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

Anonymous authors
Paper under double-blind review

Abstract

Thanks to their improved data efficiency, equivariant neural networks have gained increased interest in the deep learning community. They have been successfully applied in the medical domain where symmetries in the data can be effectively exploited to build more accurate and robust models. To be able to reach a much larger body of patients, mobile, on-device implementations of deep learning solutions have been developed for medical applications. However, equivariant models are commonly implemented using large and computationally expensive architectures, not suitable to run on mobile devices. In this work, we design and test an equivariant version of MobileNetV2 and further optimize it with model quantization to enable more efficient inference. We achieve close-to state of the art performances on the Patch Camelyon(PCam) medical dataset while being more computationally efficient.

1 Introduction

Deep learning has recently moved closer to edge devices Chen & Ran (2019), creating opportunities for many new applications where data needs to be analyzed in real time. However, this development opens new challenges to improve power consumption, computational efficiency and memory footprint Rallapalli et al. (2016). Many efforts have been directed toward compute, memory, and power efficient architecture design Sandler et al. (2018); Cai et al. (2019); Tan & Le (2019); Howard et al. (2019). Additionally, methods like compression Han et al. (2015b;a); Kuzmin et al. (2019) and quantization Meller et al. (2019); Nagel et al. (2019); Cai et al. (2020) have become popular.

These works become especially important in developing countries where the constrained resources make deploying state of the art models challenging De-Arteaga et al. (2018); Sinha et al. (2019). Computer vision and deep learning algorithms can provide low-cost solutions where human experts are not available Wahl et al. (2018). For instance, they can power automatic systems which help doctors in performing diagnosis or can be combined with drones to perform aerial imaging, e.g. to monitor disaster areas Kyrkou & Theocharides (2019). Quinn et al. (2016) develop a system for point-of-care diagnostics which uses mobile phones and microscopes to automate the diagnosis of different diseases.

In many cases, computational power is not the only limiting resource: gathering large quantities of labelled data is often prohibitively expensive. In this context, equivariance has been found to be a useful design choice, improving data efficiency through built-in knowledge about the symmetries of the problem Cohen & Welling (2016); Worrall et al. (2017); Weiler et al. (2018). Equivariant networks guarantee pre-determined transformations of their outputs under corresponding transformations of the input signals, enabling them to easily generalize over transformed signals. For these reasons, they have been successfully applied to medical imaging as well as aerial imaging, where symmetries are common Bekkers et al. (2018); Veeling et al. (2018); Winkels & Cohen (2018); Li et al. (2018); Chidester et al. (2019); Dieleman et al. (2016); Hoogeboom et al. (2018).

Unfortunately, the demands of data and compute efficiency can be at odds. Indeed, equivariant networks often rely on expensive architectures. Their data efficiency is also often exploited to build larger models without overfitting. However, this is not always affordable in real world applications, especially under computational constraints as on handheld devices. Moreover, for very small model sizes, increased weight sharing implied by equivariance can limit the number of distinct learnable filters too much, potentially harming the expressiveness of the model. It is therefore not clear yet if equivariance can still be combined with efficient architecture design. In this work: (i) We show that equivariance is a useful design choice even with limited computational resources and in small model regime. (ii) We show that quantization can be used to increase efficiency without harming equivariance, preserving the stability of the model. To the best of our knowledge, this work is the first to combine quantization with equivariant networks.

2 Related Works

Equivariance offers a principled way to design models when the problem of interest presents certain symmetries. In particular, it has proven very successful in image processing and, therefore, it has been extensively studied in this context. Cohen & Welling (2016) generalize convolutional networks beyond translations to arbitrary discrete groups but only consider $\frac{\pi}{2}$ rotations. Following works Weiler et al. (2018); Bekkers et al. (2018); Cheng et al. (2019); Bekkers (2020) extend group convolutions to arbitrary discrete rotations. Worrall et al. (2017) achieve continuous rotation equivariance using steerable filters. Cohen & Welling (2017) describe steerable networks for finite groups, covering group convolution as a special case. Weiler & Cesa (2019) extend steerable networks to all planar isometries. See Kondor & Trivedi (2018); Cohen et al. (2018) for further generalizations.

