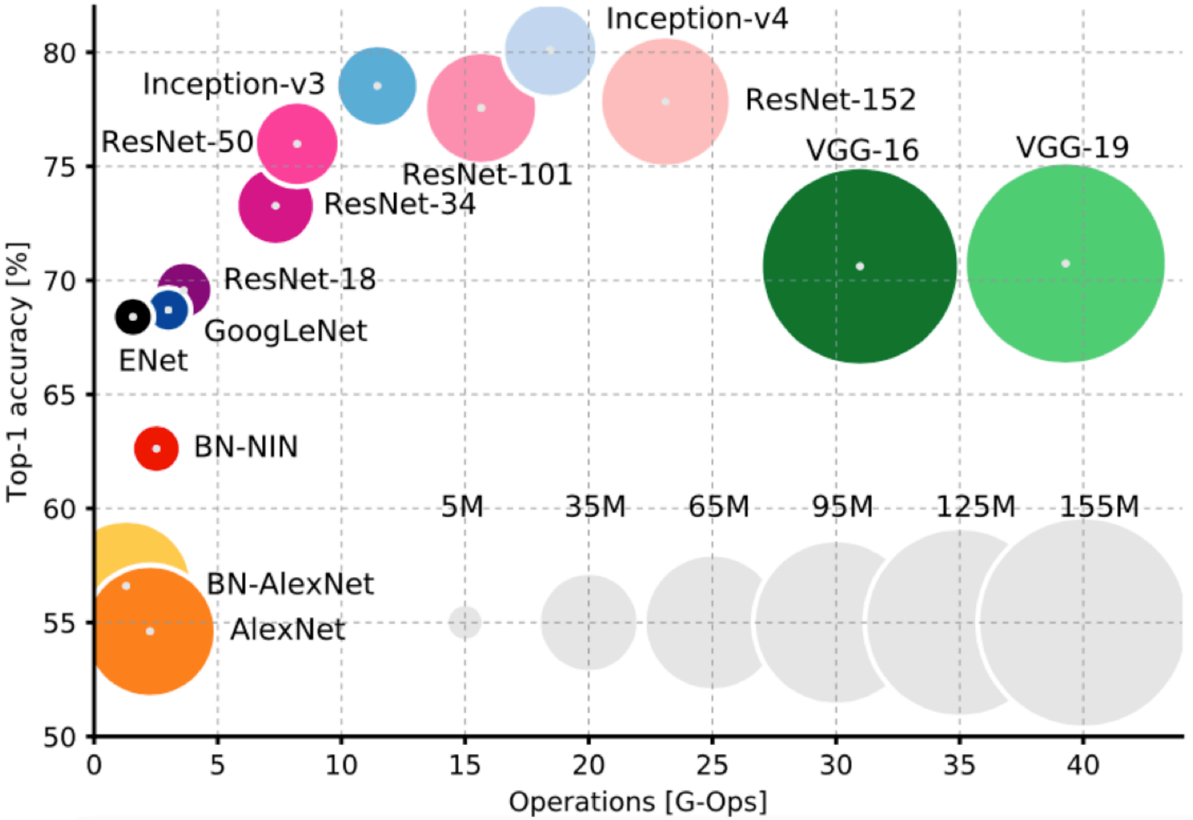
DeepCABAC: Plug&Play Compression of Neural Network Weights and Weight Updates

David Neumann, Felix Sattler, Heiner Kirchhoffer, Simon Wiedemann, Karsten Müller,
Heiko Schwarz, Thomas Wiegand, Detlev Marpe, Wojciech Samek

David Neumann
david.neumann@hhi.fraunhofer.de

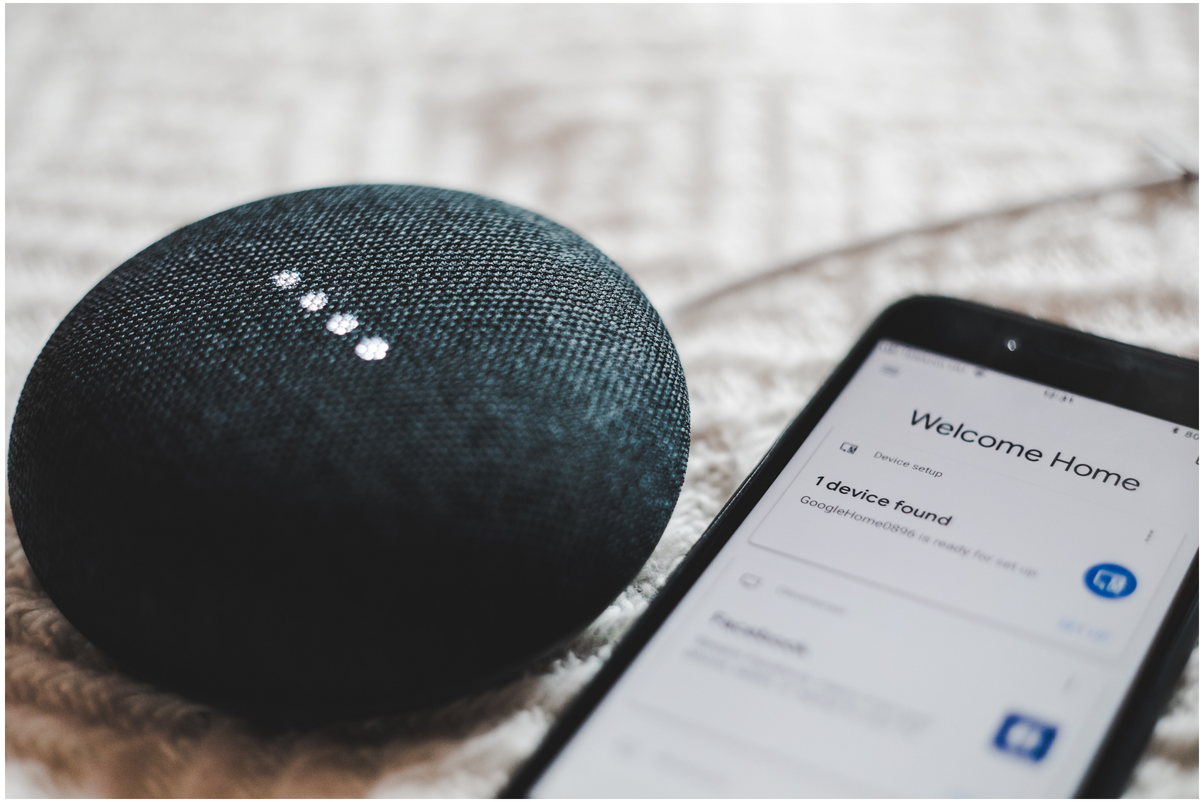
Introduction



Deep learning models contain up to multiple billions of parameters [1, 2]

