Jigsaw: Partitioning DNN for Model Extraction Defense

Hunter Baxter

August 4, 2023

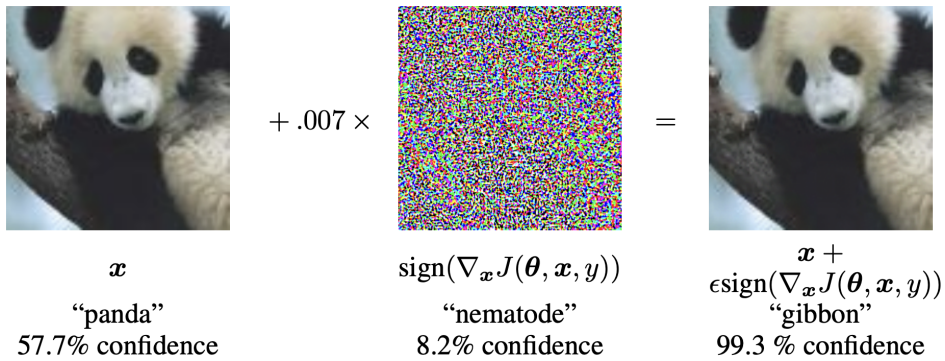


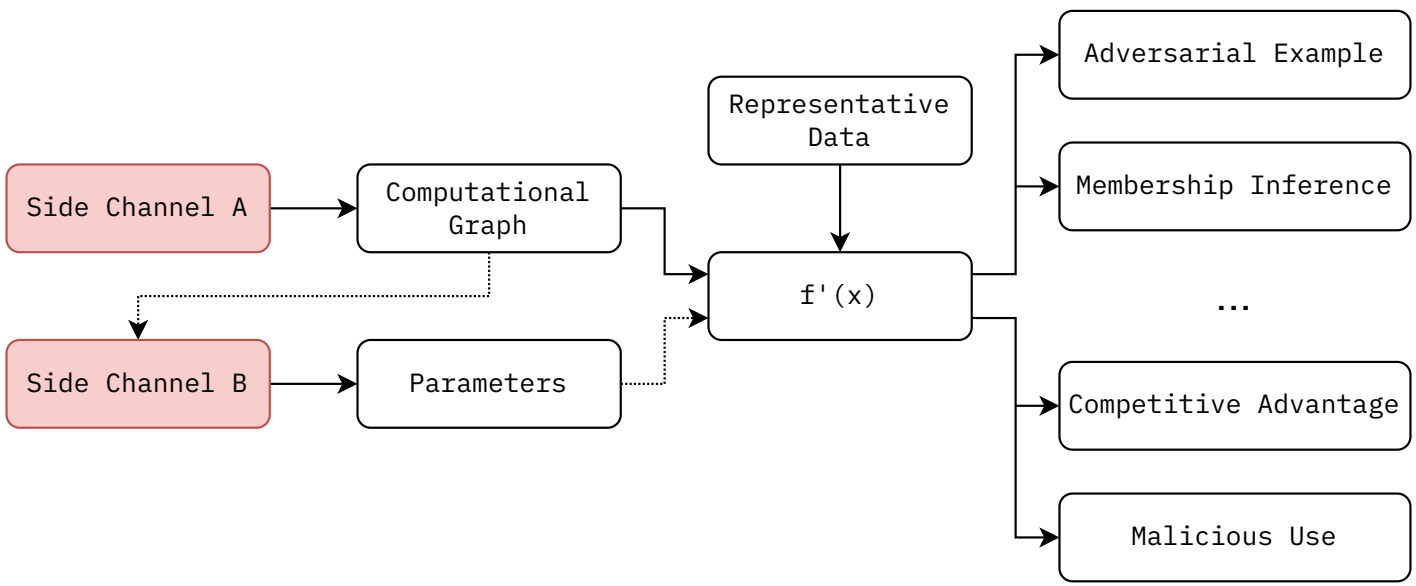
Introduction

- PI = Dr. Kevin Leach
- Deep Learning Infrastructure Security
- Developed Jigsaw
  - Input: PyTorch Neural Network
  - Output: Equivalent model distributed amongst nodes
  - Result: Reduced model extraction probability