Equivariance has recently been applied in a number of medical imaging problems. Veeling et al. (2018) builds a reflection and $\frac{\pi}{2}$ rotation equivariant DenseNet Huang et al. (2017) for histopatological tissues classification. The fully-convolutional design allows them to also use this model for efficient segmentation of whole-slide images. Following works exploit equivariance to additional rotations Bekkers (2020) or scale Worrall & Welling (2019) on the same task. In the context of segmentation, most works Linmans et al. (2018); Li et al. (2018); Chidester et al. (2019) implement versions of U-Net Ronneberger et al. (2015) only equivariant to $\frac{\pi}{2}$ rotations and reflections.

The interest in bringing deep learning on edge devices has inspired new research to improve the efficiency of the models. Recent works design new models optimizing both performances and number of FLOPS required Sandler et al. (2018); Cai et al. (2019); Tan & Le (2019); Howard et al. (2019). Other lines of research, instead, focus on optimizing existing models to reduce their cost. In particular, quantization involves reducing the precision used to store the weights and the activations, typically to 8-bit integers Jacob et al. (2017). To preserve the full-precision performances, it is often necessary to change the architecture Sheng et al. (2018), perform additional training Ullrich et al. (2017); Louizos et al. (2018) or adapt the training procedure Zhou et al. (2016). Krishnamoorthi (2018); Nagel et al. (2019) propose solutions which do not require data or additional training. In this work, we adopt the data-free quantization methods from Nagel et al. (2019).

3 Background on Equivariant Networks

We interpret a neural network Φ as a sequence of layers $\{\phi_i\}_{i=1}^n$, each mapping from an input feature space F_{i-1} to an output feature space F_i. A feature space is the vector space containing the features produced by a layer of the network. In many cases, we have prior knowledge about the symmetries of the problem. For instance, we know that medical images may appear with arbitrary orientation and that their labels are independent from their rotations. These symmetries are modeled by a group G and its action on the data, e.g. a rotation of the images. It is therefore desirable to build a neural network Φ which guarantees the following property (**equivariance**):

$$\forall g \in G, f \in F_0 \quad \Phi(g \cdot f) = g \cdot \Phi(f) \tag{1}$$

Most approaches achieve this by enforcing equivariance at each layer ϕ_i of the network. This requires the definition of an action of the group on each intermediate feature space F_i during the design of the model. Each layer, then, needs to satisfy the equivariance property Eq. (1) with respect to the group action on its own input and output feature spaces. In the case of $2D$-convolutional neural networks, a feature map $f \in \mathbb{R}^{c\times h\times w}$ can be interpreted as a *feature field*, i.e. a function [1] $f : \mathbb{R}^2 \rightarrow \mathbb{R}^c$ associating a c-dimensional *feature vector* to each point on the plane $\mathbb{R}^2$. The spatial structure of the feature maps implicitly defines the action of the translation group on them (by translating the points on the plane) and the use of convolution guarantees translation equivariance.

[1] The input images and features are assumed to be continuous signals over $\mathbb{R}^2$ and not discretized on a grid.

We are generally interested in exploiting a larger class of symmetries; for example, we can additionaly consider the group of N discrete rotations containing $\{r\frac{2\pi}{N}\}_{r=0}^{N-1}$. Assuming translation equivariance is already guaranteed by the use of convolution, we denote with G only the rotations. More precisely, the c channels of the feature field are split into c/N blocks of size N. The group element $r\frac{2\pi}{N}$, then, cyclically shifts the channels in the same block by r positions. The resulting equivariant model is effectively a group convolutional network. Fig 1 shows the action of the group of 4 rotations over a convolutional feature field with $c = N = 4$ channels.

Enforcing rotation equivariance in a convolutional layer requires its filters to live in a lower dimensional vector space. An equivariant filter can therefore be built by finding a basis for this space and learning the coefficients to linearly combine it. For further details on the constraints on the layers of a network required by equivariance to the isometries of the plane, see Weiler & Cesa (2019).