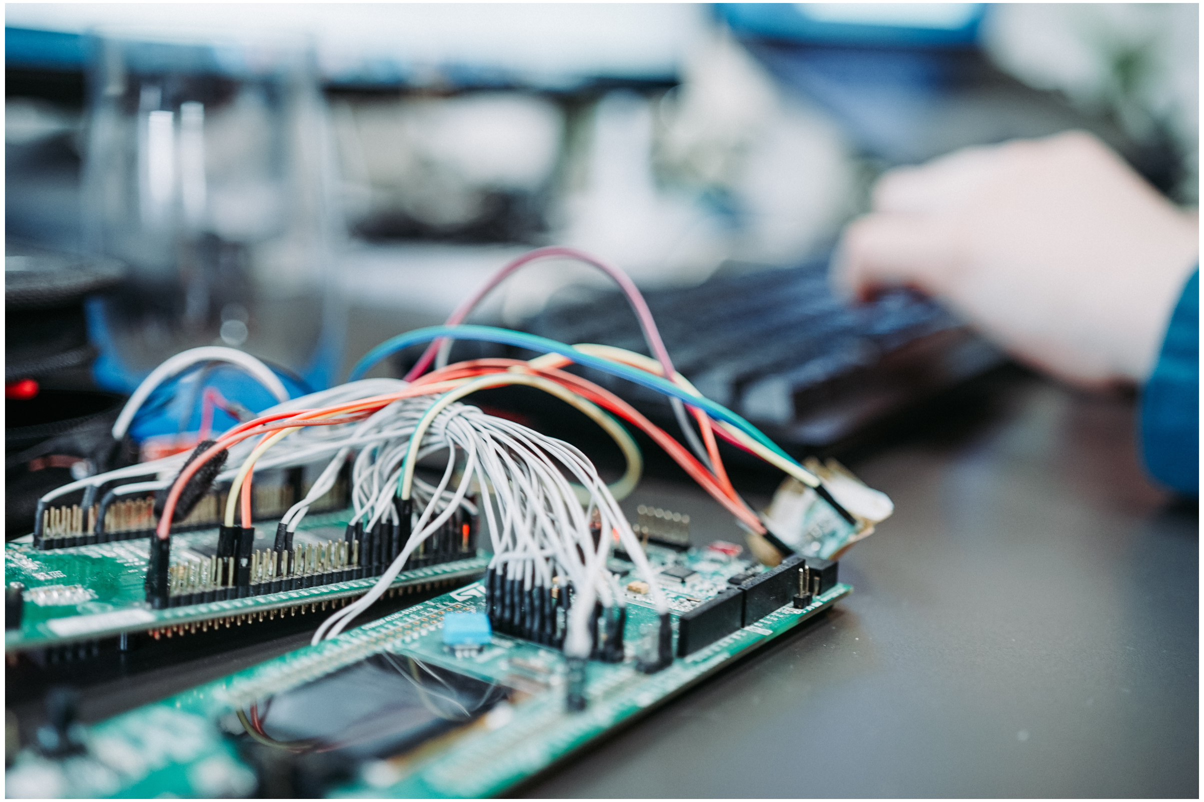
David Neumann
david.neumann@hhi.fraunhofer.de

Introduction



Mobile phones and IoT devices

Photo by BENCE BOROS on Unsplash (https://unsplash.com/photos/anapPhJFRhM)



Embedded Devices

Photo by Zan on Unsplash (https://unsplash.com/photos/wGqz5YSqsfk)

David Neumann
david.neumann@hhi.fraunhofer.de

Introduction

Mobile Connections & 5G

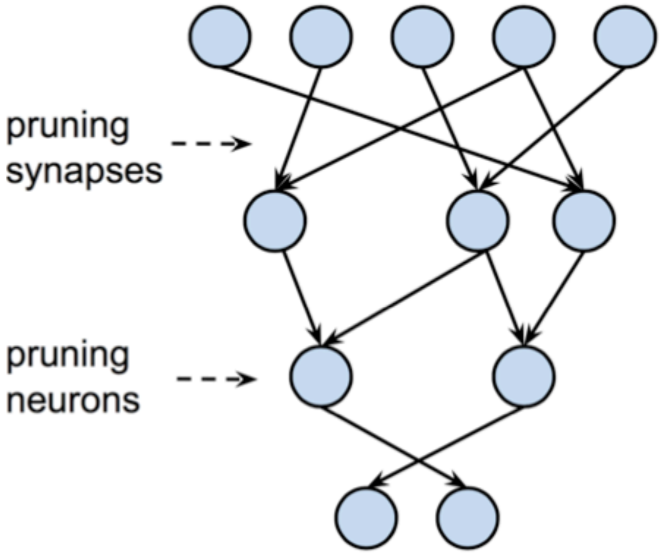
Photo by Mika Baumeister on Unsplash (https://unsplash.com/photos/gwWkv06WYFY)

Bandwidth-constrained communication channels

Photo by Jordan Harrison on Unsplash (https://unsplash.com/photos/40XgDxBfYXM)