Model Extraction

- An adversarial attack where a private model $f(x)$ is completely reconstructed
- Model Extraction accelerates malicious activity

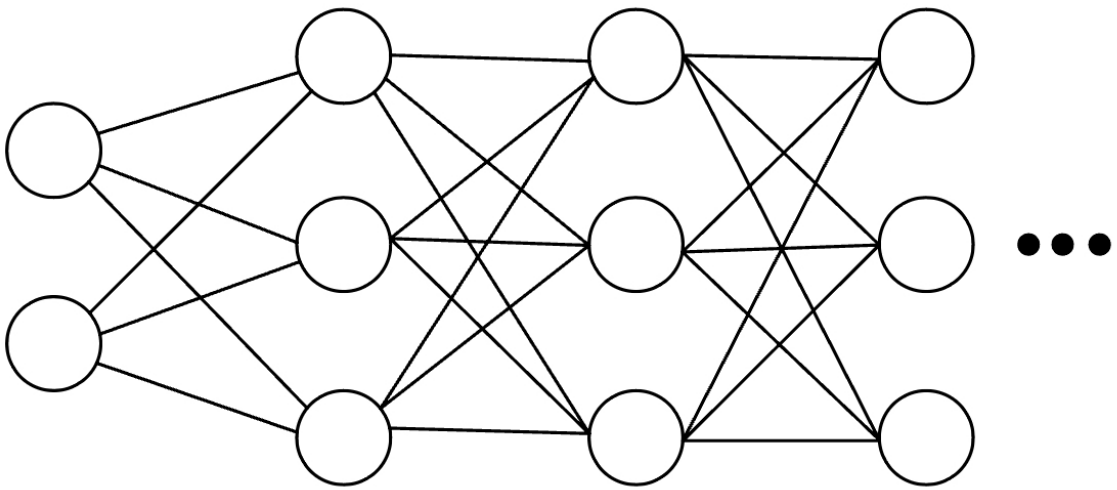




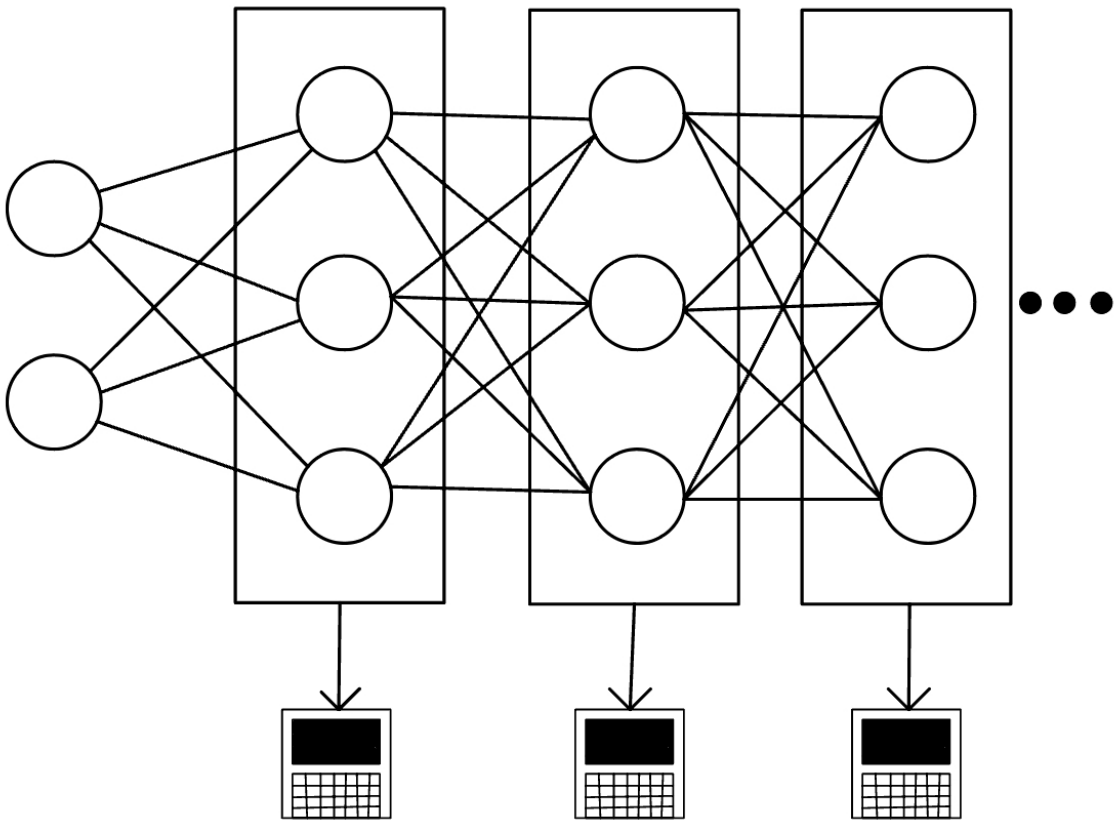
Prior Work

- Strong Isolation
  - Limited Scalability
  - Limited Performance
  - Side-Channel Vulnerabilities
  - Re-engineered Solutions
- Jigsaw uses a moving target defense approach
- Information gain at t may not work at $t + \delta$
- Probabilistic, not strong, isolation