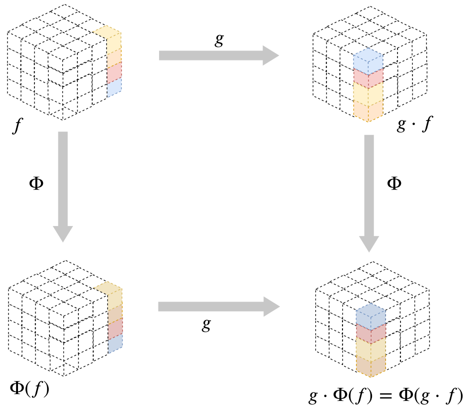
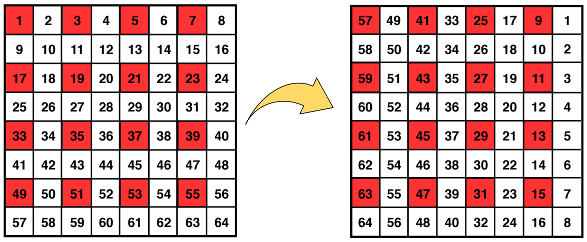


Figure 1: Action of the group of 4 rotations on a convolutional feature f (1st row): $g = \frac{\pi}{2}$ moves the pixels on the plane and transforms each feature vector $f(x) \in \mathbb{R}^4$ by circularly shifting its channels. A similar action is defined on the output of a neural network Φ (2nd row). If Φ is equivariant, it commutes with these two actions (bottom right).

Unfortunately, because images are sampled on a pixel grid, only rotations by multiples of $\frac{\pi}{2}$ are perfect symmetries. In order to consider rotations by arbitrary angles smaller than $\frac{\pi}{2}$, a common approach is defining filters in terms of a finite basis of steerable continuous filters Worrall et al. (2017); Weiler et al. (2018); Cheng et al. (2019); Weiler & Cesa (2019). However, it is important that the continuous basis is band-limited: as discussed in Weiler & Cesa (2019), sampling high-frequency filters on a grid can introduce non-equivariant elements in the basis due to aliasing.

4 Method

In order to study how equivariance is affected when limited computational resources are available, we experiment with MobileNetV2 Sandler et al. (2018). To implement an equivariant version of MobileNetV2, we build an equivariant depth-wise convolution. In our equivariant model, we preserve the same architecture, i.e. we use the same number of channels and the same filter sizes. Rotations by angles $\theta < \frac{\pi}{2}$ move the pixels in the corners outside the borders of the pixel grid. In order to enforce the rotational symmetry of the images, we apply a circular binary mask to the input of a model. Moreover, the global spatial pooling at the end of MobileNetV2 is combined with a similar mask to have a circular field of view. The use of strided convolution can harm the stability of the model under rotations of the same input image. For instance, Fig 2 shows that a strided 3×3 convolution over an even-sized feature map samples different points when the feature map is rotated by $\frac{\pi}{2}$. In order to preserve perfect $\frac{\pi}{2}$ rotation equivariance, we use odd-sized inputs and adapt stride and padding of convolution layers to maintain odd-sized pixel grids.



1	2	**3**	4	**5**	6	**7**	8
9	10	11	12	13	14	15	16
17	18	**19**	20	**21**	22	**23**	24
25	26	27	28	29	30	31	32
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

Figure 2: Strided convolution on an even-sized grid samples different points when the input is rotated, breaking equivariance.

4.1 Data-Free Quantization

An efficient architecture design is often not enough to run deep networks on edge devices. This means that other methods to reduce the computational costs are needed. In this work, we use the *data free quantization* (DFQ) methods from Nagel et al. (2019) which we show being compatible with an equivariant design. Because an equivariant network can be converted to a conventional one after training, we can easily apply these techniques on our models. In particular, we will use *cross-layer range equalization* and *high bias absorption*.

