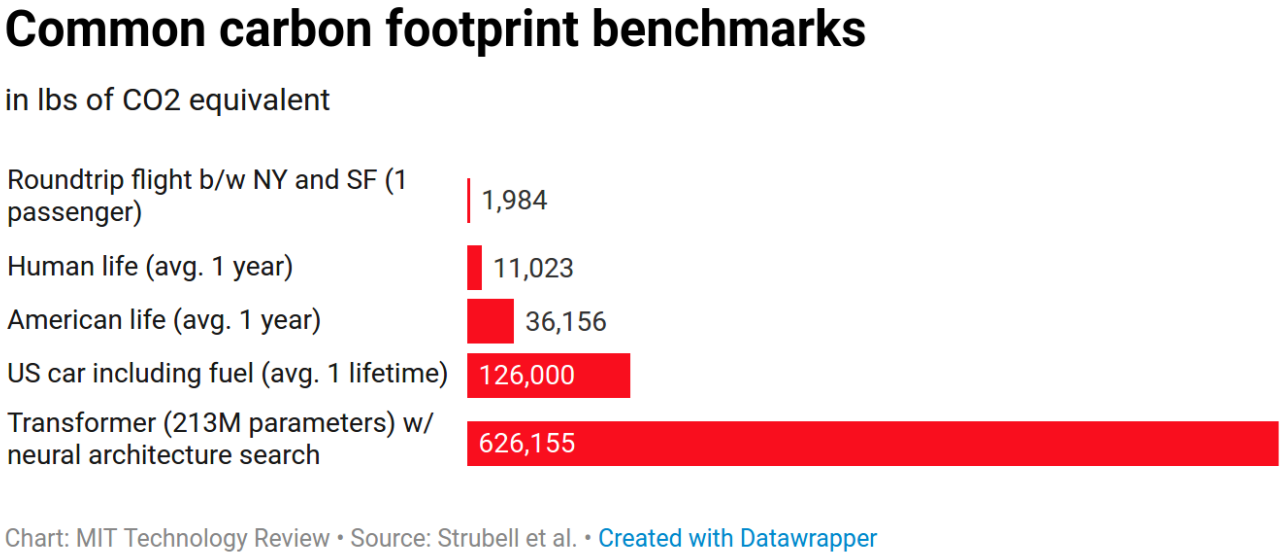
BNNs – Binary Neuronal Networks –

Making AI-Training Energy-efficient

Prof. Dr. Christoph Meinel
Dean, Institute Director, and CEO
Hasso Plattner Institute, Germany

Deep Learning Models **Spend Lots of Energy**

An extreme example during training of very large CNNs:



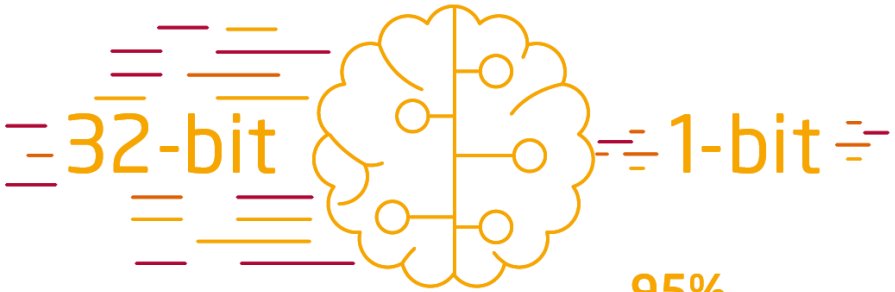
Strubell, Emma, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP." In the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Florence, Italy. July 2019

Prof. Dr. Ch. Meinel
Dean and Director
HPI, Germany

HPI clean-IT Initiative – Energy-efficient AI Training: Deep Learning with **Binary Neuronal Networks**

Binary neuronal networks – BNNs

- State of the art deep neuronal networks are trained and operate on 32-bit models
- Design and Training of deep neuronal networks on binary-level (1-bit) is possible

