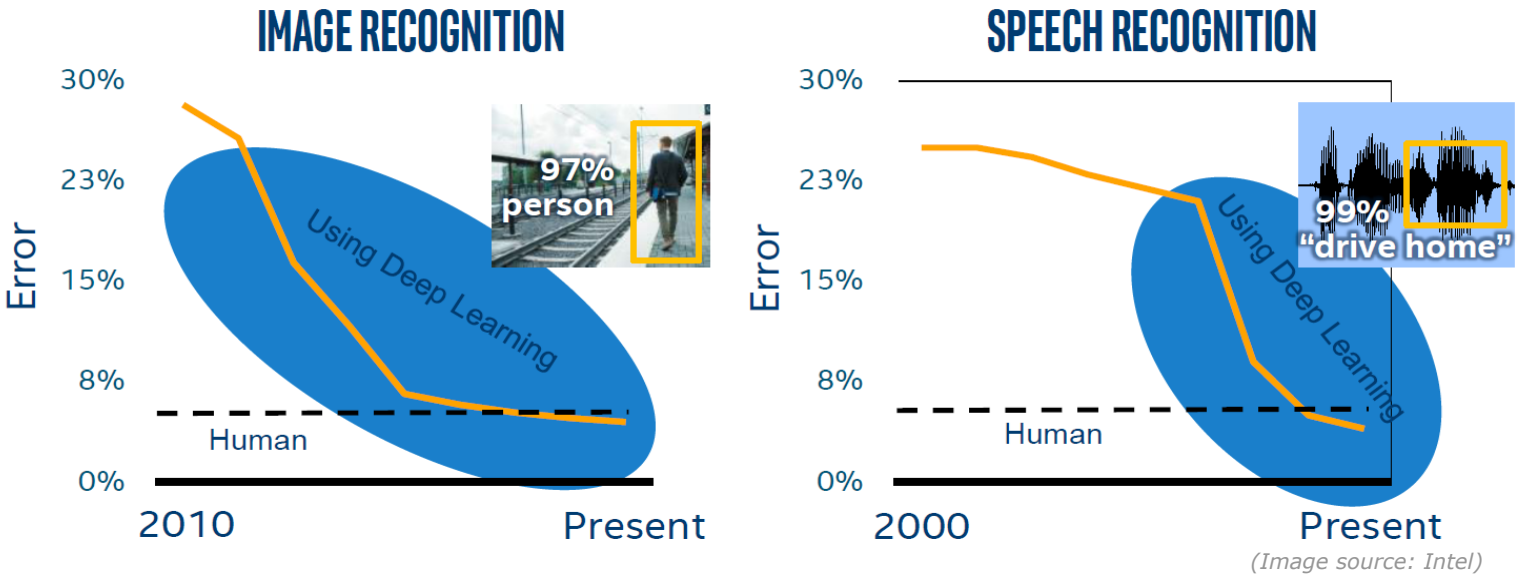
Deep Model Compression – Compact Network Design

Dr. habil. Haojin Yang
Senior Researcher
Hasso Plattner Institute, Germany

The current AI technologies

Deep learning breakthroughs since 2006

- AI achieved human performance in many perceptual recognition tasks



(Image source: Intel)