Cross-layer range equalization Exploiting the commutative property of the ReLU function with respect to scaling of its input, it is possible to adapt the weights of a trained network for better quantization performance. Nagel et al. (2019) use this property to equalize the range of values of the weights attached to each channel in a layer in order to maximize the channel precision after quantization.

Using the same notation as Nagel et al. (2019), consider a pair of fully connected layers such that:

$$\boldsymbol{y} = W^{(2)} f(W^{(1)} \boldsymbol{x} + \boldsymbol{b}^{(1)}) + \boldsymbol{b}^{(2)} \ .$$

Equalization involves scaling the weights with a positive diagonal matrix S as

$$\boldsymbol{y} = W^{(2)} S f(S^{-1} W^{(1)} \boldsymbol{x} + S^{-1} \boldsymbol{b}^{(1)}) + \boldsymbol{b}^{(2)} \tag{2}$$

The diagonal of S contains elements $s_i = \frac{\sqrt{r_i^{(1)} r_i^{(2)}}}{r_i^{(2)}}$, where $r_i^{(1)}$ and $r_i^{(2)}$ are respectively the ranges of values of the incoming weights $W_{i,:}^{(1)}$ and the outgoing weights $W_{:,i}^{(2)}$. In a convolutional network, this is applied for each pixel and the matrices $W^{(*)}$ correspond to the convolutional kernels. Assuming that the model is equivariant to $N = 4$ discrete rotations and that the transformation law described in Sec 3 is used in all feature maps, this scaling is equivariant only if all the N channels in the same block are scaled by the same factor. Because of equivariance, the filters which map to different channels in the same block are rotations of each other and, therefore, share the same values. A similar argument holds for the outgoing filters. As a result, both terms $r_i^{(1)}$ and $r_i^{(2)}$ are shared within each N-dimensional block. For $N > 4$, this does not theoretically hold anymore because the rotation of the filters involves interpolation. Nevertheless, because we apply band-limiting, rotated filters usually contain similar values. Indeed, we observe that equivariance is only marginally affected in practice.

High-bias absorption The equalization performed in Eq. (2) scales also the bias. This can increase the range of values it takes and, therefore, potentially, the ranges of the activations. If the distribution of the inputs is concentrated in the positive domain of ReLU, it is possible to exploit its linear behavior to shift the input distribution and absorb part of the bias in the following layer. Using batch normalization statistics, we can estimate the distribution and correct the bias without additional data. Because an equivariant batch normalization layer shares statistics across each channel in the same block, it is guaranteed that the bias values of the channels in the same field are shifted by the same amount, preserving the equivariance.

5 Experiments

Table 1: Test accuracy on PCam

Model	Full-Precision	Quantized (INT8)	
Conventional MobileNetV2	86.05	85.10	-1.1%
Equivariant MobileNetV2	88.90	88.16	-0.8%
Equivariant DenseNet Veeling et al. (2018)	89.8	-	-

We evaluate our classification models on the PatchCamelyon (PCam) dataset Veeling et al. (2018). It contains 96 × 96px images extracted from histopathologic scans and labeled based on the presence of metastatic tissue in their central 32 × 32px region. The surrounding pixels do not influence the label but only provide context. We consider two models: a conventional and an equivariant MobileNetV2, both adapted as described in the previous section for a fair comparison. The input images are cropped to 95 × 95px to have odd-size. Moreover, we reduce the stride in two layers to adapt the model to the lower resolution of this dataset. In the equivariant model, we consider the group of 12 rotations as Weiler et al. (2018); Bekkers et al. (2018); Weiler & Cesa (2019) found only smaller improvements by using more.

We train both models for 300 epochs and select the set of weights with the lowest loss on the validation set. As in Veeling et al. (2018), we augment the training set with $\frac{\pi}{2}$ rotations and reflections. To study the effect of quantization on equivariant models, we first apply the equalization techniques described in Sec 4.1 and then reduce the weights and activations precision to INT8. The test accuracies of both the conventional and the equivariant models, before and after quantization are shown in Tab 1. As expected, enforcing equivariance leads to a significant improvements in the full-precision model. The same improvement is preserved when the models are quantized, proving the quantization techniques applied are compatible with equivariance.

