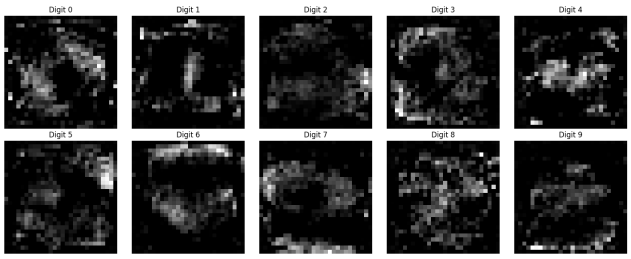
BrainLeaks: On the Privacy-Preserving Properties of Neuromorphic Architectures against Model Inversion Attacks

Supplementary Material

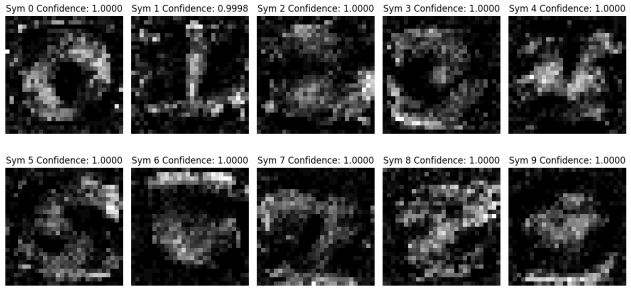
This document provides supplementary results to complement the analysis in the main paper. It contains samples of reconstructed inputs for each dataset, generated from inversion attacks against both ANN and SNN models.

MNIST

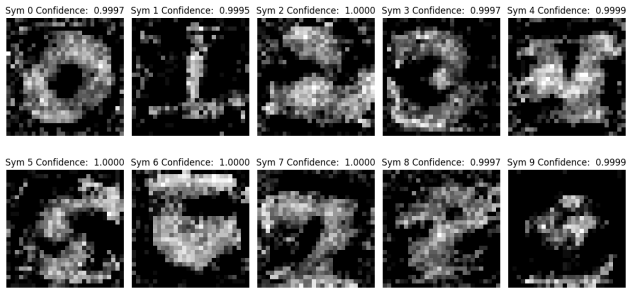
Figure 1 displays reconstructed digit images from the MNIST dataset following inversion attacks on the ANN and SNN models.



(a) ANN Inversion



(b) SNN Inversion (BrainLeaks-v1)

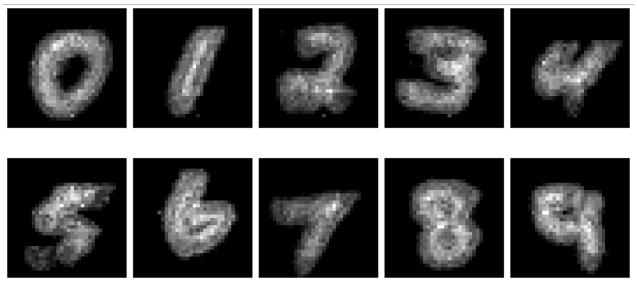


(c) SNN Inversion (BrainLeaks-v2)

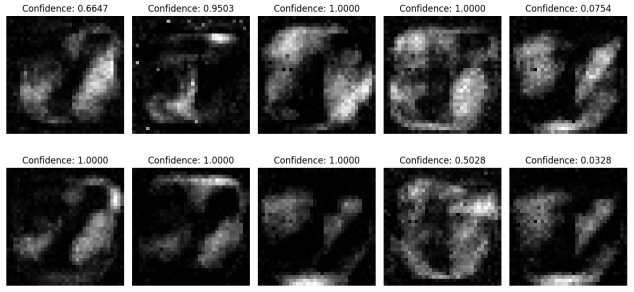
Figure 1. Reconstructed digits from MNIST dataset. (a) Samples reconstructed from the ANN model using the Fredrikson *et al*. inversion attack (b,c) Samples reconstructed from the SNN model using the BrainLeaks-v1 and BrainLeaks-v2 inversion attacks.

Neuromorphic-MNIST

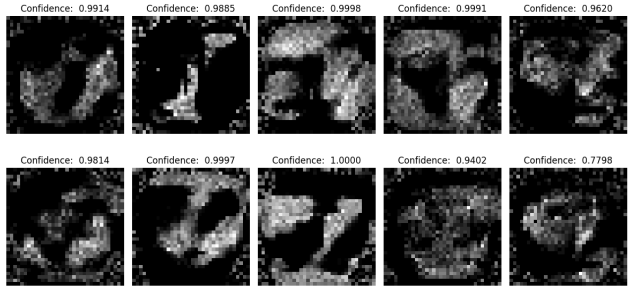
Figure 2 displays reconstructed digits from the N-MNIST dataset using BrainLeaks attacks on the SNN model. Note that the visualization is achieved by employing rate decoding on the spiky reconstructed samples.



