SIEVE: Speculative Inference on the Edge with Versatile Exportation

Arm Research Summit

Babak Zamirai

University of Michigan, Ann Arbor

September 2019

DNNs on Mobile Devices

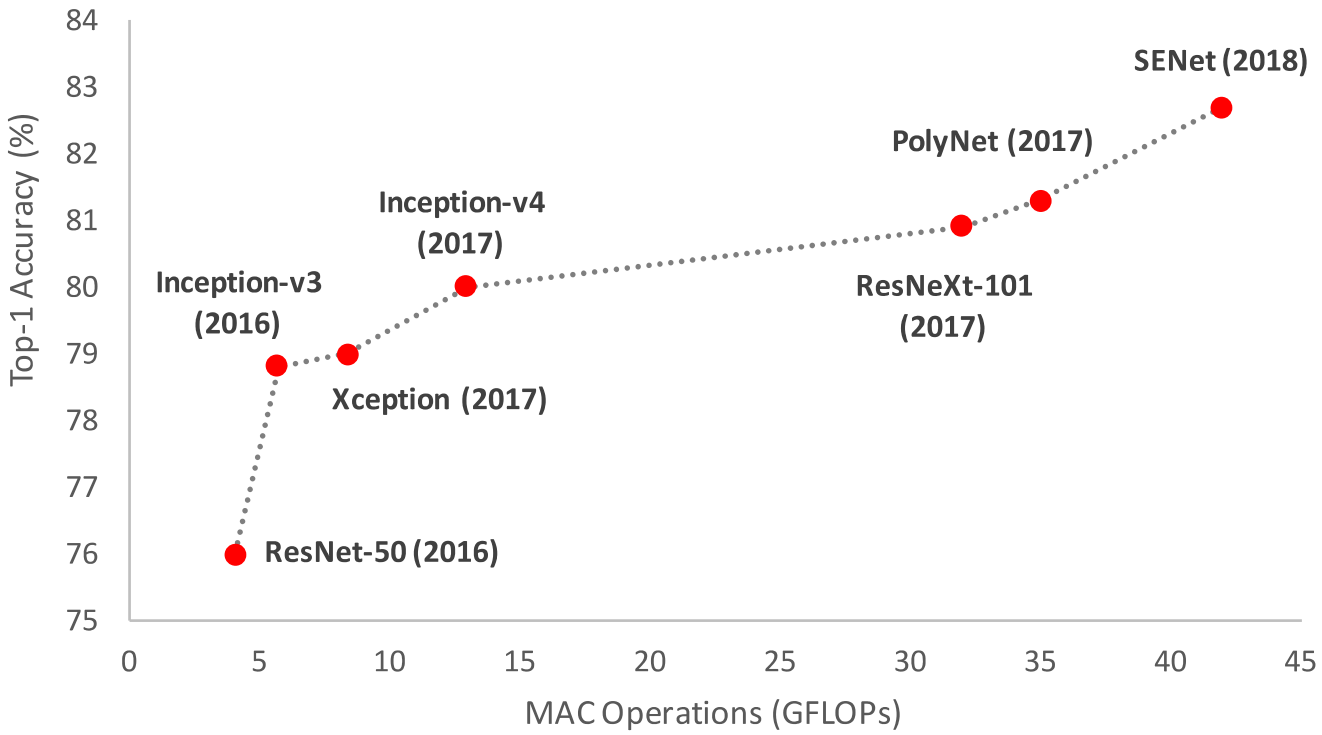
Translation

Facial Recognition