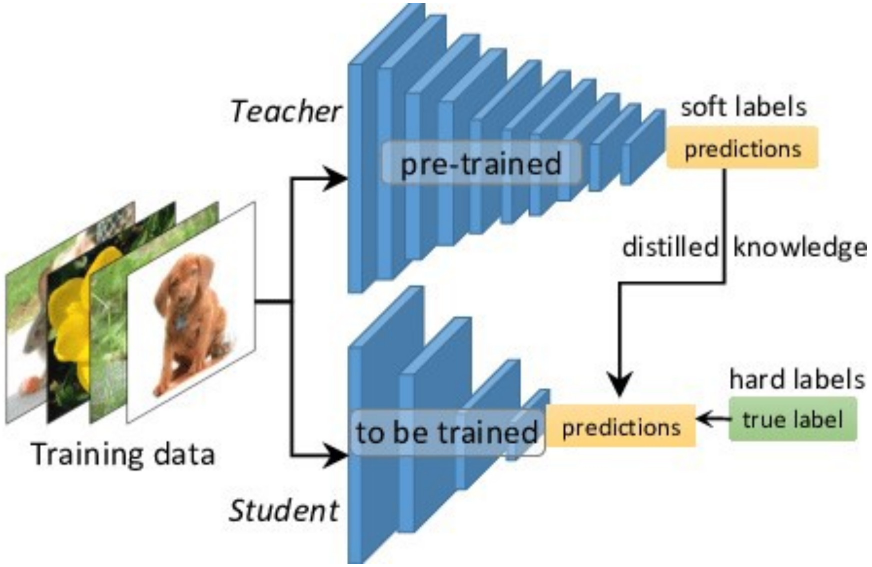
David Neumann
david.neumann@hhi.fraunhofer.de

Introduction



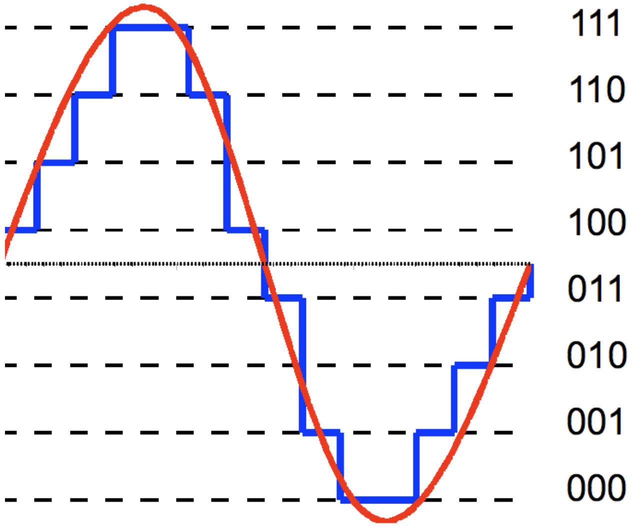
Pruning

Han, Song, et al. "Learning both weights and connections for efficient neural network." NIPS. 2015



Distillation

Image by Prakhar Ganesh (https://towardsdatascience.com/knowledge-distillation-simplified-dd4973dbc764)



Trained Quantization

Image by Hyacinth on Wikimedia Commons (https://commons.wikimedia.org/wiki/File:3-bit_resolution_analog_comparison.png)

David Neumann
david.neumann@hhi.fraunhofer.de

Introduction

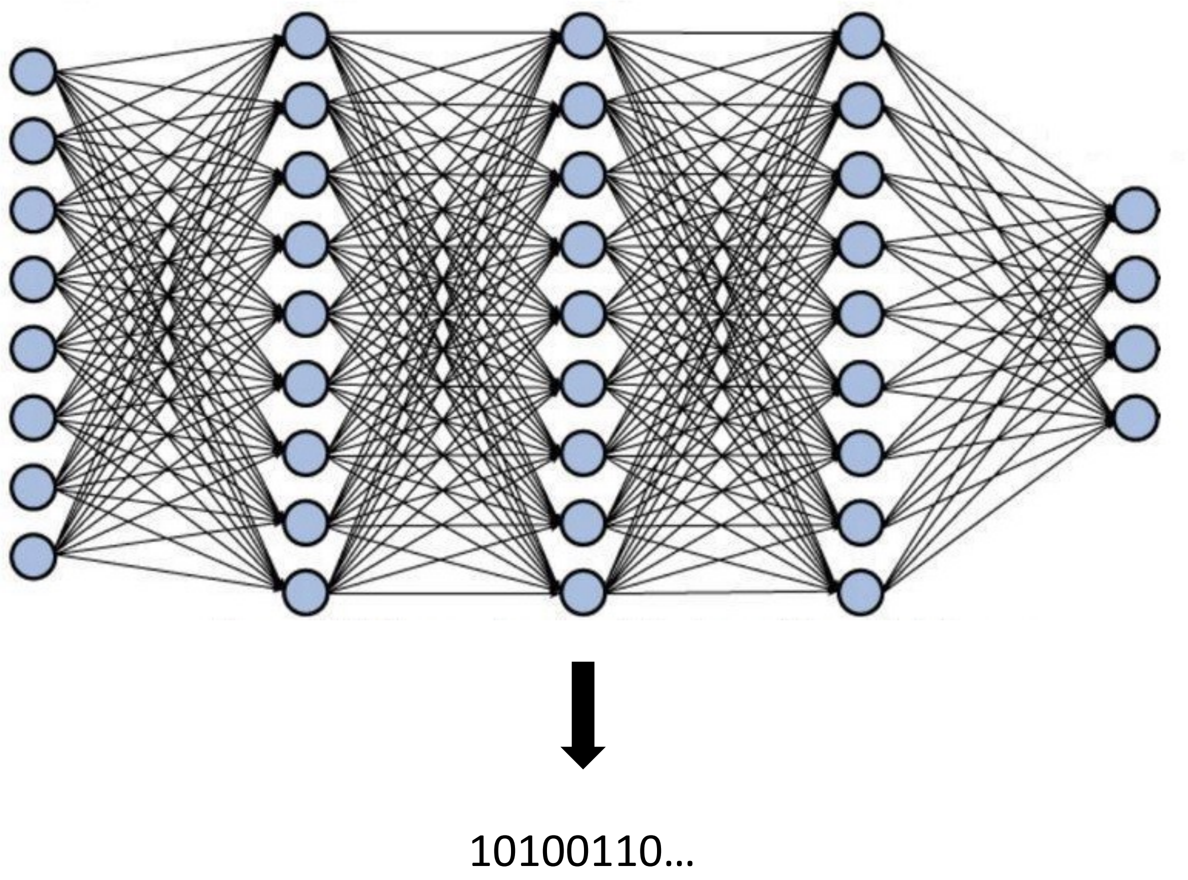
- Proposed solutions:
  - Highly optimized for one application area and/or
  - Require expensive re-training of the neural network