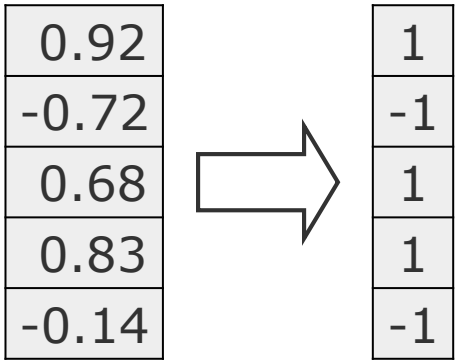


Energy-efficient AI with BNN

Prof. Dr. Ch. Meinel
Dean and Director
HPI, Germany

Low-bit Neural Networks

- The extreme case **Binary Neuronal Networks** only use +1 and -1 for weights and inputs instead of 32-bit floating point numbers



- Up to 32x model compression and 58x speedup during inference[1]
- **More than 1000x energy saving on dedicated hardware[2]**

Energy-efficient AI with BNN

Prof. Dr. Ch. Meinel
Dean and Director
HPI, Germany

[1] Rastegari, Mohammad, et al. "Xnor-net: Imagenet Classification using Binary Convolutional Neural Networks." European conference on computer vision. Springer, Cham, 2016.

[2] Mishra, Asit, et al. "WRPN: Wide Reduced-Precision Networks." International Conference on Learning Representations. 2018.

Challenges of **Low-bit Networks**

- Loss of accuracy compared to 32-bit networks
  - for example, directly binarizing a network trained on ImageNet, leads to a loss in accuracy of about 10% [1]
- We belief that we can shrink that gap

The goal of our ongoing research work:

- To achieve the same accuracy with binary networks as with "traditional" CNN

Prof. Dr. Ch. Meinel
Dean and Director
HPI, Germany

[1] Joseph Bethge, Haojin Yang, Marvin Bornstein, Christoph Meinel. "BinaryDenseNet: Developing an Architecture for Binary Neural Networks." Proceedings of the IEEE International Conference on Computer Vision Workshops. 2019.

Future Potential of **Low-bit Networks**

If the gap between 32-bit CNNS and BNNs is closed:

- We can deploy dedicated hardware on servers and achieve huge energy savings
- Networks can run on mobile and embedded devices without a loss of accuracy

https://commons.wikimedia.org/wiki/File:Raspberry_Pi_3_B%2B_(39906369025).png

Prof. Dr. Ch. Meinel
Dean and Director
HPI, Germany

Our Research Insights on **Binary Neuronal Networks**

- The clipping threshold t_{clip} should be considered a hyperparameter and values between 1.2 and 1.3 leads to better results than the value of 1 that was used in most previous work [1]
- A scaling of channels after a binary convolution according to Rastegari et al. [2] can be absorbed by BatchNorm layers [3]
- A tighter approximation of the sign function does not necessarily achieve better results [4]

[1] Bethge, Joseph, Haojin Yang, and Christoph Meinel. "Training accurate binary neural networks from scratch." 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019.

[2] Rastegari, Mohammad, et al. "Xnor-net: Imagenet classification using binary convolutional neural networks." European conference on computer vision. Springer, Cham, 2016.

[3Joseph Bethge, Haojin Yang, Marvin Bornstein, Christoph Meinel. "BinaryDenseNet: Developing an Architecture for Binary Neural Networks." Proceedings of the IEEE International Conference on Computer Vision Workshops. 2019.

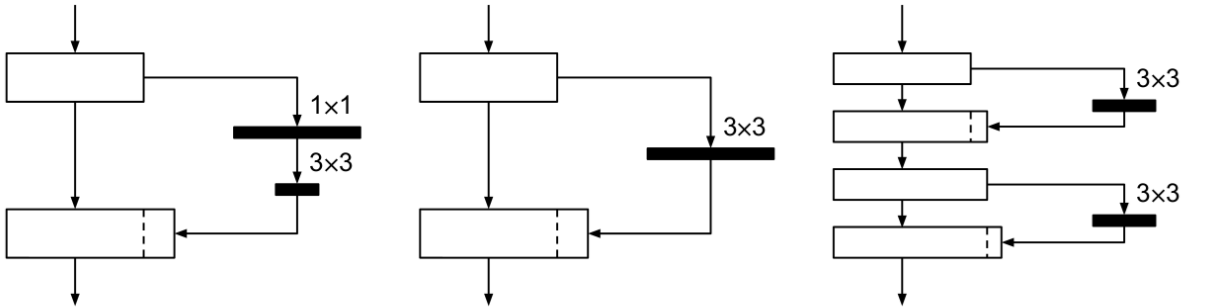
[4] Bethge, Joseph, Bethge, Joseph, Christian Bartz, Haojin Yang, Ying Chen, Christoph Meinel. "Training competitive binary neural networks from scratch." arXiv preprint arXiv:1812.01965 (2018).

