**Abstract:**

Neural network models are high-value intellectual property—trained over large datasets with significant computational resources. Any compromise can lead to IP theft, privacy violations and commercial losses. This thesis proposes a comprehensive design for building a secure and high-performance computing platform for neural networks, addressing key challenges from encryption to side-channel resilience. Securing neural processing units (NPUs) requires enforcing key security properties such as confidentiality, integrity, authentication, freshness and resistance to side-channel attacks. To preserve model confidentiality, encryption of weights and biases is essential. However, performing encryption in shared components, such as the CPU in a system-on-chip (SoC) can expose the system to cache side-channel attacks (CSCA), particularly when widely trusted encryption routines like AES are implemented in software. In our first contribution, we review CSCA countermeasures and show that address-based defenses, while obfuscatory, can still be broken in n^(O(log(log(n)))) time. To make such attacks more practical, we propose a theoretical classification framework along with a two-phase attack strategy that reduces the required attempts to just 51,000, underscoring the urgency for a fast, hardware-based secure encryption engine. Consequently, we designed a lightweight, high-throughput encryption engine that employs ephemeral keys and fewer rounds, yet achieves AES-equivalent security with up to 5X lower latency.

Next, we turn to the system-level vulnerabilities in NPUs, particularly their reliance on off-chip DRAM and system buses, which remain unencrypted and vulnerable to tampering, replay, cold-boot and bus snooping attacks. We analyze and model memory access behavior across popular neural network dataflows and develop lightweight hardware that enables integrity and freshness verification at the layer level—achieving a 20% speedup. We also address memory- and timing-based side-channel attacks. We categorize existing defenses and show their limitations through practical attacks. To counter them, we propose a novel compression- and binning-based obfuscation technique, which applies multiplicative noise to model parameters. This expands the attack search space to over 10^(70) . This work presents a unified design framework for secure NPUs—spanning encryption, memory protection, and side-channel protection — all without sacrificing performance, making it both practical and resilient against evolving attack vectors.