Qualcomm AI research

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

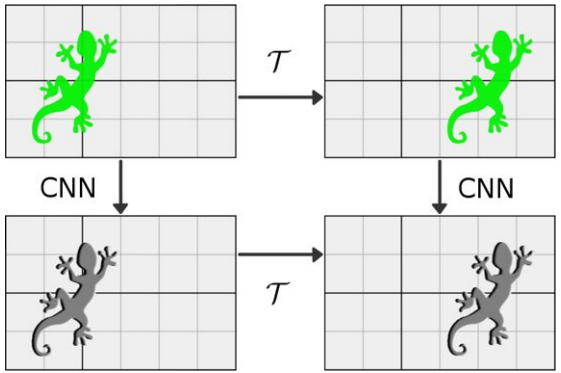
Our Solution

- **Compute Efficiency**:
  - Use light weight models
  - Weight and activation quantization
- **Data Efficiency**:
  - Equivariance: exploit data symmetry to achieve improved generalization

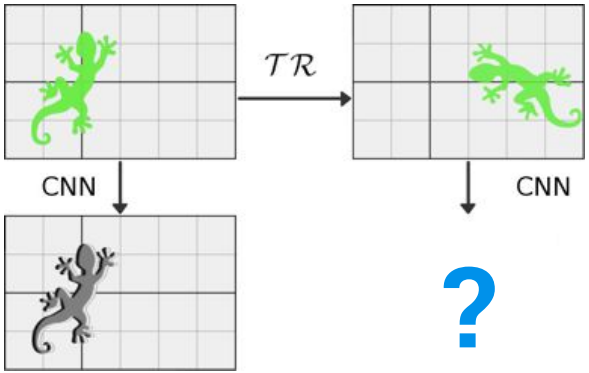
Equivariance

A short introduction

Conventional CNNs: Translation Equivariant



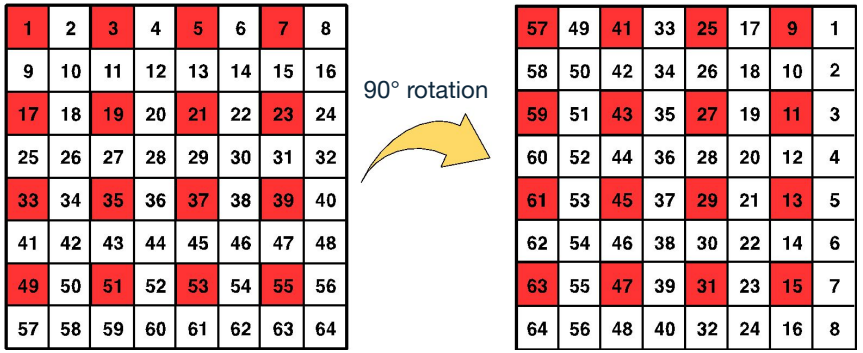
Rotations?



Images adapted from Maurice Weiler

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

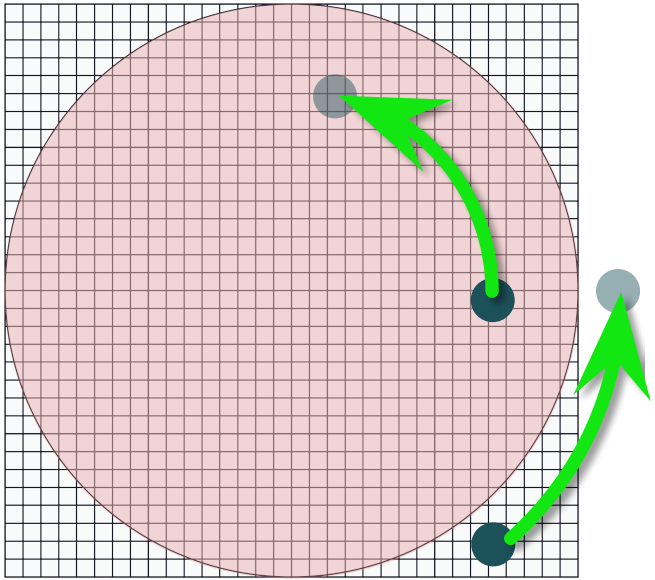


Strided convolution over image with even size breaks equivariance to 90° rotations

[1] Weiler, M. and Cesa, G. General E(2)-Equivariant Steerable CNNs. 2019.
[2] Cohen, T. and Welling, M. Group Equivariant Convolutional Networks, 2016.

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

[3] Nagel, M., van Baalen, M., Blankevoort, T., and Welling, M. Data-Free Quantization through Weight Equalization and Bias Correction. 2019.

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	-	-

[4] Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., and Welling, M. Rotation equivariant CNNs for digital pathology, 2018.

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

[3] Nagel, M., van Baalen, M., Blankevoort, T., and Welling, M. Data-Free Quantization through Weight Equalization and Bias Correction. 2019.

Qualcomm

Thank you

Follow us on:

For more information, visit us at:

www.qualcomm.com &
www.qualcomm.com/blog