- Desired solution:
  - General purpose
  - Easy-to-use
  - Fast and efficient
  - High compression gains
  - Must not harm the performance of the neural network

David Neumann
david.neumann@hhi.fraunhofer.de

DeepCABAC

- DeepCABAC [12] is a general-purpose neural network compression algorithm
- Was adopted to the current working draft of the MPEG-7 part 17 standardization efforts
- Is based on **c**ontext-based **a**daptive **b**inary **a**rithmetic **c**oding (CABAC) [13], widely used in video coding standards



David Neumann
david.neumann@hhi.fraunhofer.de

Fraunhofer HHI

DeepCABAC

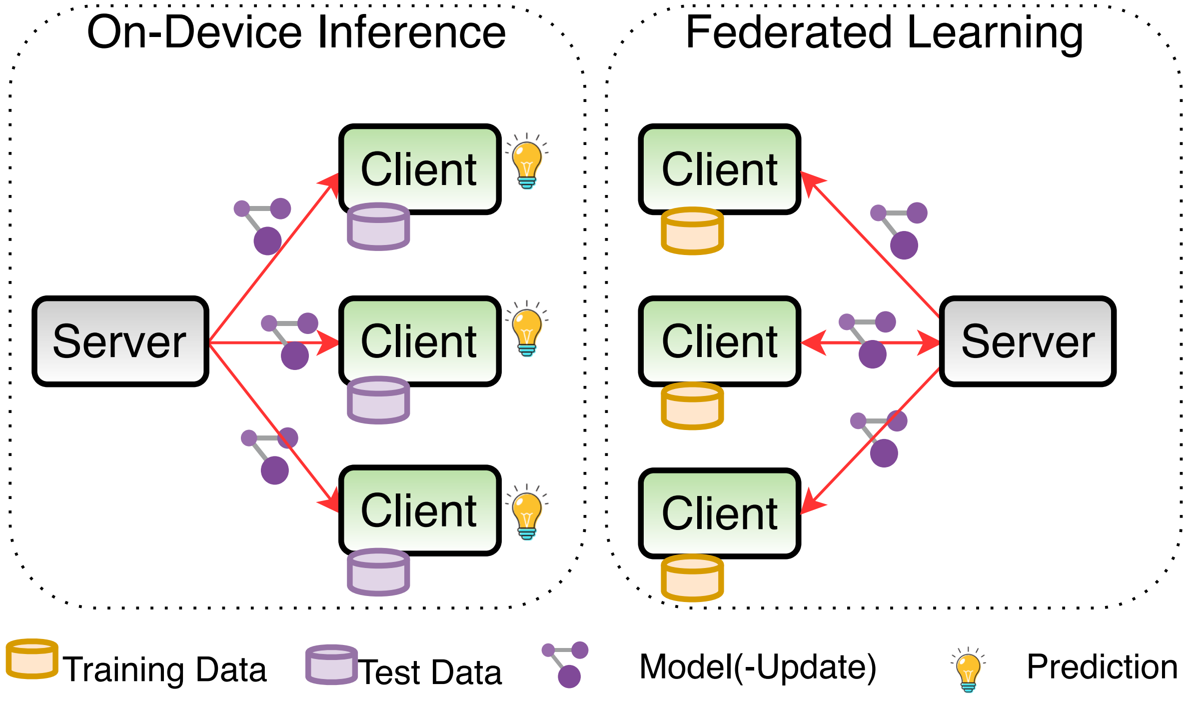
- Why is DeepCABAC ideal as a universal compression method?
    1) CABAC was designed for lossless compression of integers, i.e. can be combined with any quantization scheme
    2) Achieves high compression gains
    3) Adaptive towards any kind of tensor-shaped data
    4) Fast and efficient and does not need the model to undergo expensive re-training
    5) Can be used in a plug & play fashion, i.e. it can be easily integrated into existing deep learning pipelines, e.g. federated learning

David Neumann
david.neumann@hhi.fraunhofer.de

Compression Scenarios in Federated Learning

1) A fully-trained model needs to be communicated, e.g. when a model was trained on central server and needs to be deployed on-device

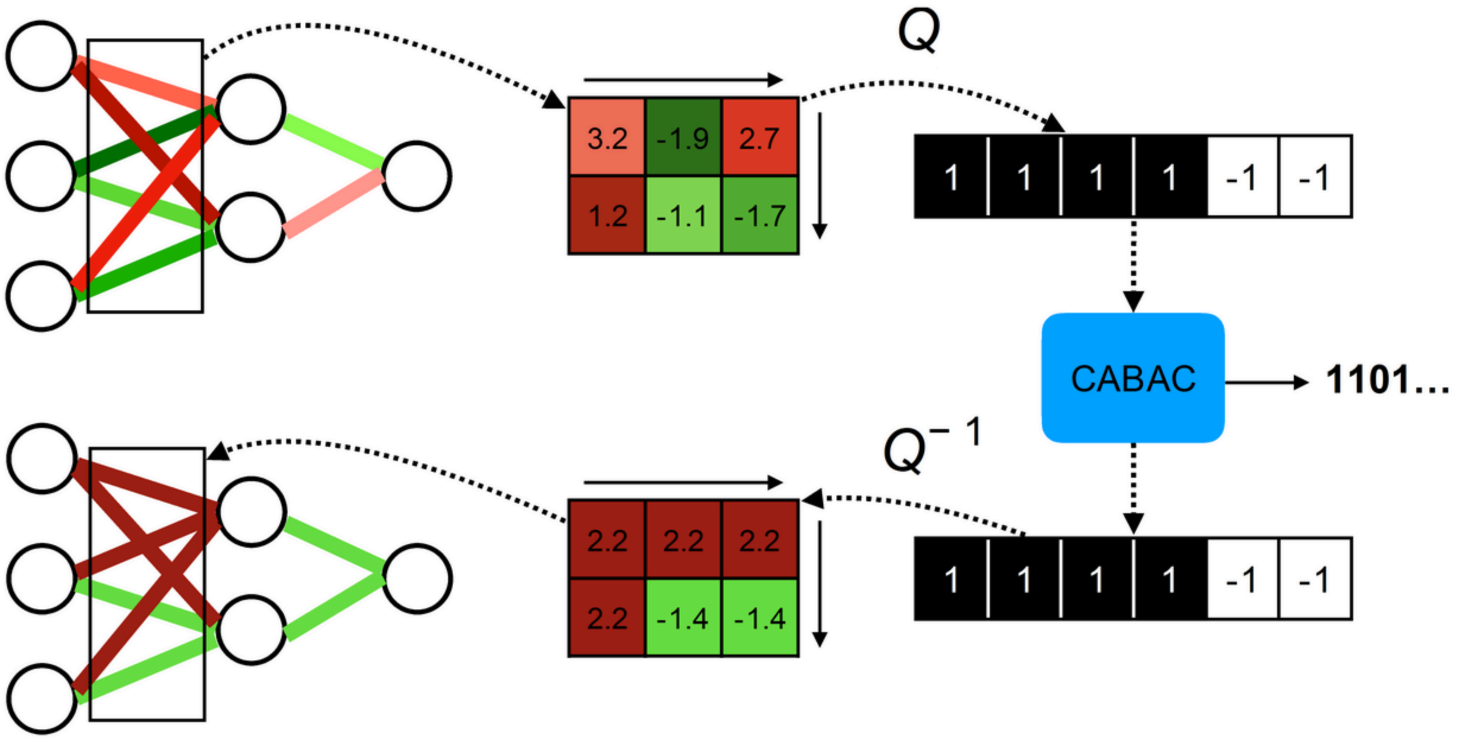
2) The recipient already possesses an out-of-date version and only the element-wise difference needs to be communicated