Speech Recognition

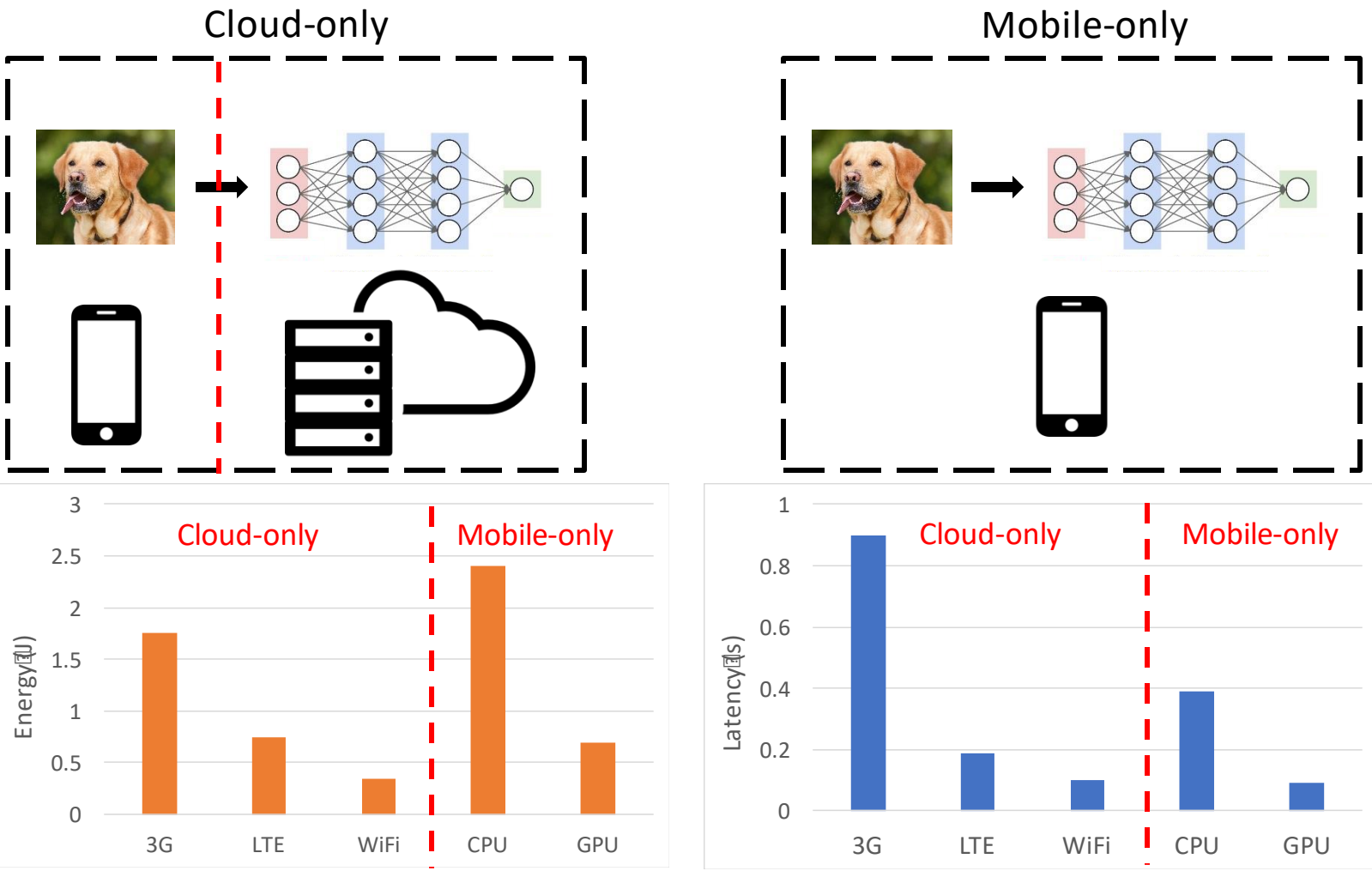
Image Captioning

DNN Complexity Trend



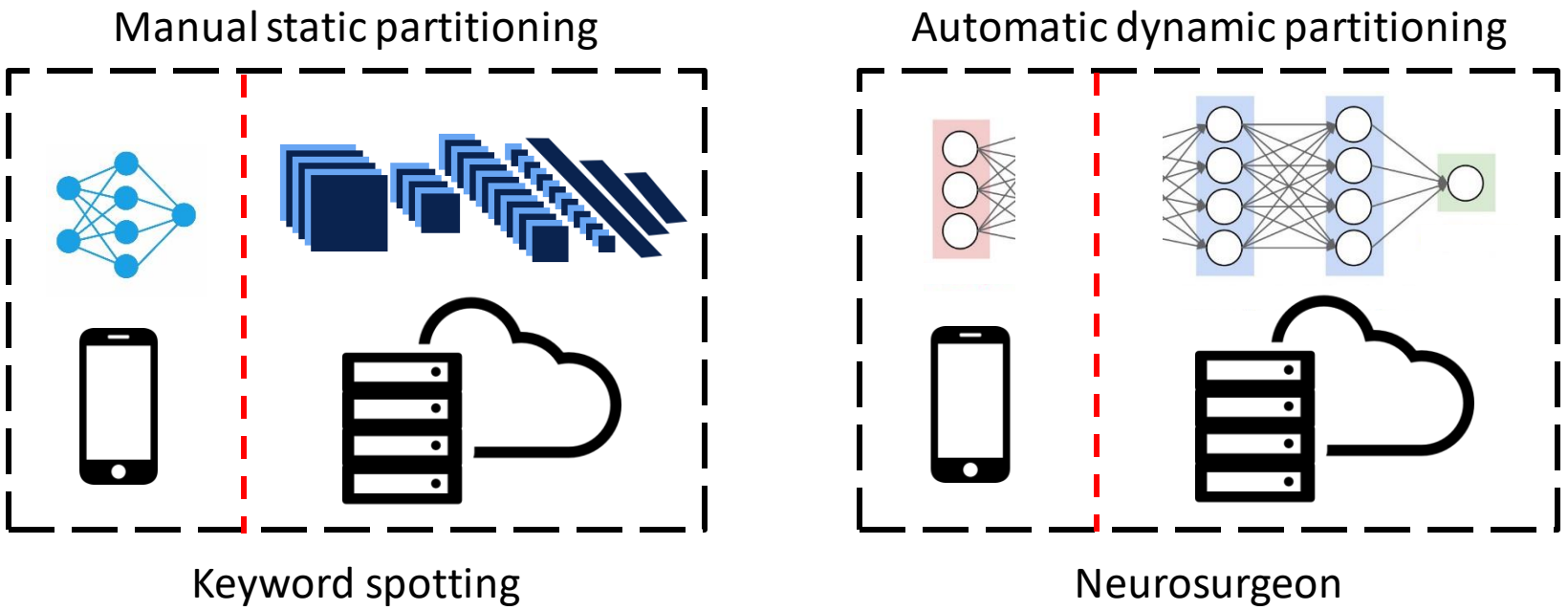
- Computational complexity grows fast
  - ✓ Accuracy improvement
  - ✗ Excessive pressure on mobile hardware

Mobile Deep Learning



- No superior approach

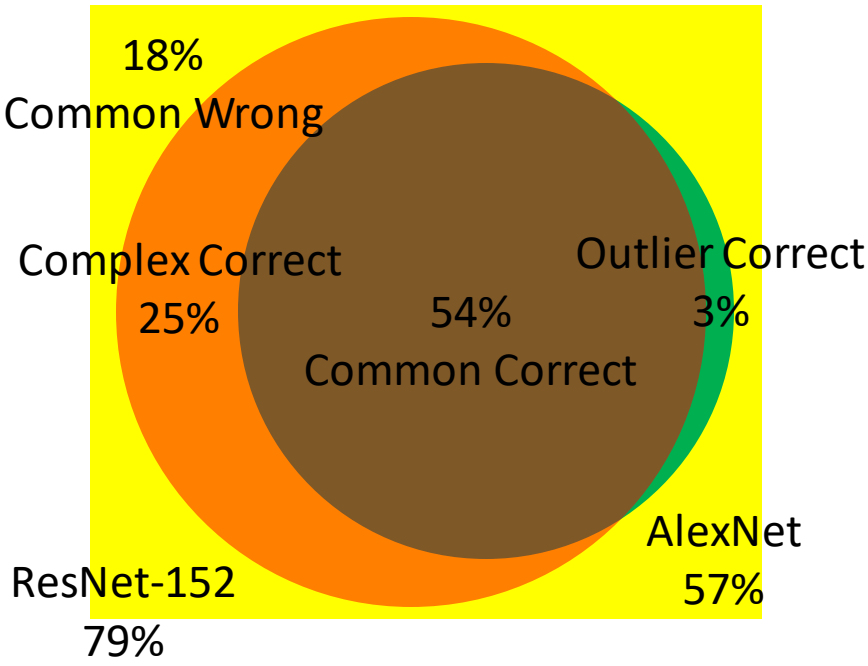
Collaborative Intelligence



- Data-invariant
- Overuse the cloud

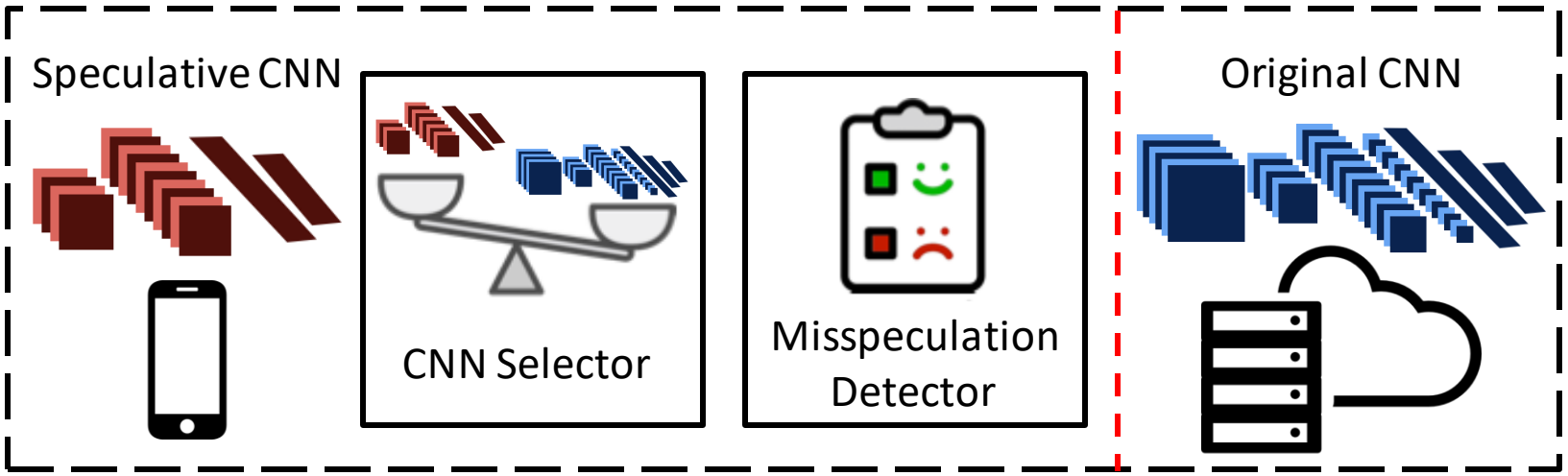
* Kang, Yiping, et al. "Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge." *ASPLOS*. 2017.

Input Variation



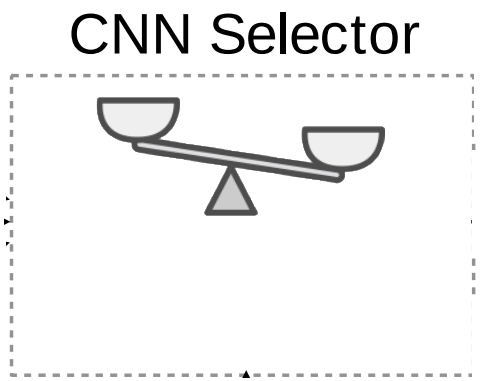
- There is no single best DNN for all inputs
- Combine multiple DNNs
  - Lower computational complexity
  - Higher accuracy

SIEVE

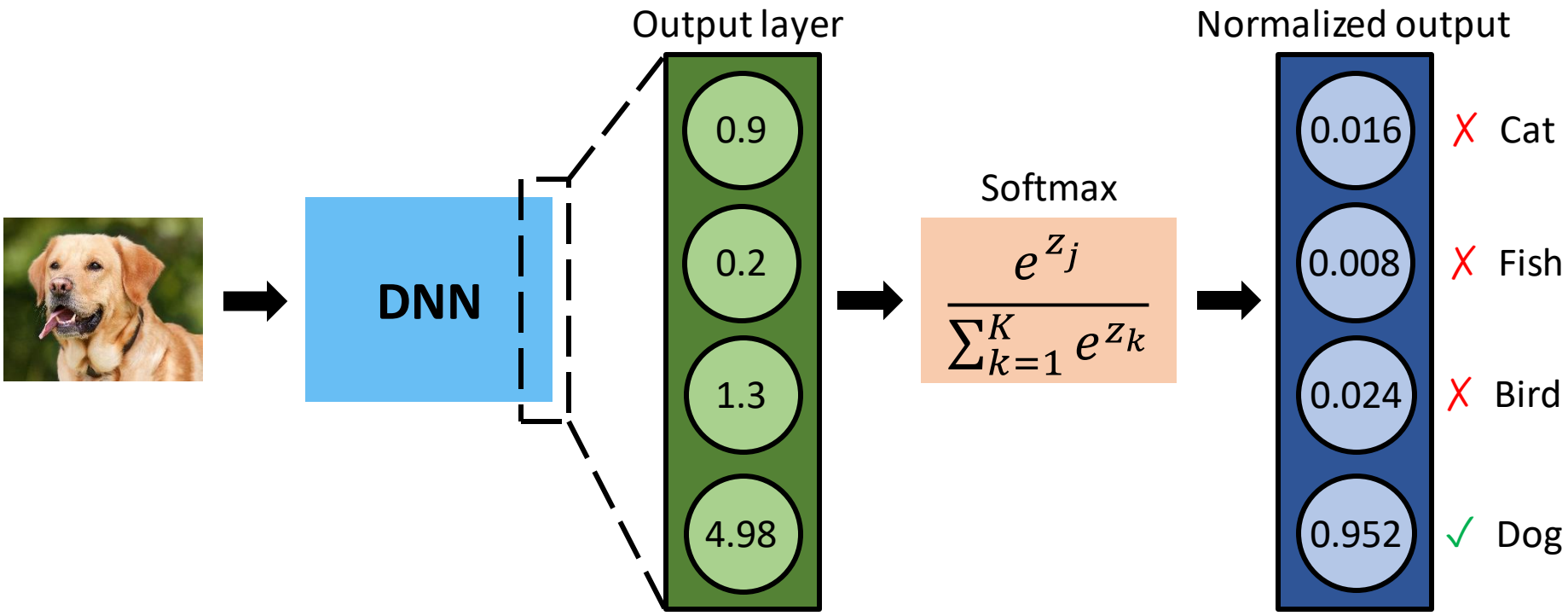


- Speculative hybrid server-edge system
- Choose between speculative CNN and original one
- Detect and recover unreliable speculations