Prof. Dr. Ch. Meinel
Dean and Director
HPI, Germany

Our BNN Models: **BinaryDenseNet**

BinaryDenseNet: a DenseNet adapted for Binary Networks

- Replace bottlenecks
- Add shortcuts



(a) DenseNet (bottleneck)

(b) DenseNet (no bottleneck)

(c) BinaryDenseNet (our suggestion)

Energy-efficient AI with BNN

Prof. Dr. Ch. Meinel
Dean and Director
HPI, Germany

[3] Joseph Bethge, Haojin Yang, Marvin Bornstein, Christoph Meinel. "BinaryDenseNet: Developing an Architecture for Binary Neural Networks." Proceedings of the IEEE International Conference on Computer Vision Workshops. 2019.

Our BNN Models: **BinaryDenseNet**

Model size	Method	Top-1/Top-5 accuracy
∼4.0MB	XNOR-ResNet18 [25]	51.2%/73.2%
	TBN-ResNet18 [30]	55.6%/74.2%
	Bi-Real-ResNet18 [24]	56.4%/79.5%
	BinaryResNetE18	58.1%/80.6%
	BinaryDenseNet28	**60.7%/82.4%**
∼5.1MB	TBN-ResNet34 [30]	58.2%/81.0%
	Bi-Real-ResNet34 [24]	62.2%/83.9%
	BinaryDenseNet37	62.5%/83.9%
	*BinaryDenseNet37-dilated**	**63.7%/84.7%**
7.4MB	*BinaryDenseNet45*	63.7%/84.8%
46.8MB	Full-precision ResNet18	69.3%/89.2%
249MB	Full-precision AlexNet	56.6%/80.2%

Energy-efficient AI with BNN

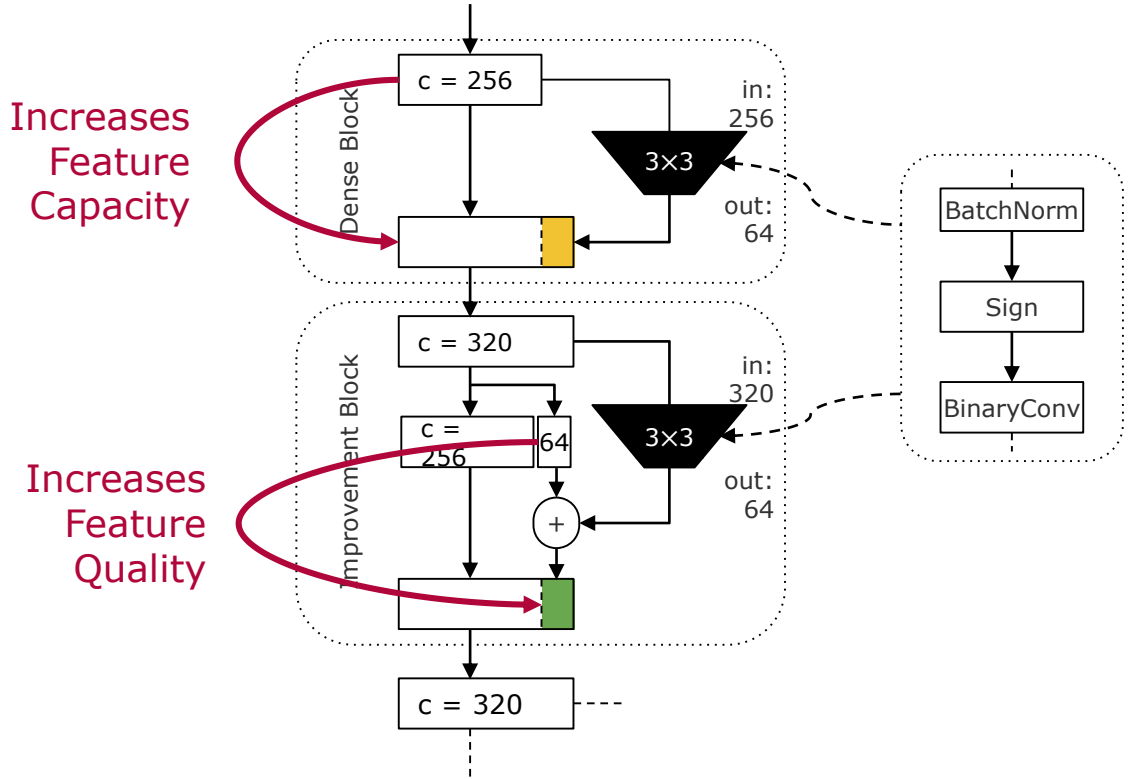
Prof. Dr. Ch. Meinel
Dean and Director
HPI, Germany

