Extracting Machine Learning Models from IoT ML Accelerators via Power & EM Attacks

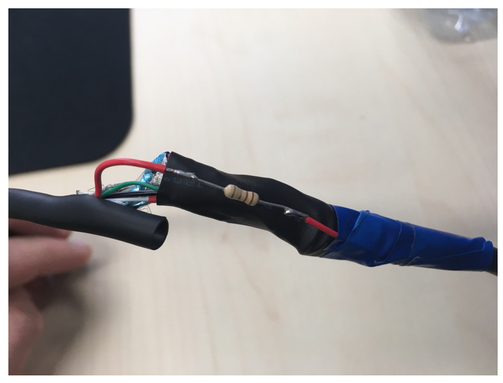
Alexander Treff*, Thore Tiemann, Okan Seker, and Thomas Eisenbarth
University of Lübeck, Institute for IT Security

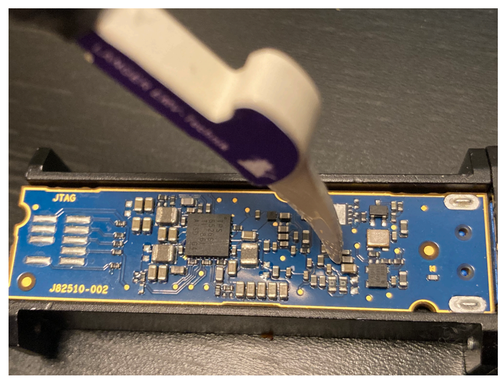
ML Hardware Accelerators

- ML is ubiquitous nowadays
- Transition to IoT and other edge devices (e.g., for surveillance cameras, drones, etc.)
  - Need for dedicated ML hardware accelerators
    - Example: Intel Neural Compute Stick 2 (NCS2)

Side Channel Attacks

- Cheap setup (<500$ oscilloscope)
- Add resistor to USB3.0 extension cable (for power measurements)
- Remove stick shielding for EM measurements