6 Conclusion

Deep neural networks are known to require large amount of data to train and, often, complex and expensive architectures to obtain state-of-the-art performances. However, the limited resources available in developing countries make deploying deep learning solutions challenging. In this work, we try to solve this problem by combining two so far independent lines of research, quantization and equivariance, to achieve improved generalization and efficient inference. In particular, we show that equivariant networks can be efficiently implemented and quantized without losing their desirable properties or reducing their expressive power.

References

Erik J Bekkers. B-spline CNNs on lie groups. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=H1gBhkBFDH`.

Erik J. Bekkers, Maxime W Lafarge, Mitko Veta, Koen A.J. Eppenhof, Josien P.W. Pluim, and Remco Duits. Roto-translation covariant convolutional networks for medical image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018.

Han Cai, Ligeng Zhu, and Song Han. ProxylessNAS: Direct neural architecture search on target task and hardware. In *International Conference on Learning Representations*, 2019. URL `https://arxiv.org/pdf/1812.00332.pdf`.

Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. Zeroq: A novel zero shot quantization framework, 2020.

J. Chen and X. Ran. Deep learning with edge computing: A review. *Proceedings of the IEEE*, 107 (8):1655–1674, Aug 2019. ISSN 1558-2256. doi: 10.1109/JPROC.2019.2921977.

Xiuyuan Cheng, Qiang Qiu, Robert Calderbank, and Guillermo Sapiro. RotDCF: Decomposition of convolutional filters for rotation-equivariant deep networks. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=H1gTEj09FX`.

Benjamin Chidester, That-Vinh Ton, Minh-Triet Tran, Jian Ma, and Minh N. Do. Enhanced Rotation-Equivariant U-Net for Nuclear Segmentation. pp. 0–0, 2019.

Taco S. Cohen and Max Welling. Group Equivariant Convolutional Networks. 48, 2016.

Taco S. Cohen and Max Welling. Steerable CNNs. In *International Conference on Learning Representations (ICLR)*, 2017.

Taco S. Cohen, Mario Geiger, and Maurice Weiler. A general theory of equivariant CNNs on homogeneous spaces. *arXiv preprint arXiv:1811.02017*, 2018.

Maria De-Arteaga, William Herlands, Daniel B. Neill, and Artur Dubrawski. Machine learning for the developing world. *ACM Trans. Manage. Inf. Syst.*, 9(2), August 2018. ISSN 2158-656X. doi: 10.1145/3210548. URL `https://doi.org/10.1145/3210548`.

Sander Dieleman, Jeffrey De Fauw, and Koray Kavukcuoglu. Exploiting cyclic symmetry in convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 2016.

Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. 2015a. cite arxiv:1510.00149Comment: Published as a conference paper at ICLR 2016 (oral).

Song Han, Jeff Pool, John Tran, and William Dally. Learning both Weights and Connections for Efficient Neural Network. In C Cortes, N D Lawrence, D D Lee, M Sugiyama, and R Garnett (eds.), *Advances in Neural Information Processing Systems 28*, pp. 1135–1143. Curran Associates, Inc., 2015b.

Emiel Hoogeboom, Jorn W. T. Peters, Taco S. Cohen, and Max Welling. HexaConv. In *International Conference on Learning Representations (ICLR)*, 2018.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for MobileNetV3. 2019.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:2261–2269, 2017. doi: 10.1109/CVPR.2017.243.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew G. Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. *CoRR*, abs/1712.05877, 2017. URL `http://arxiv.org/abs/1712.05877`.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning (ICML)*, 2018.

Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. 2018. URL `http://arxiv.org/abs/1806.08342`.

Andrey Kuzmin, Markus Nagel, Saurabh Pitre, Sandeep Pendyam, Tijmen Blankevoort, and Max Welling. Taxonomy and evaluation of structured compression of convolutional neural networks, 2019.