CNN Selection

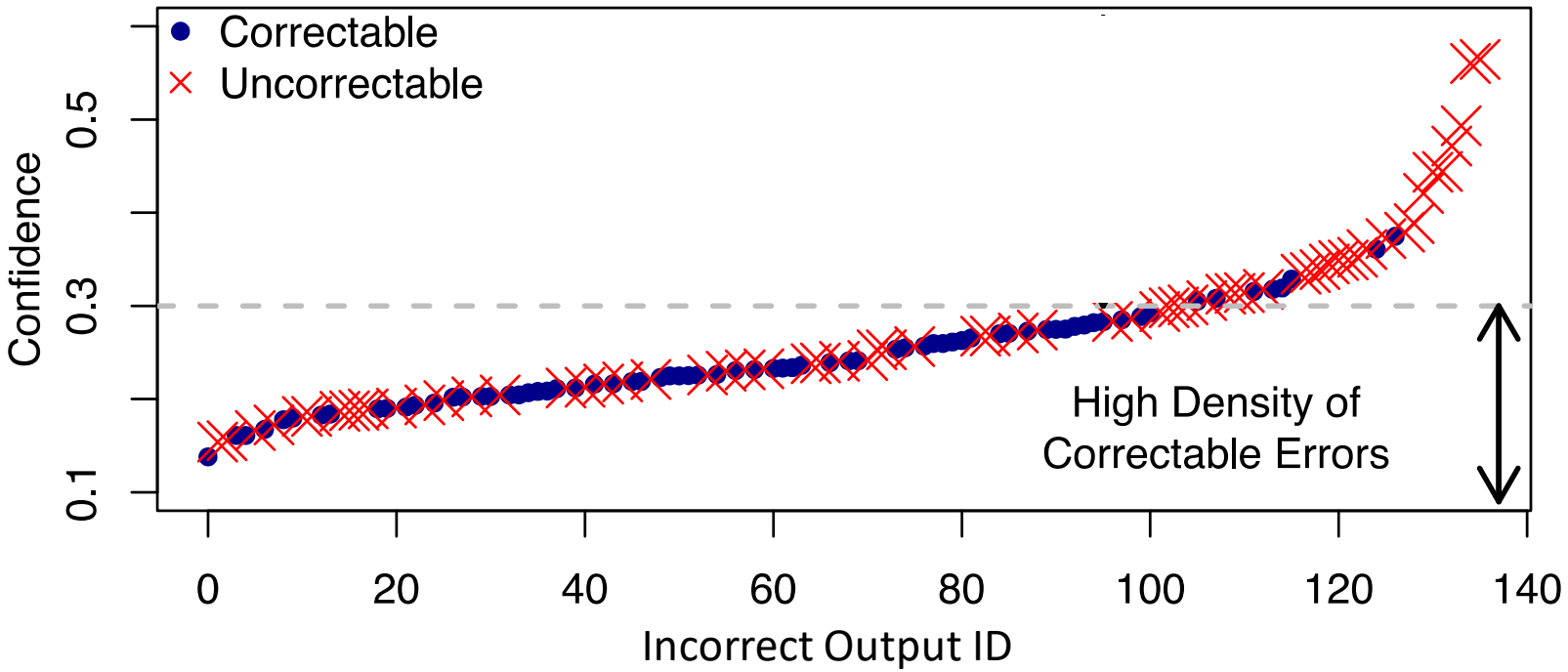


Confidence



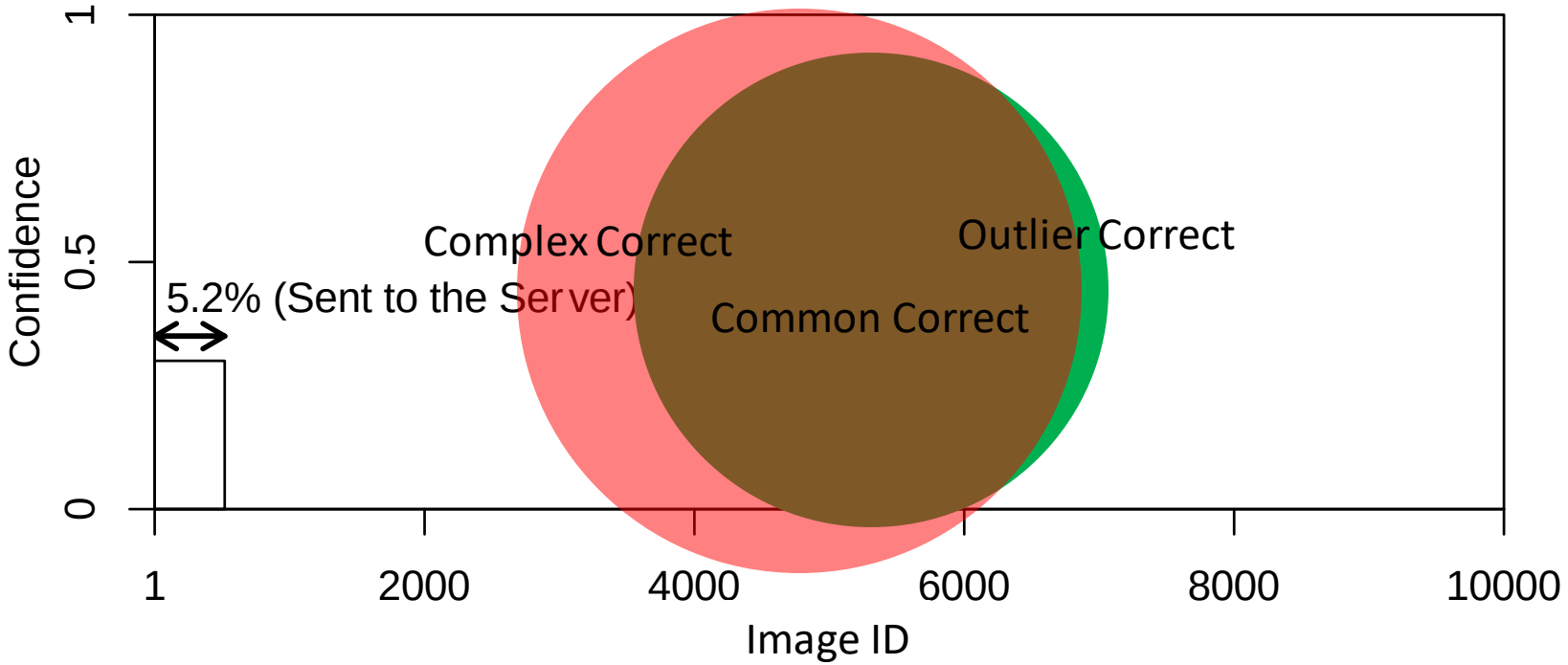
- An estimation of confidence

Unreliable Outputs



- Threshold on confidence of speculative CNN
- Distinguish correctable and uncorrectable errors

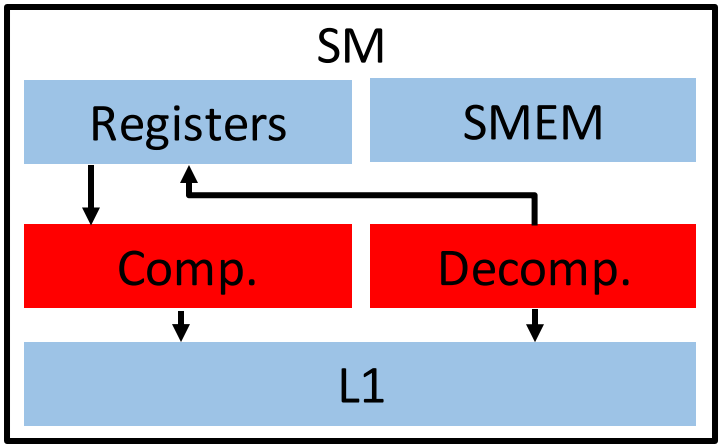
Misspeculation Detection



- Recompute a small portion of outputs
- Recover most of correctable errors
- Some exclusive corrects on the speculative CNN