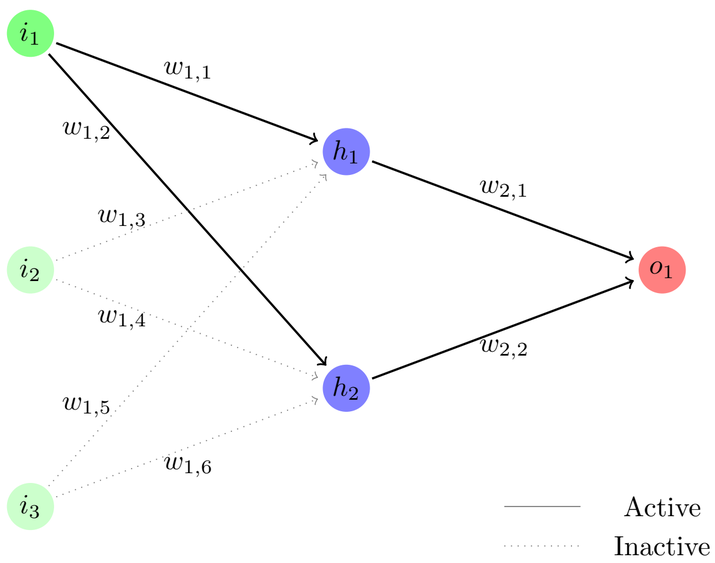


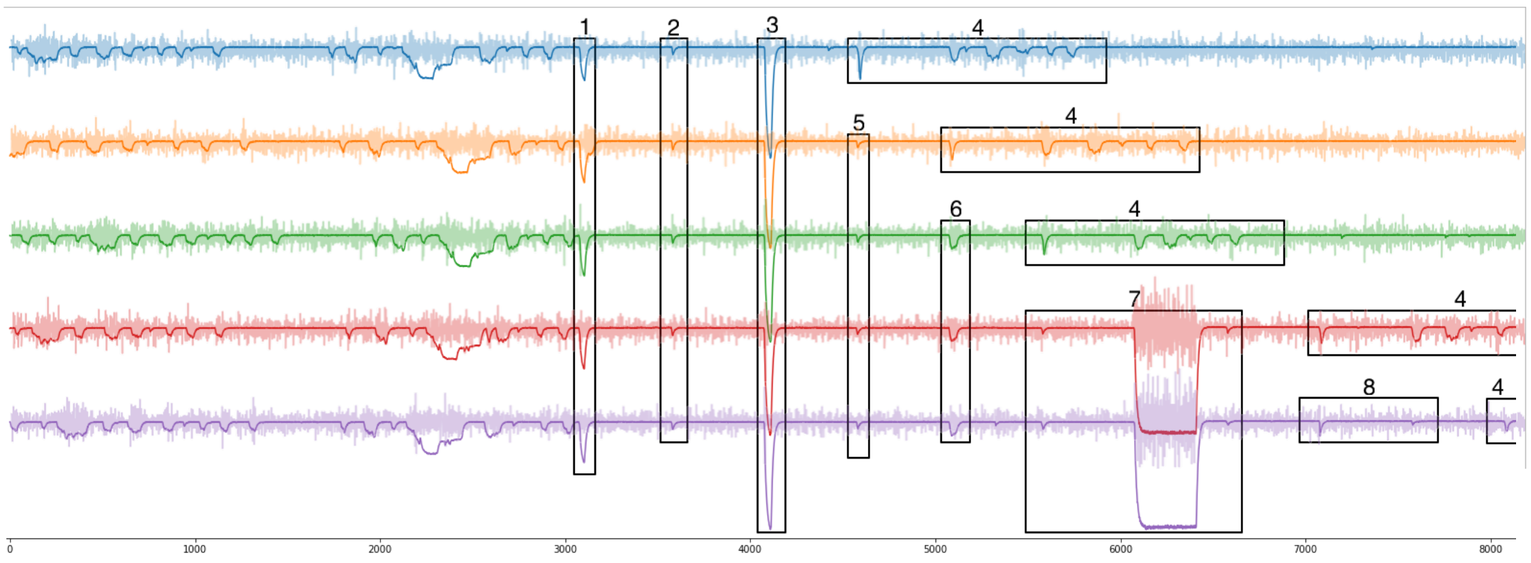


Attack Procedure

- Multiply randomly chosen input with hidden weight (set all other inputs to zero)
- Then, take the hamming weight of the multiplication as a basis for DPA
- Target one byte of computed weight per time
- Correct byte guess correlates well with acquired power traces
- Repeat for all other weights per layer, by adapting the inputs

- IEEE754 float recovery verified on ARM Cortex-M4 board with 32-bit floats.
  - Small imprecisions are fine for floating point numbers





1, 3: small/medium Convolution layer
2, 5: MaxPool layer
6: Transpose layer
7: large Dense layer
8: small Dense layer (hardly noticeable)
4: ”end sequence”, USB traffic

Model Structure Recovery

- Different layer types have different power consumption

- Simple Power Analysis already allows the recovery of a model’s structure to certain extent:
  - Pooling layer result in a very small peak
  - Peaks for Convolution/Dense result in larger peaks
    - Depends on number of multiplications

- Recovery of structure and (approximate) size of shapes

Open Problems / Outlook

- Batina et al. successfully reconstructed models on ARM Cortex-M3 [1], Chmielewski and Weissbart recovered structure on Jetson Nano [2].

- Recovery of weights and more details on hyperparameters on NCS2 is still WIP
- NCS2 is highly parallelized
  - How to identify single operations?

- Different targets: Google Coral, Jetson Nano, …

UNIVERSITÄT ZU LÜBECK
INSTITUTE FOR IT SECURITY

Bibliography

[1] Batina, L., Bhasin, S., Jap, D., & Picek, S. (2019). CSI NN: Reverese Engineering of Neural Network Structures Through Electromagnetic Side Channel. ***28th USENIX Security Symposium (USENIX Security 19), 2019***. https://www.usenix.org/conference/usenixsecurity19/presentation/batina

[2] Chmielewski, L., & Weissbart, L. (2021). On Reverse Engineering Neural Network Implementations on GPU. ***2nd AIHWS Workshop in conjunction with ACNS 2021***. https://eprint.iacr.org/2021/720