Dr. Haojin Yang
HPI, Germany

AI applications

Chatbots in Social Media and Retail

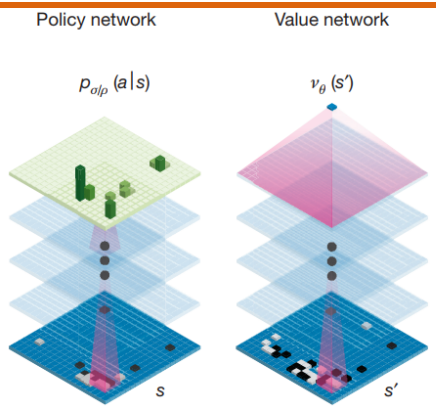
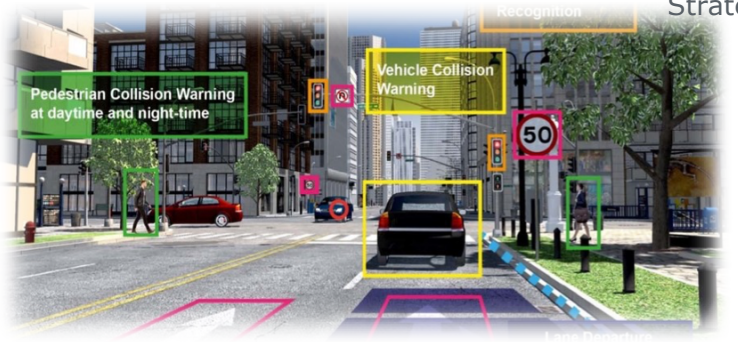
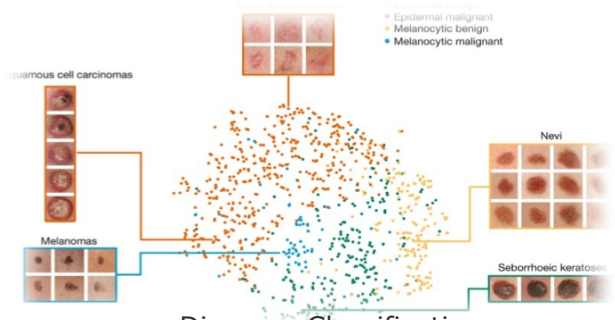


Image from Silver et al., 2016.
Strategy game

Image by Emily MacKenzie, 2015
machine translation



Autonomous Driving



Diseases Classification

The promising and challenging future of AI

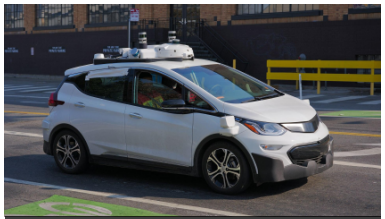
AI competition has entered the second half…

- the AI hype in media *vs.* the difficulty of AI technology landing
- continue to make breakthroughs and fulfill its promises, e.g., in autonomous driving
- the impact of large-scale AI computing on the environment
- **especially important to improve the efficiency of AI models**
- **reducing or even decoupling its strong dependence on high-performance computing hardware**

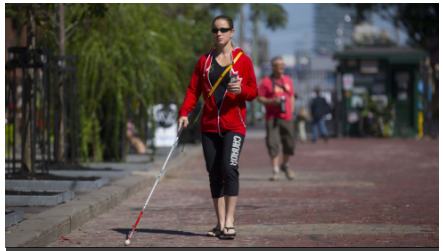
Dr. Haojin Yang
HPI, Germany

Deploy AI models on client and edge devices

- VGG-Net has 16/19 layers, **24M** nodes, **14M** parameters and, **15B** connections
  - model size **550MB**
  - memory: 24M ∗ 4 bytes ≈ **96MB / image** (inference only)

