

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 Al Research Max Welling Qualcomm Al 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

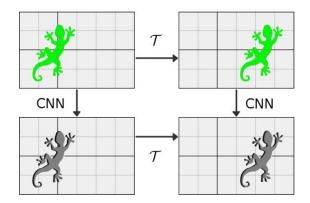
• Data Efficiency:

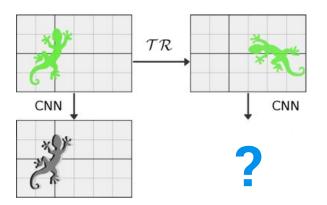
• Equivariance: exploit data symmetry to achieve improved generalization



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

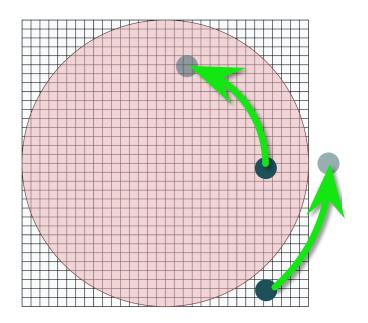
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | |
|----|----|----|----|----|----|----|----|--------------|
| 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 90° rotation |
| 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | |
| 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | \sim |
| 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | / |
| 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | |
| 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | |
| 57 | 58 | 59 | 60 | 61 | 62 | 63 | 64 | |

| 57 | 49 | 41 | 33 | 25 | 17 | 9 | 1 |
|----|----|----|----|----|----|----|---|
| 58 | 50 | 42 | 34 | 26 | 18 | 10 | 2 |
| 59 | 51 | 43 | 35 | 27 | 19 | 11 | 3 |
| 60 | 52 | 44 | 36 | 28 | 20 | 12 | 4 |
| 61 | 53 | 45 | 37 | 29 | 21 | 13 | 5 |
| 62 | 54 | 46 | 38 | 30 | 22 | 14 | 6 |
| 63 | 55 | 47 | 39 | 31 | 23 | 15 | 7 |
| 64 | 56 | 48 | 40 | 32 | 24 | 16 | 8 |

Strided convolution over image with even size breaks equivariance to 90° rotations

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

| Model | Full-Precision | Quantized (INT8) | |
|--|-----------------------|------------------|-------|
| Conventional MobileNetV2 | 84.67 ± 1.91 | 84.32 ± 1.76 | -0.4% |
| Equivariant MobileNetV2 | 89.19 ± 0.79 | 88.94 ± 0.66 | -0.3% |
| Equivariant DenseNet Veeling et al. (2018) | 89.8 | - 1 | - |

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

Qualcom

Thank you

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.