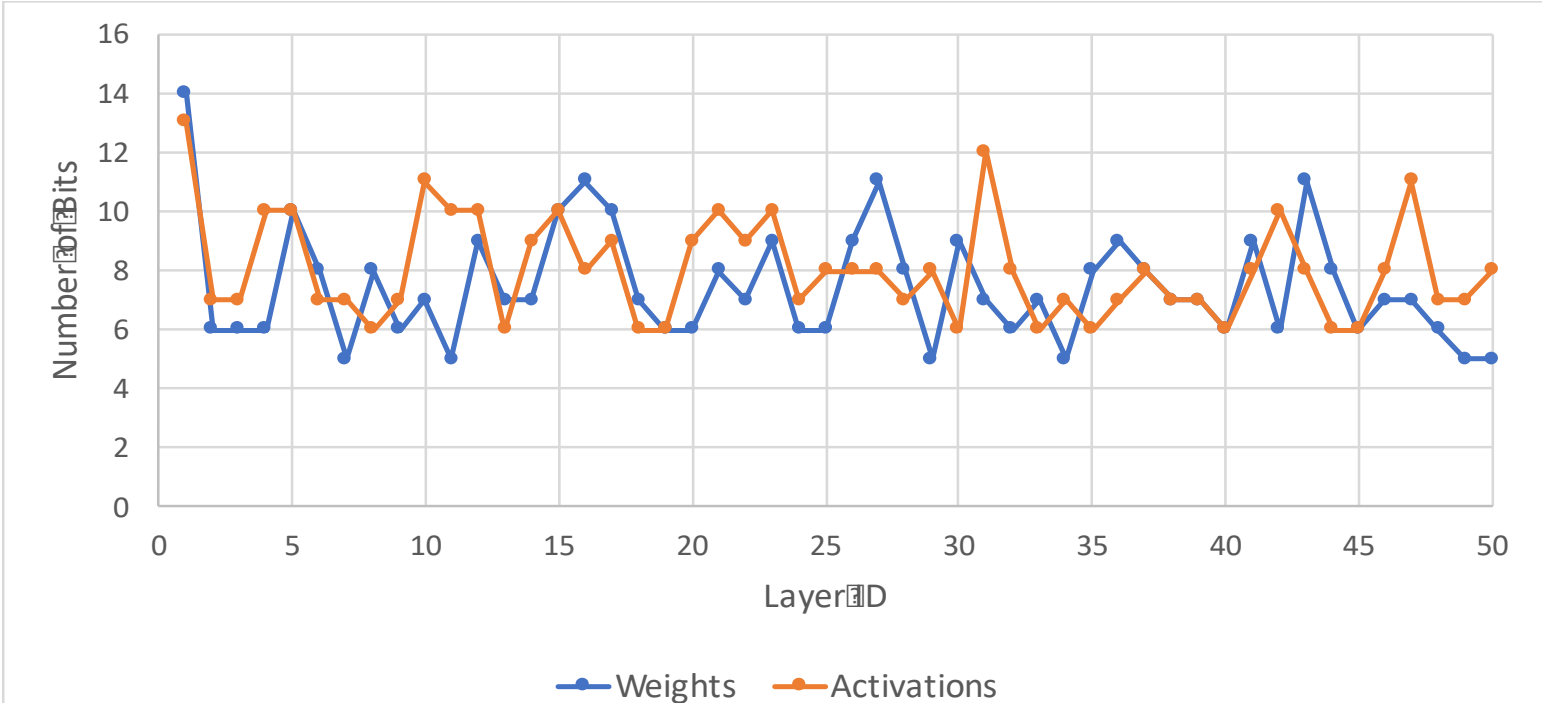
Speculative CNN Design

- Reduced precision memory
- Full precision compute
- ✓ Compressor and decompressor
  - ✓ Floating-point reformatting
  - ✓ 0.05% area and 2.06% power consumption overhead
- ✓ No rigorous CNN modifications and training