(a) Ground Truth



(b) SNN Inversion (BrainLeaks-v1)



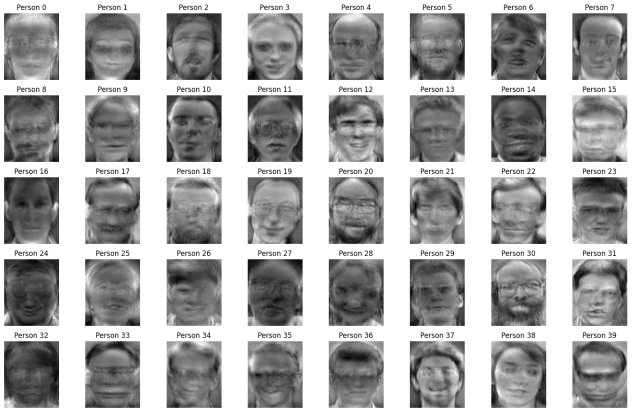
(c) SNN Inversion (BrainLeaks-v2)

Figure 2. Reconstructed digits from the N-MNIST dataset using attacks on the SNN model.(a) Rate-decoded visualization of original N-MNIST samples. (b) Samples reconstructed using the BrainLeaks-v1 inversion attack (c) Samples reconstructed using the BrainLeaks-v2 inversion attack

AT&T Face Database

Figures 3 and 4 show reconstructed faces from the AT&T face database using different inversion methods. Figure 3 contains ANN reconstructions along with the samples from ground truth dataset. Figure 4 shows SNN reconstructions using BrainLeaks-v1 and BrainLeaks-v2 attacks.

(a) Ground Truth



(b) ANN Inversion

Figure 3. Reconstructed faces from the AT&T Face dataset. (a) Original face images from the dataset. (b) Samples reconstructed from the ANN model using the Fredrikson *et al*. inversion attack

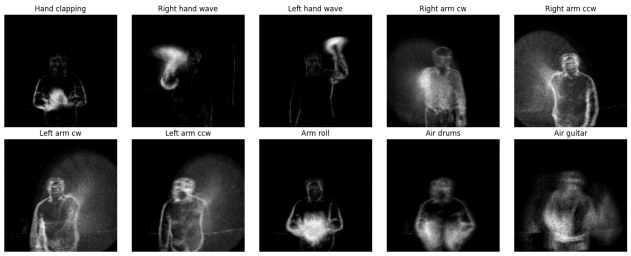
(a) SNN Inversion (BrainLeaks-v1)

(b) SNN Inversion (BrainLeaks-v2)

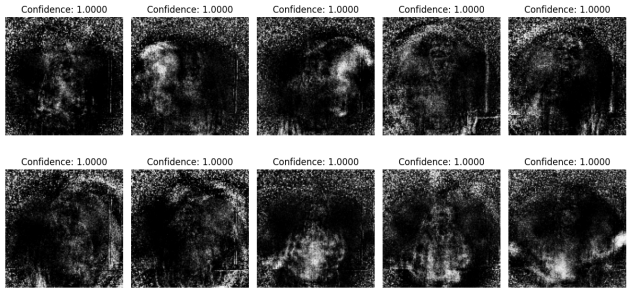
Figure 4. Reconstructed faces from the AT&T Face dataset using attacks on the SNN model. (a) Samples reconstructed using the BrainLeaks-v1 inversion attack (b) Samples reconstructed using the BrainLeaks-v2 inversion attack.

IBM DVS Gesture Dataset

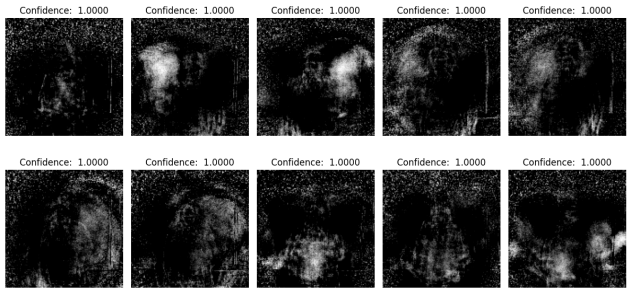
Figure 5 shows reconstructed gestures from the IBM DVS Gesture dataset following BrainLeaks attacks on the SNN model.



(a) Ground Truth



(b) SNN Inversion (BrainLeaks-v1)



(c) SNN Inversion (BrainLeaks-v2)

Figure 5. Reconstructed gestures from the IBM DVS Gesture dataset using attacks on the SNN model. (a) Rate-decoded visualization of original IBM DVS Gesture samples. (b) Samples reconstructed using the BrainLeaks-v1 inversion attack. Bottom: (c) Samples reconstructed using the BrainLeaks-v2 inversion attack. Bottom: