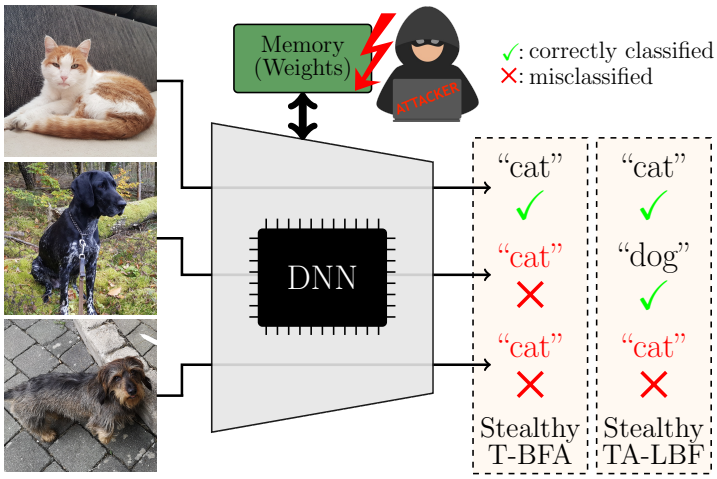
Security of Deep Spiking Neural Networks Against Hardware Bit-Flip Attacks



Deep neural networks (DNNs) are shown to be vulnerable against adversarial weight bit-flip attacks through hardware-induced fault-injection methods on the memory systems where network parameters are stored. These attacks influence the inference process of a DNN by flipping specific bits of the weight representations stored in the memory and hence altering the weights of the DNN. There are several attack algorithms, such as the stealthy targeted bit-flip attack (T-BFA) or targeted attack with limited bit-flips (TA-LBF), that aim to identify such vulnerable bits. To the contrary, there exists minimal countermeasures that one can apply to increase robustness against these attacks (see our recent work [1]). In this project we will explore the effectiveness of existing bit-flip attacks and investigate potential countermeasures using spiking neural network (SNN) architectures.

[1] O. Özdenizci & R. Legenstein, "Improving robustness against stealthy weight bit-flip attacks by output code matching", IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2022.

Goals & Tasks

- Review of the state-of-the-art on adversarial weight bit-flip attacks on SNNs.
- Implementing and simulating bit-flip attack scenarios with SNNs.
- Exploring and extending novel defense algorithms to adversarial weight attacks.

Qualifications

- Interest in machine/deep learning security.
- Experience with Python based deep learning frameworks such as TensorFlow or PyTorch are beneficial.
- Registered to one of the following:
  - ☐ **Bachelor Thesis**
  - ✓ **Seminar Project**
  - ✓ **Master Thesis**

Contact

Ozan Özdenizci
ozan.ozdenizci@igi.tugraz.at