David Neumann
david.neumann@hhi.fraunhofer.de

Fraunhofer HHI

How Does DeepCABAC Work?



Compression and Decompression of a neural network

David Neumann
david.neumann@hhi.fraunhofer.de

How Does DeepCABAC Work?

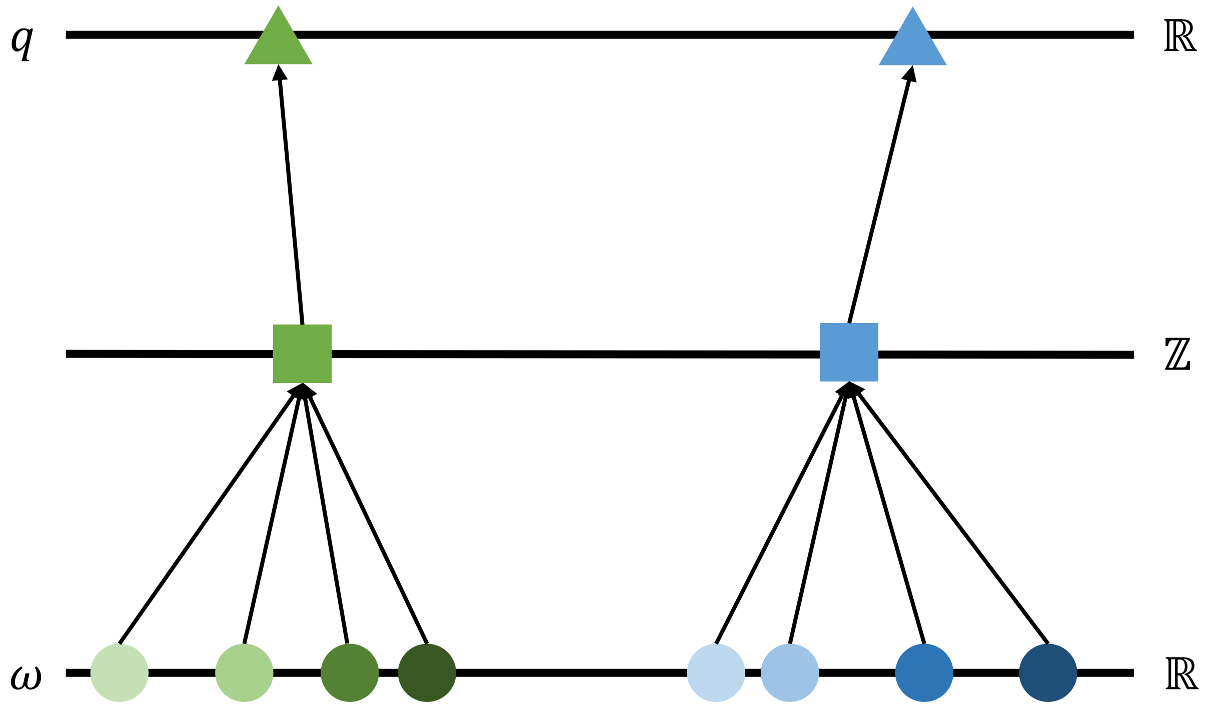
Quantization → Binarization → Context Modelling → Entropy Coding

$\underbrace{\text{Quantization}}_{\textbf{Lossy}}$ $\underbrace{\text{Binarization → Context Modelling → Entropy Coding}}_{\textbf{Lossless}}$

David Neumann
david.neumann@hhi.fraunhofer.de

How Does DeepCABAC Work?

- Quantization
  - Continuous input data need to be converted to discrete inputs
  - CABAC does not prescribe any specific way of quantization