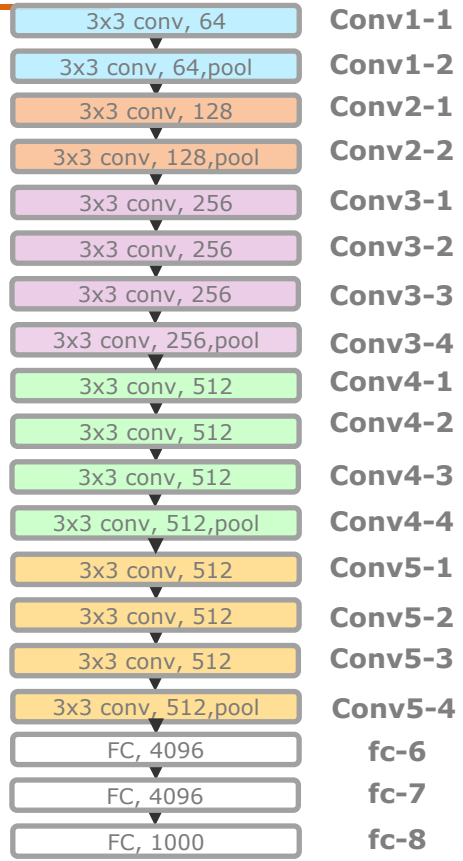


Autonomous driving



Assistance apps

Low power devices



Simonyan et al. VGG-Net, ICLR'15

Squeezing deep models

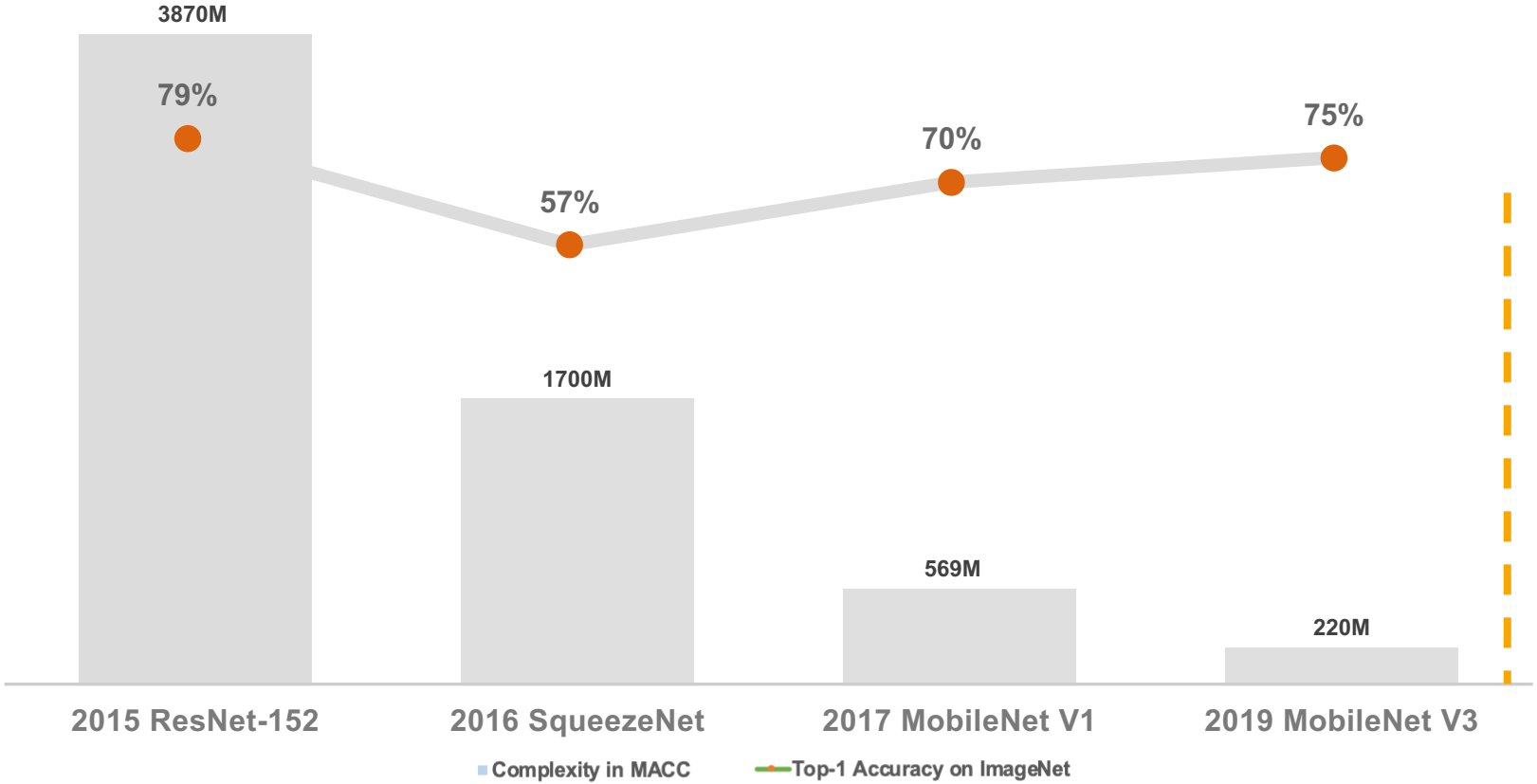
Model Compression

- **compact network design**
- knowledge distillation
- quantization and pruning technique
- binary neural networks

