing Countries

- Eg. aid for medical diagnosis
- Challenges in deploying SOTA solutions
  - Constrained computational resources
  - Limited data available
- Compute Efficiency
- Data Efficiency
  - Better generalization with less data
Our Solution

• **Compute Efficiency:**
  - Use light weight models
  - Weight and activation quantization

• **Data Efficiency:**
  - Equivariance: exploit data symmetry to achieve improved generalization
Equivariance

A short introduction

Conventional CNNs: Translation Equivariant

Rotations?
Architecture

• MobileNetV2

• Equivariant version based on [1]
  • Group convolutional design [2]
  • Preserve computational cost
  • Reduce trainable parameters
  • Equivariance to 12 rotations

• Strided conv: adapt padding and input resolution
  • Avoid artifacts on 90° rotations
  • Improve overall stability also on continuous angles

---

Architecture

• Preserve rotational symmetry of data and features

• Circular mask
  • Input images
  • Global spatial pooling
Quantization

- Reduce precision of weighs and activations from FLOAT32 to INT8
- Optimize models with the data-free quantization techniques from [3]:
  - Cross-layer range equalization
  - High-bias absorption
- Does not break 90° rotation equivariance
- Equivariance to <90° rotations marginally affected

Results on Patch Camelyon (PCam) [4]

Table 1: Test accuracy on PCam

<table>
<thead>
<tr>
<th>Model</th>
<th>Full-Precision</th>
<th>Quantized (INT8)</th>
<th>Diff (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional MobileNetV2</td>
<td>84.67 ± 1.91</td>
<td>84.32 ± 1.76</td>
<td>-0.4%</td>
</tr>
<tr>
<td>Equivariant MobileNetV2</td>
<td>89.19 ± 0.79</td>
<td>88.94 ± 0.66</td>
<td>-0.3%</td>
</tr>
<tr>
<td>Equivariant DenseNet[4] Veeling et al. (2018)</td>
<td>89.8</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Conclusion

• Combine two independent lines of research to improve data and compute efficiency

• Equivariance in small architecture regime

• Quantization techniques [3] are compatible with equivariant networks

Thank you

Follow us on: f /twitter /linkedin /instagram
For more information, visit us at:
www.qualcomm.com &
www.qualcomm.com/blog

Nothing in these materials is an offer to sell any of the components or devices referenced herein.

©2018-2019 Qualcomm Technologies, Inc. and/or its affiliated companies. All Rights Reserved.

Qualcomm is a trademark of Qualcomm Incorporated, registered in the United States and other countries. Other products and brand names may be trademarks or registered trademarks of their respective owners.

References in this presentation to “Qualcomm” may mean Qualcomm Incorporated, Qualcomm Technologies, Inc., and/or other subsidiaries or business units within the Qualcomm corporate structure, as applicable. Qualcomm Incorporated includes Qualcomm’s licensing business, QTL, and the vast majority of its patent portfolio. Qualcomm Technologies, Inc., a wholly-owned subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of Qualcomm’s engineering, research and development functions, and substantially all of its product and services businesses, including its semiconductor business, QCT.