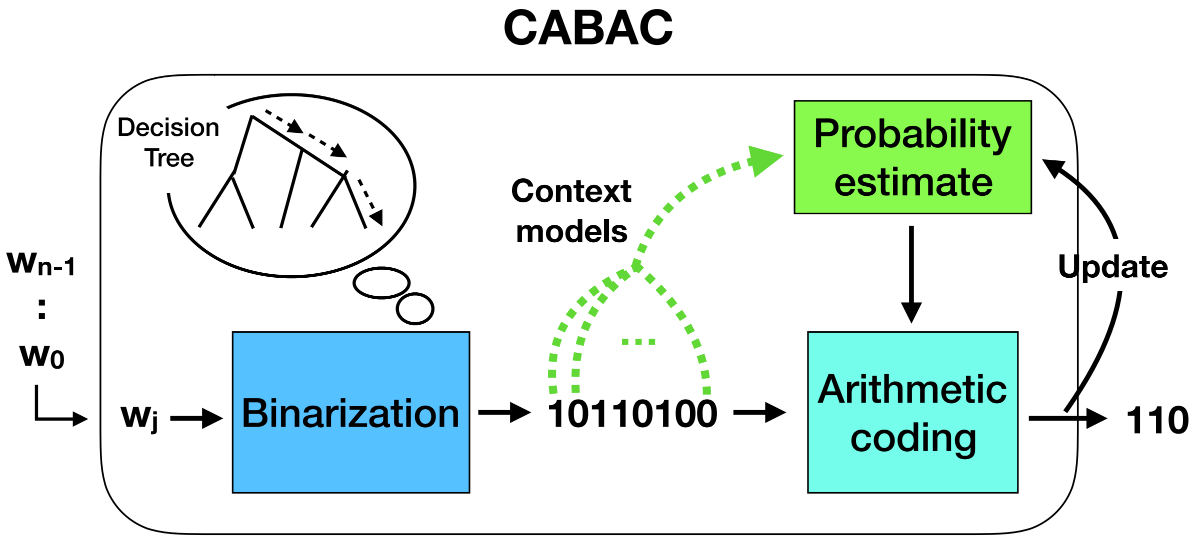


David Neumann
david.neumann@hhi.fraunhofer.de

Fraunhofer HHI

How Does DeepCABAC Work?

- Binarization
  - CABAC expects discrete inputs (integers)
  - Represents each unique input value as a sequence of binary decisions

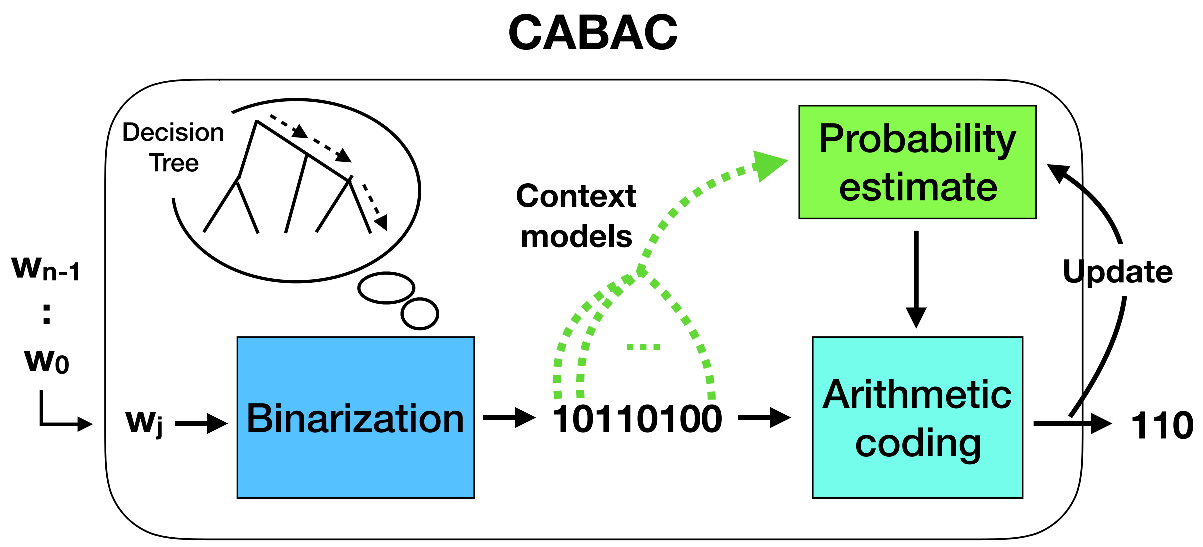


David Neumann
david.neumann@hhi.fraunhofer.de

Fraunhofer HHI

How Does DeepCABAC Work?

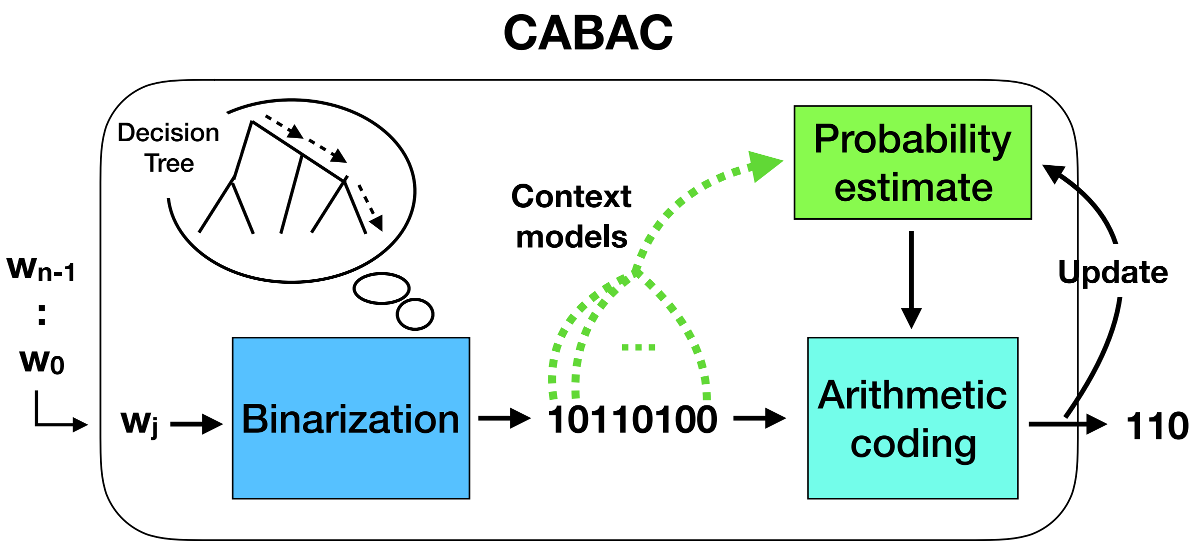
- Context Modelling
  - For each of the decisions in the binarization, a probability model is used
  - This context model is updated on-the-fly based on how the current input data is distributed
  - Local distribution estimation means no prior for the data distribution is needed



David Neumann
david.neumann@hhi.fraunhofer.de

How Does DeepCABAC Work?