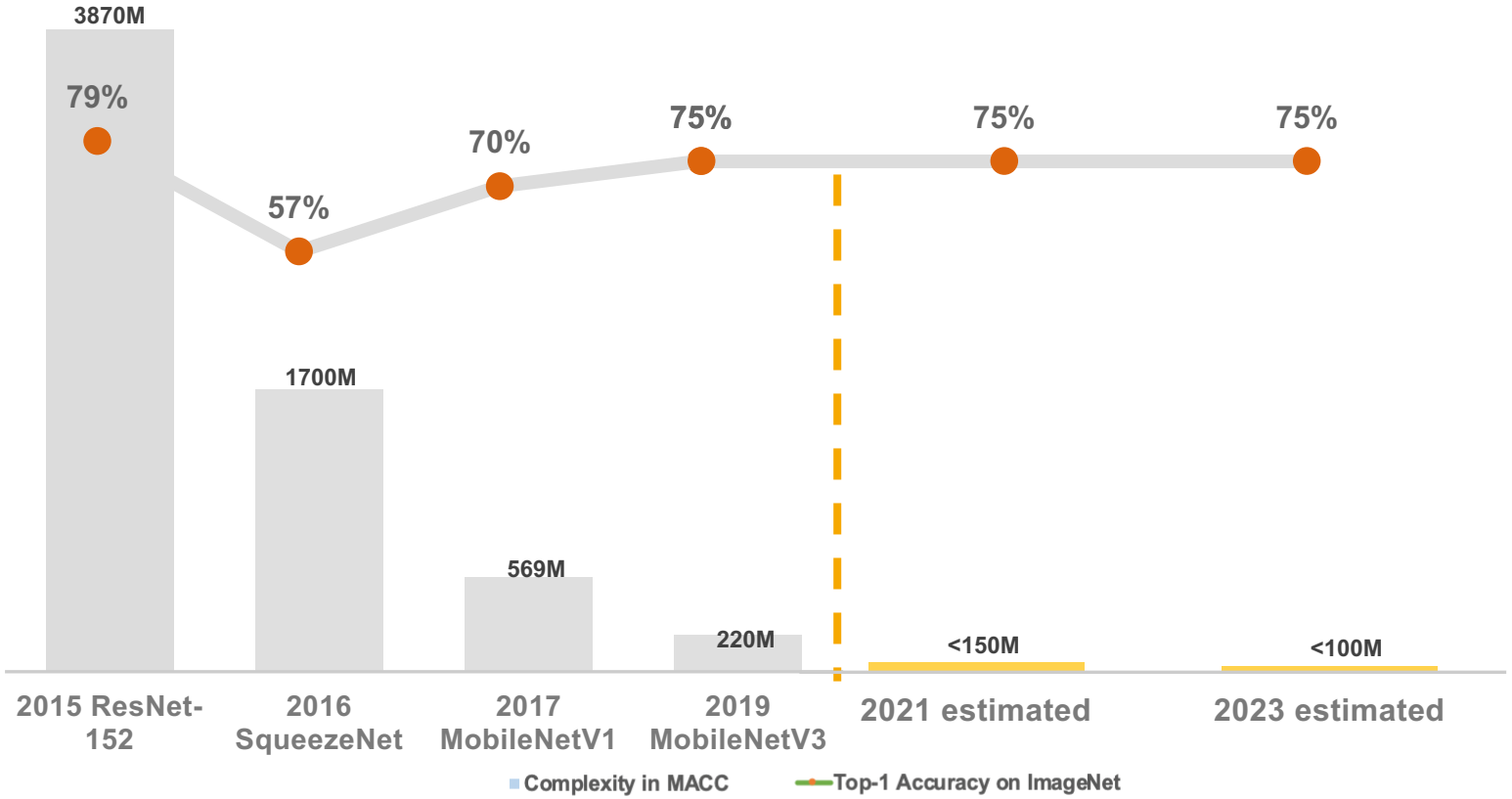
Dr. Haojin Yang
HPI, Germany

Evolution of image classification models

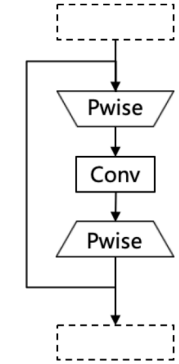
Evolution of image classification models