[3] Joseph Bethge, Haojin Yang, Marvin Bornstein, Christoph Meinel. "BinaryDenseNet: Developing an Architecture for Binary Neural Networks." Proceedings of the IEEE International Conference on Computer Vision Workshops. 2019.

Our BNN Models: **MeliusNet**

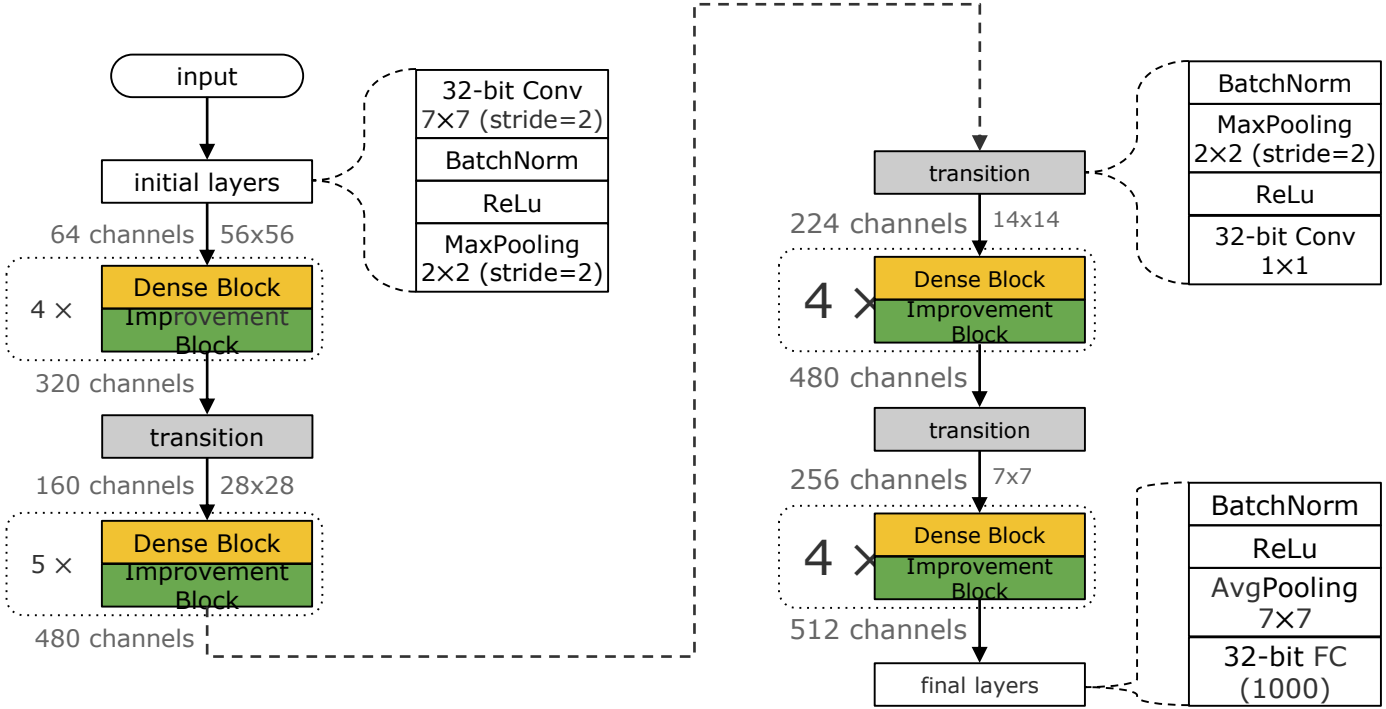
- Using 1 bit for weights and inputs leads to lower quality and capacity
- Number of possible values for **weights** is reduced from 2^{32} to 2
  - → leads to quantization error
  - → lower feature **quality**
- Value range of **inputs** is similarly reduced
  - → fine granular differences can no longer exist, only -1 and +1
  - → lower feature **capacity**
- Idea: solve both challenges through a specific architecture design

Our BNN Models: **MeliusNet**



Bethge, Joseph, Bethge, Joseph, Christian Bartz, Haojin Yang, Ying Chen, Christoph Meinel. "MeliusNet: Can Binary Neural Networks Achieve MobileNet-level Accuracy?." To appear in WACV 2021.