- Entropy Coding
  - Encodes the bit sequence with minimal redundancy
  - Arithmetic coding is extremely efficient



David Neumann
david.neumann@hhi.fraunhofer.de

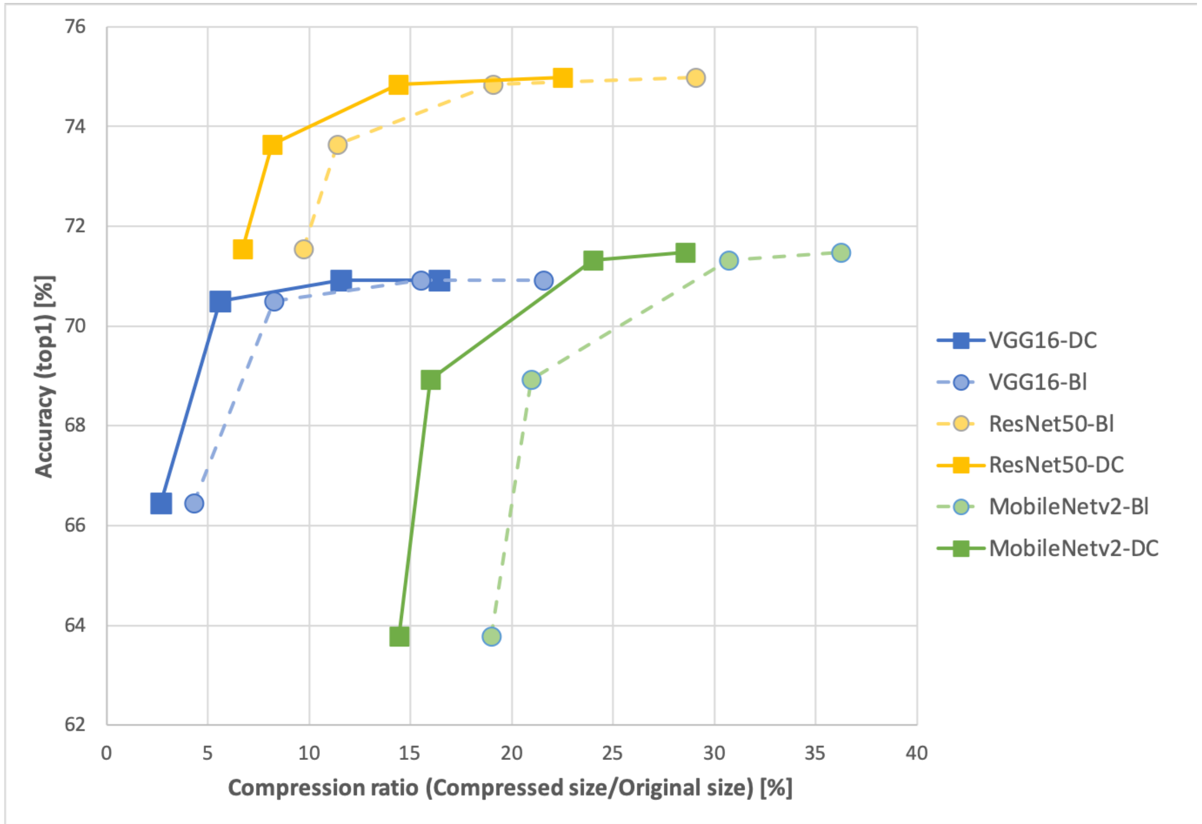
Experiments

- We performed two kinds of experiments:
  1) Compression of pre-trained neural networks to gauge general compression gains achievable using DeepCABAC
  2) A federated learning use-case where different neural network architectures where trained on CIFAR-10 using 10 clients

David Neumann
david.neumann@hhi.fraunhofer.de

Experiments

- Full-network compression
  - „DC“ denotes networks compressed with DeepCABAC
  - „Bl“ denotes a baseline compression algorithm (bZip)
  - DeepCABAC consistently attains better rate-distortion curves



David Neumann
david.neumann@hhi.fraunhofer.de

Fraunhofer HHI

Experiments

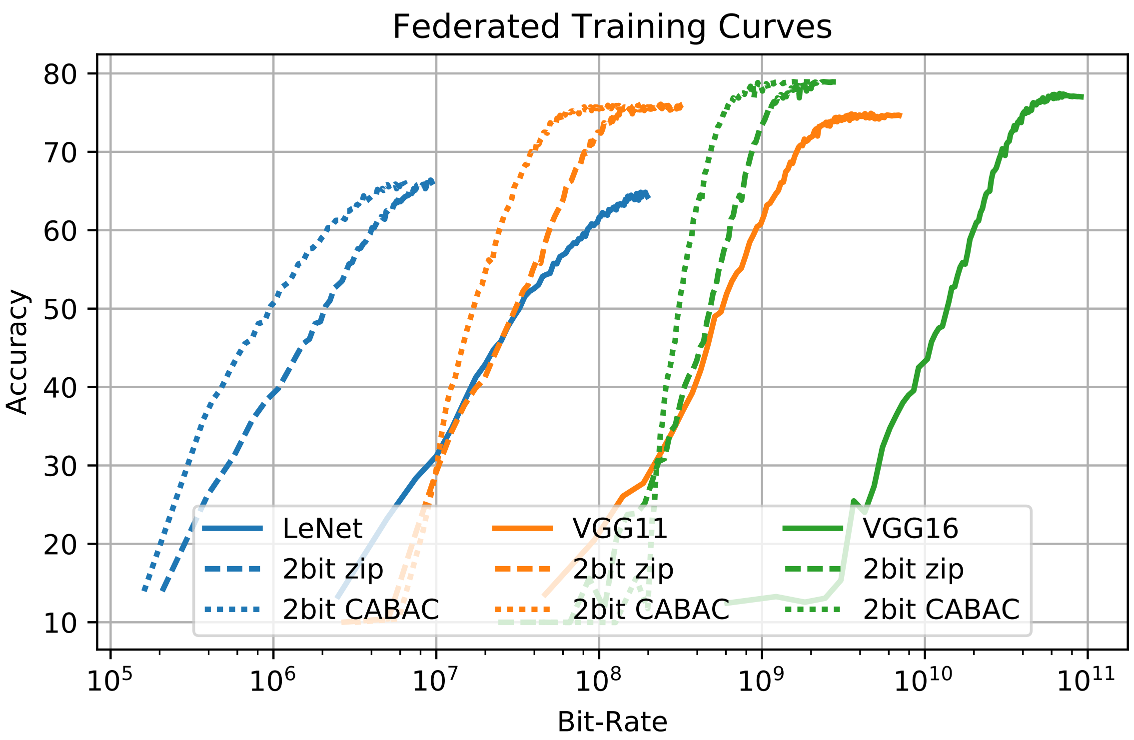
- Compression ratios achieved at no loss of accuracy when applying DeepCABAC to a wide set of NN architectures trained on different tasks