Christos Kyrkou and Theocharis Theocharides. Deep-learning-based aerial image classification for emergency response applications using unmanned aerial vehicles. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

Xiaomeng Li, Lequan Yu, Chi-Wing Fu, and Pheng-Ann Heng. Deeply supervised rotation equivariant network for lesion segmentation in dermoscopy images. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, - and - Skin Image Analysis - First International Workshop, OR 2.0 2018, 5th International Workshop, CARE 2018, 7th International Workshop, CLIP 2018, Third International Workshop, ISIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16 and 20, 2018, Proceedings*, pp. 235–243, 2018. doi: 10.1007/978-3-030-01201-4_25.

Jasper Linmans, Jim Winkens, Bastiaan S. Veeling, Taco S. Cohen, and Max Welling. Sample efficient semantic segmentation using rotation equivariant convolutional networks. *CoRR*, abs/1807.00583, 2018.

Christos Louizos, Matthias Reisser, Tijmen Blankevoort, Efstratios Gavves, and Max Welling. Relaxed quantization for discretized neural networks. *CoRR*, abs/1810.01875, 2018. URL `http://arxiv.org/abs/1810.01875`.

Eldad Meller, Alexander Finkelstein, Uri Almog, and Mark Grobman. Same, same but different: Recovering neural network quantization error through weight factorization. In *International Conference on Machine Learning, ICML 2019*, 2019.

Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-Free Quantization through Weight Equalization and Bias Correction. 2019.

John A Quinn, Rose Nakasi, Pius K. B. Mugagga, Patrick Byanyima, William Lubega, and Alfred Andama. Deep convolutional neural networks for microscopy-based point of care diagnostics. In Finale Doshi-Velez, Jim Fackler, David Kale, Byron Wallace, and Jenna Wiens (eds.), *Proceedings of the 1st Machine Learning for Healthcare Conference*, volume 56 of *Proceedings of Machine Learning Research*, pp. 271–281, Children's Hospital LA, Los Angeles, CA, USA, 18–19 Aug 2016. PMLR. URL `http://proceedings.mlr.press/v56/Quinn16.html`.

Swati Rallapalli, Hang Qiu, Archith John Bency, S. Karthikeyan, Ramesh Govindan, B. S. Manjunath, and Rahul Urgaonkar. Are very deep neural networks feasible on mobile devices. 2016.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018. ISSN 10636919. doi: 10.1109/CVPR.2018.00474.

Tao Sheng, Chen Feng, Shaojie Zhuo, Xiaopeng Zhang, Liang Shen, and Mickey Aleksic. A quantization-friendly separable convolution for mobilenets. *CoRR*, abs/1803.08607, 2018. URL `http://arxiv.org/abs/1803.08607`.

Anugraha Sinha, Naveen Kumar, Murukesh Mohanan, MD Muhaimin Rahman, Yves Quemener, Amina Mim, and Suzana Ilić. Quantized deep learning models on low-power edge devices for robotic systems, 2019.

Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. 2019.

Karen Ullrich, Edward Meeds, and Max Welling. Soft weight-sharing for neural network compression, 2017. URL `https://arxiv.org/abs/1702.04008`.

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant CNNs for digital pathology. June 2018.

Brian Wahl, Aline Cossy-Gantner, Stefan Germann, and Nina R Schwalbe. Artificial intelligence (ai) and global health: how can ai contribute to health in resource-poor settings? *BMJ Global Health*, 3(4), 2018. doi: 10.1136/bmjgh-2018-000798. URL `https://gh.bmj.com/content/3/4/e000798`.

Maurice Weiler and Gabriele Cesa. General $E(2)$-Equivariant Steerable CNNs. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Maurice Weiler, Fred A. Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant CNNs. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Marysia Winkels and Taco S. Cohen. 3D G-CNNs for pulmonary nodule detection. In *Conference on Medical Imaging with Deep Learning (MIDL)*, 2018.

Daniel E. Worrall and Max Welling. Deep scale-spaces: Equivariance over scale. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Daniel E. Worrall, Stephan J. Garbin, Daniyar Turmukhambetov, and Gabriel J. Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv preprint arXiv:1606.06160*, 2016. URL `https://arxiv.org/abs/1606.06160`.