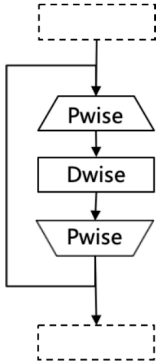
Evolution of image classification models



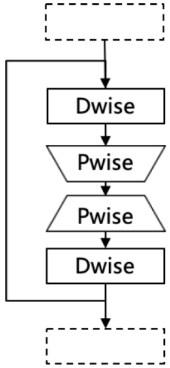
SOTA compact networks



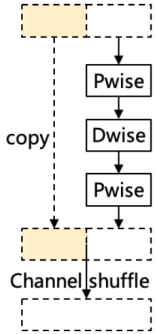
(a) Bottleneck block [17]



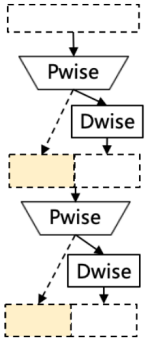
(b) Inverted residual block [38]



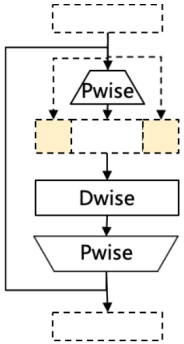
(c) Sandglass block [49]



(d) ShuffleBlock v2 [31]

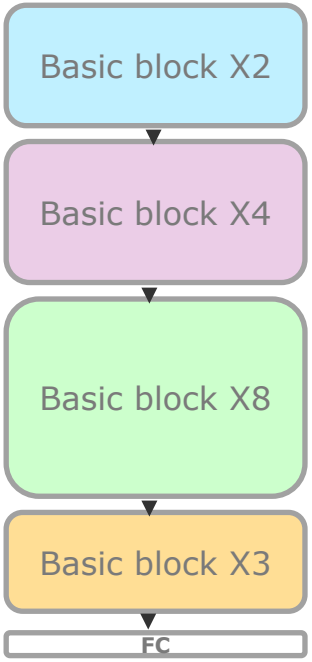


(e) Ghost module [14]



(f) Asymmetrical bottleneck

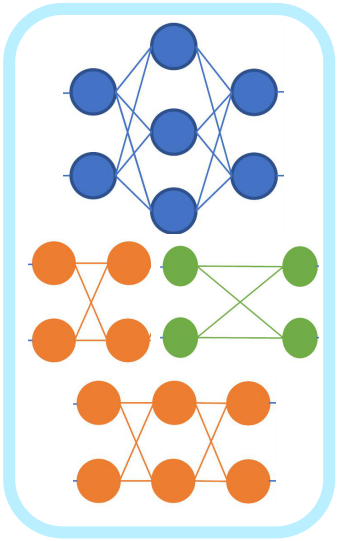
Basic blocks



Network

Compact network design powered by NAS

Neural Architecture Search



Search Space

network architecture sampling

Random

Bayesian

Grid

RL etc.

Search Strategy

picked networks

return performance estimate

Model training

Evaluation

Performance Estimation Strategy

Outlook

- SOTA compact networks: *MobileNetV3*, *GhostNet* based on handcrafted block design and NAS
- hardware aware search
  - E.g. Once-for-All *Cai et. al., 2020*
- computation resources often fail to explore the large search space, e.g., 10^{20}
  - proxy metrics for NAS, e.g., convergence speed

Dr. Haojin Yang
HPI, Germany

Reference

- **MobileNetV3**, *Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., ... & Adam, H. (2019). Searching for mobilenetv3. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1314-1324).*
- **MobileNetV2**, *Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4510-4520)*
- **GhostNet**, *Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., & Xu, C. (2020). GhostNet: More features from cheap operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1580-1589).*
- **Once-for-all**, Cai, H., Gan, C., Wang, T., Zhang, Z., & Han, S. (2019). Once-for-all: Train one network and specialize it for efficient deployment. *arXiv preprint arXiv:1908.09791*.
- **ShuffleNet**, *Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6848-6856)*
- **SandglassBlock**, *Daquan, Z., Hou, Q., Chen, Y., Feng, J., & Yan, S. (2020). Rethinking bottleneck structure for efficient mobile network design. arXiv preprint arXiv:2007.02269.*

Dr. Haojin Yang
HPI, Germany

Thank you for your Interest!

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