Models	Original Size [MB]	Original Accuracy (top 1 [%])	bZip (CR [%])	DeepCABAC (CR [%])	Accuracy (top 1 [%])
VGG16	553.43	70.93	15.52	**11.98**	70.92
ResNet50	102.23	74.98	29.09	**22.52**	74.99
MobileNet-v2	14.15	71.47	36.24	**28.57**	71.48
Audio-Net	467.27	58.27	15.15	**10.93**	59.51
FCAE	304.72	30.13 PSNR	39.28	**30.63**	30.17 PSNR

David Neumann
david.neumann@hhi.fraunhofer.de

Experiments

- Federated Learning
  - Convergence speed with respect to communicated bits
  - Solid lines denote no compression
  - Dashed lines denote a baseline compression algorithm (bZip)
  - Dotted lines denote DeepCABAC compression
  - For both compression methods, the weights were quantized to 2 bits



David Neumann
david.neumann@hhi.fraunhofer.de

Experiments

- Federated learning with 10 clients on the CIFAR-10 dataset
- Compression results for 2-bit nearest neighbor quantization encoded

Models	Total Communication	Original Accuracy (top 1 [%])	bZip (CR [%])	DeepCABAC (CR [%])	Accuracy (top 1 [%])
LeNet	553.43 MB	64.84	4.84	**3.29**	66.39
VGG11	6.98 GB	74.91	4.90	**2.76**	76.15
VGG16	90.86 GB	77.44	3.36	**2.30**	78.98

David Neumann
david.neumann@hhi.fraunhofer.de

Conclusion

- Several specialized solutions have been proposed for different use-cases
- There is a need for general and easy-to-use compression methods
- We addressed this issue and presented DeepCABAC, a universal compression tool, which achieves competitive compression rates with no or minimal loss of accuracy
- We demonstrated that DeepCABAC can easily be integrated with distributed training pipelines

David Neumann
david.neumann@hhi.fraunhofer.de

Where to go from here?

- Another talk on DeepCABAC
  - https://slideslive.com/38917367/deepcabac-contextadaptive-binary-arithmetic-coding-for-deep-neural-network-compression

- Want to learn more about Neural Network Compression and Federated Learning?
  - http://efficient-ml.org

- DeepCABAC on GitHub
  - https://github.com/fraunhoferhhi/DeepCABAC

David Neumann
david.neumann@hhi.fraunhofer.de

References

- **[1]** Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, „Language models are unsupervised multitask learners,“ 2019.

- **[2]** Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro, “Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism,“ *arXiv e-prints*, p. arXiv:1909.08053, Sep 2019.

- **[3]** Anusha Lalitha, Osman Cihan Kilinc, Tara Javidi, and Farinaz Koushanfar, „Peer-to-peer Federated Learning on Graphs,“ *arXiv e-prints*, p. arXiv:1901.11173, Jan 2019.

David Neumann
david.neumann@hhi.fraunhofer.de

References

- **[4]** Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al., „Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.

- **[5]** S. Caldas, J. Konečny, HB. McMahan, and A. Talwalkar, „Expanding the Reach of Federated Learning by Reducing Client Resource Requirements," *arXiv e-prints*, p. arXiv:1812.07210, Dec 2018.

- **[6]** Yann LeCun, John S. Denker, and Sara A. Solla, „Optimal brain damage," in *Advances in Neural Information Processing Systems 2*, D. S. Touretzky, Ed., pp. 598– 605. Morgan-Kaufmann, 1990.

David Neumann
david.neumann@hhi.fraunhofer.de

References

- **[7]** Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean, „Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

- **[8]** Song Han, Huizi Mao, and William J. Dally, „Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding," *arXiv e-prints*, p. arXiv:1510.00149, Oct 2015.

- **[9]** J. Konečný, HB. McMahan, F. X. Yu, P. Richtárik, A. Theertha Suresh, and D. Bacon, „Federated Learning: Strategies for Improving Communication Efficiency," *arXiv e-prints*, p. arXiv:1610.05492, Oct 2016.

David Neumann
david.neumann@hhi.fraunhofer.de

References

- **[10]** F. Sattler, S. Wiedemann, K. Müller, and W. Samek, „Sparse binary compression: Towards distributed deep learning with minimal communication," in *2019 International Joint Conference on Neural Networks (IJCNN)*, July 2019, pp. 1–8.

- **[11]** F. Sattler, S. Wiedemann, K. Müller, and W. Samek, „Robust and communication-efficient federated learning from non-i.i.d. data," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2019.

- **[12]** S. Wiedemann, H. Kirchhoffer, S. Matlage, P. Haase, A. Marban, T. Marinc, D. Neumann, T. Nguyen, H. Schwarz, T. Wiegand, D. Marpe, and W. Samek, "Deepcabac: A universal compression algorithm for deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–1, 2020.

David Neumann
david.neumann@hhi.fraunhofer.de

References

- **[13]** D. Marpe, H. Schwarz, and T. Wiegand, „Context-based adaptive binary arithmetic coding in the h.264/avc video compression standard,“ *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 13, no. 7, pp. 620–636, July 2003.

- **[14]** HB. McMahan, E. Moore, D. Ramage, Seth Hampson, and B. Agüera y Arcas, „Communication-Efficient Learning of Deep Networks from Decentralized Data,“ *arXiv e-prints*, p. arXiv:1602.05629, Feb 2016.

- **[15]** Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, „Imagenet: A large-scale hierarchical image database,“ in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

David Neumann
david.neumann@hhi.fraunhofer.de

References

- **[16]** Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, „Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

- **[17]** Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut, „Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

- **[18]** G. Hinton, O. Vinyals, and J. Dean, „Distilling the Knowledge in a Neural Network," *arXiv preprint arXiv:1503.02531*, 2015.

David Neumann
david.neumann@hhi.fraunhofer.de