Precision Reduction

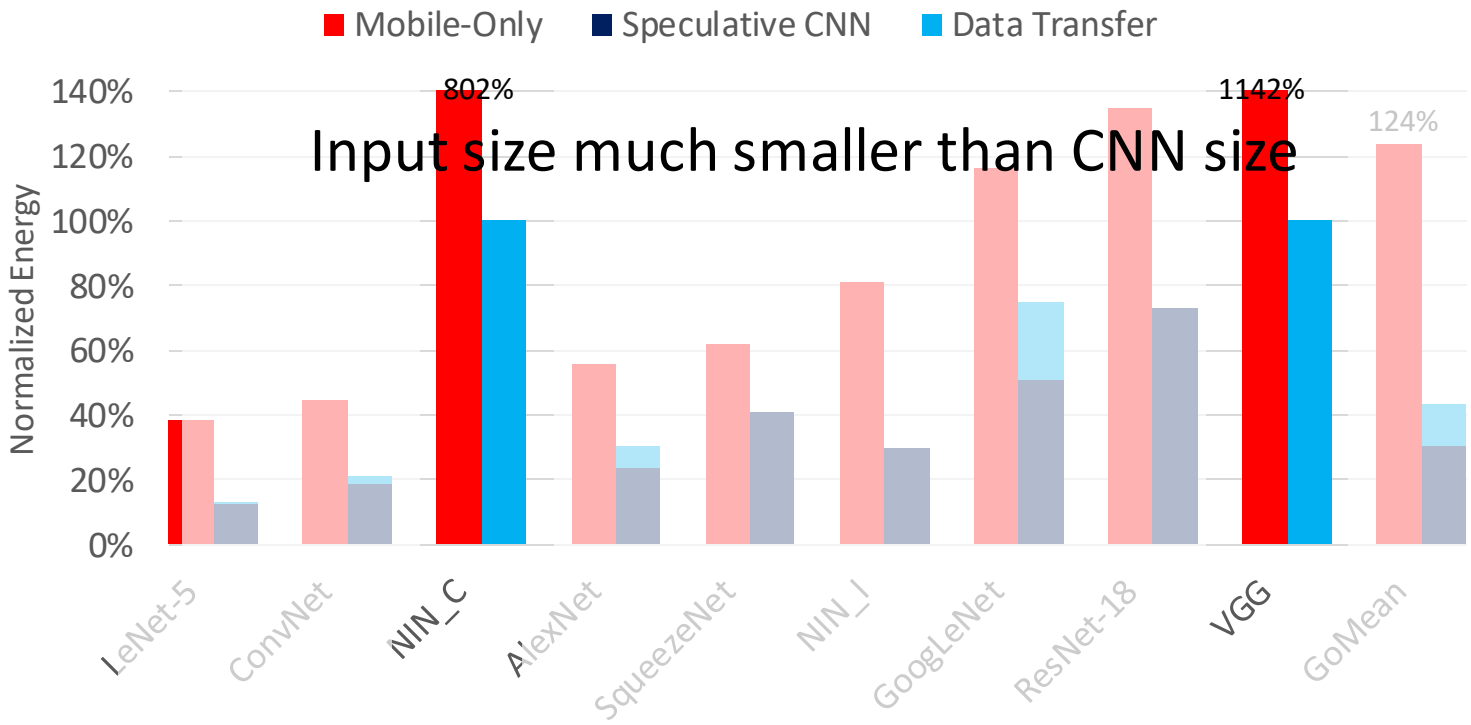
- Specific format per layer
- Aggressive compression
  - Full precision compute
  - Recovery on the cloud



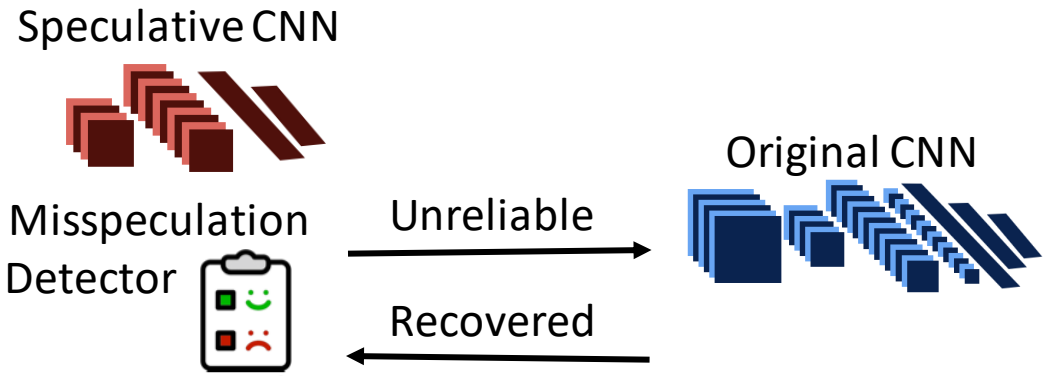
Experimental Methodology

- 9 CNNs on MNIST, CIFAR-10 and ImageNet
- 3G, LTE and WiFi connections
- Mobile: NVIDIA Jetson TK1
  - NVIDIA Tegra K1 SoC
  - Quad-core Arm A15
  - Kepler mobile GPU
- Server: NVIDIA TITAN X GPU
- Caffe framework