Our BNN Models: MeliusNet



Energy-efficient AI with BNN

Prof. Dr. Ch. Meinel
Dean and Director
HPI, Germany

Bethge, Joseph, Christian Bartz, Haojin Yang, Ying Chen, Christoph Meinel. "MeliusNet: Can Binary Neural Networks Achieve MobileNet-level Accuracy?." To appear in WACV 2021.

Our BNN Models: **MeliusNet**

Method	Bitwidth (W/A)	ImageNet (≈18 layers)			ImageNet (≈34 layers)		
		Top-1 Acc.	Model Size	OPs ($\cdot 10^8$)	Top-1 Acc.	Model Size	OPs ($\cdot 10^8$)
BWN[35]	1/32	60.8	4MB	18.1	-	-	-
TTQ[47]	2/32	**66.6**	5.3MB	18.1	-	-	-
HWGQ[7]	1/2	59.6	4MB	∼2.4	64.3	5.1MB	∼3.4
LQ-Net[44]	1/2	62.6	4MB	∼2.4	**66.6**	5.1MB	∼3.4
SYQ[11]	1/2	55.4	4MB	∼2.4	-	-	-
DoReFa[46]	2/2	62.6	5.3MB	∼2.4	-	-	-
Ensemble[48]	(1/1)×6	61.0	-	-	-	-	-
Circulant-CNN[29]	(1/1)×4	61.4	-	-	-	-	-
ABC-Net[28]	(1/1)×5	65.0	8.7MB	7.8	-	-	-
GroupNet[28]	(1/1)×5	**67.0**	9.2MB	2.68	**70.5**	15.3MB	4.13
BNN[22]	1/1	42.2		1.57	-		-
XNOR-Net[35]	1/1	51.2		1.59	-		-
Bi-RealNet[31]	1/1	56.4		1.63	62.2		1.93
XNOR-Net++[6]	1/1	57.1		1.59	-		-
Bi-RealNet (our baseline)	1/1	60.6	∼4MB	1.14	63.7	∼5.1MB	1.43
BinaryDenseNet[5]	1/1	60.7		2.58	62.5		2.71
Strong Baseline[32]	1/1	60.9		1.82	-		-
BinaryDenseNet (our baseline)	1/1	62.6		2.09	64.2		2.20
MeliusNetA,B (ours)	1/1	**63.4**		1.62	**65.7**		1.96
32-bit baseline (ResNet)	32/32	69.3	46.8MB	18.1	73.3	87.2MB	36.6

Bethge, Joseph, Bethge, Joseph, Christian Bartz, Haojin Yang, Ying Chen, Christoph Meinel. "MeliusNet: Can Binary Neural Networks Achieve MobileNet-level Accuracy?." To appear in WACV 2021.

Energy-efficient AI with BNN

Prof. Dr. Ch. Meinel
Dean and Director
HPI, Germany

BMXNet 2
Open Source Framework for Binary Neuronal Networks

- BMXNet 2 is based on mxnet [1]
- Contains reproducible models and demos to provide a strong basis for research and industry
- Can be used to find new network architectures, test new ideas, ...

https://github.com/hpi-xnor/BMXNet-v2

Energy-efficient AI with BNN

Prof. Dr. Ch. Meinel
Dean and Director
HPI, Germany

[1] https://mxnet.incubator.apache.org/

BMXNet 2
Showcases and Demo Applications

- Android demo app:
  https://github.com/hpi-xnor/android-image-classification
  - ImageNet classification based on a ResNetE
- Human Pose detection demo for a Raspberry Pi

https://github.com/hpi-xnor/BMXNet-v2

Prof. Dr. Ch. Meinel
Dean and Director
HPI, Germany

Thank You for Your Interest!

Christoph Meinel
Hasso Plattner Institute
Campus Griebnitzsee, Potsdam
www.hpi.de