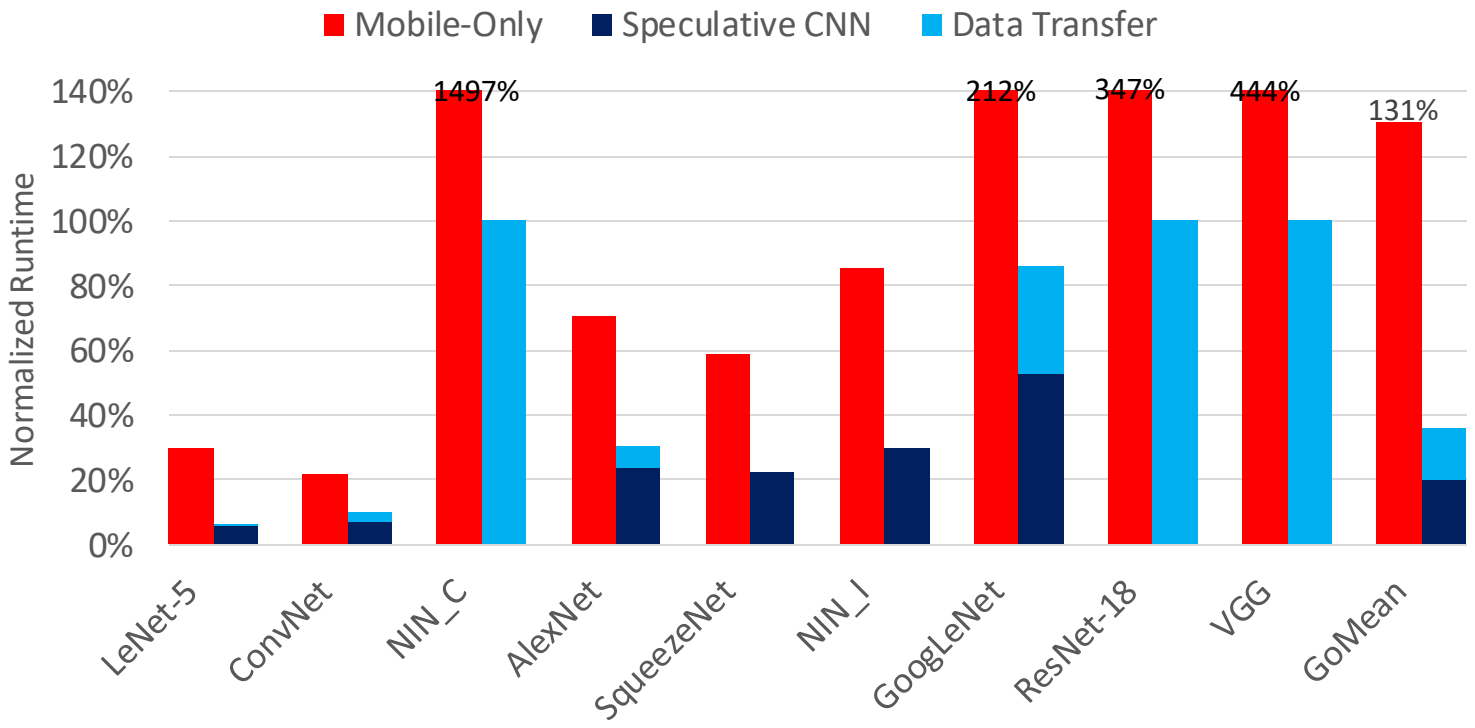
Energy Consumption



- Reduced by:
- 3G 91%
- LTE 57%
- WiFi 26%



Performance



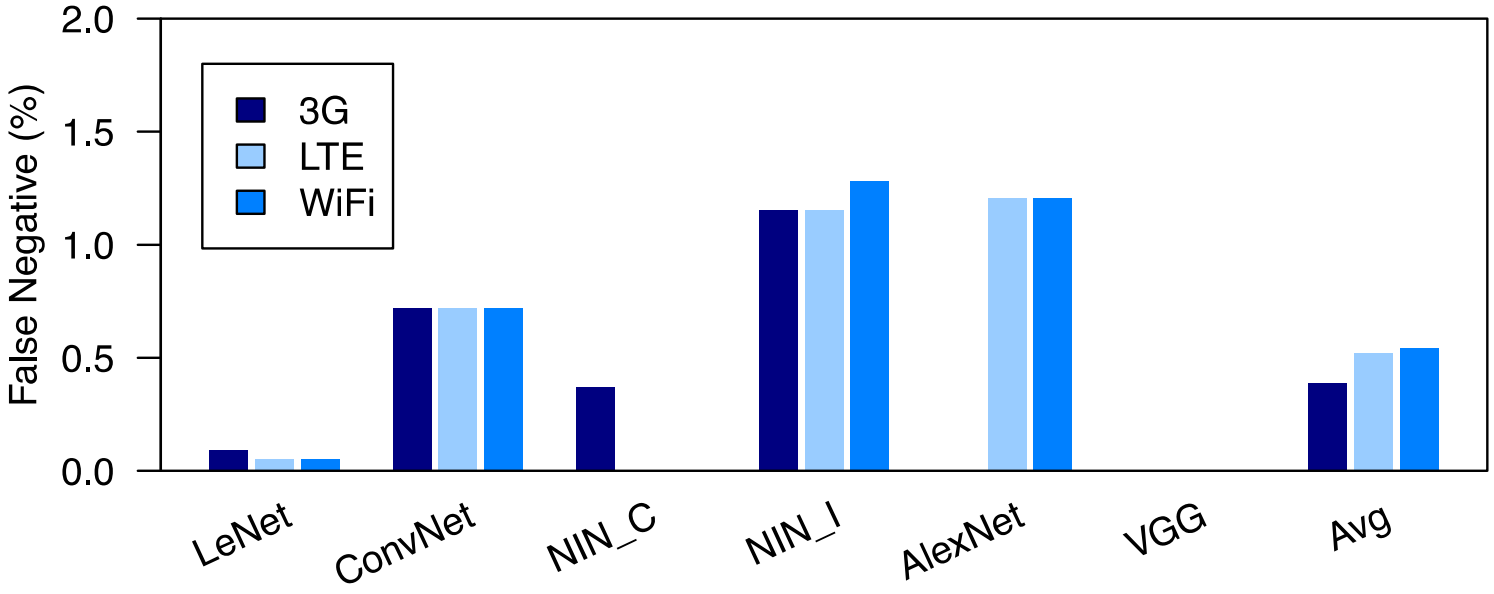
- Speedup:
- 3G 12.3x
- LTE 2.8x
- WiFi 2.0x

Conclusion

- Intelligent hybrid cloud-edge deep learning system
- Speculative inference on device
  - Aggressive precision reduction
- Dynamic computation partitioning technique
- Recover low-confidence inferences
- Using 3G, LTE and WiFi
  - 91%, 57% and 26% energy savings
  - 12.3x, 2.8x and 2.0x speedup

Q & A

Misspeculation Detector



- False negative < 2.0%
  - Correctable error
- False positive < 0.5%
  - Useless recovery