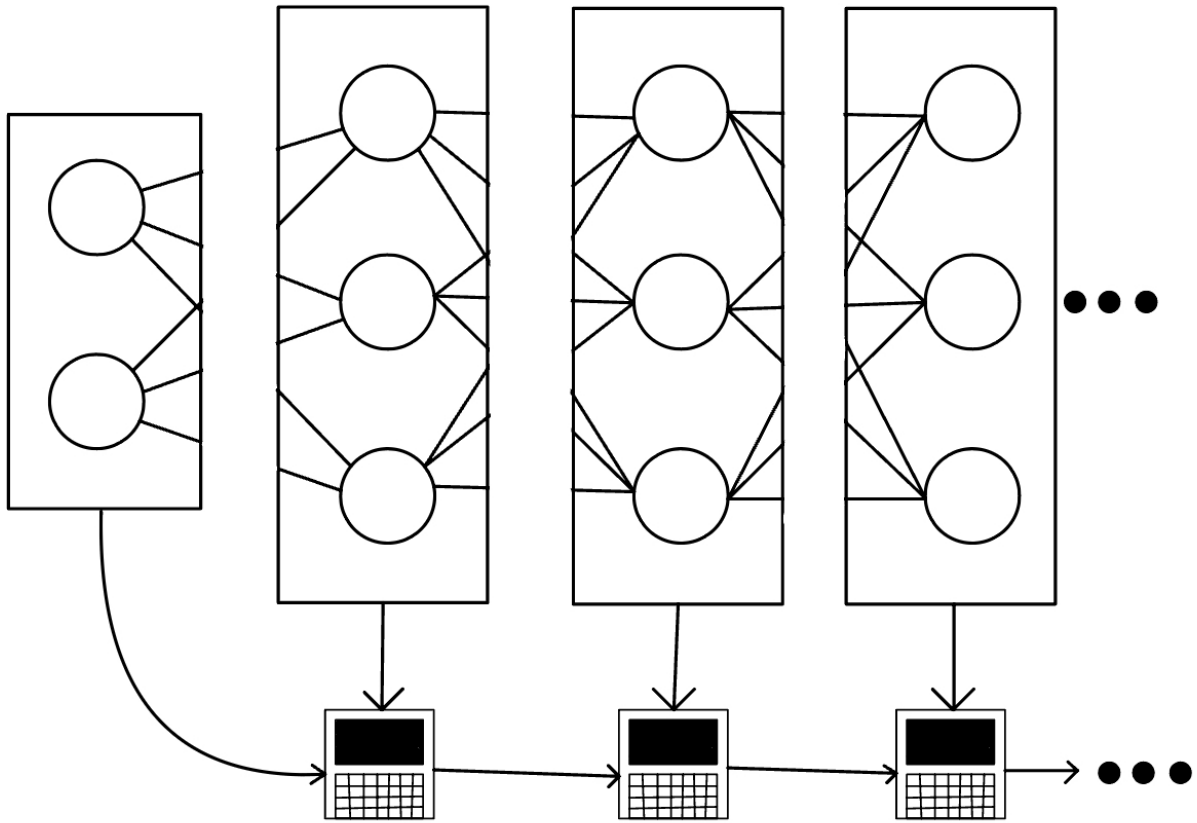
Jigsaw



Jigsaw

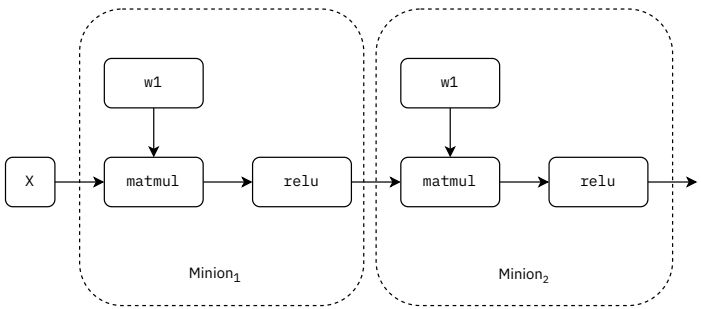


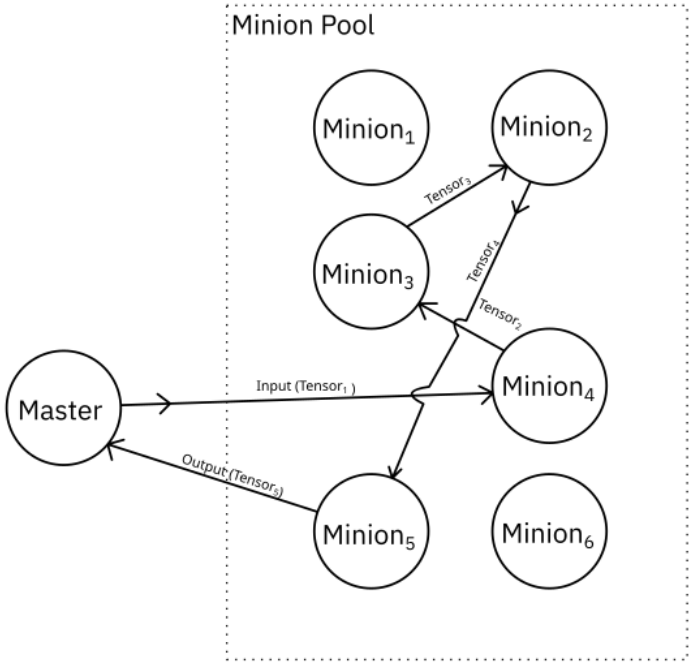
Jigsaw

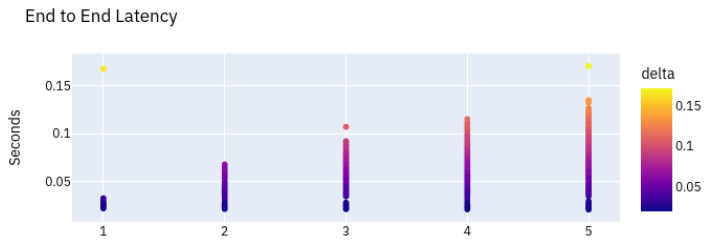


Jigsaw

- Splits neural network amongst a pool of minion nodes
- $\sum_{a=1}^{m} \binom{m}{a} a! \binom{k-1}{a-1}$ possible schedules
- Linear increase in resources to fully extract the model







ISIS Concluding Remarks

- Lessons learned?
  - Rapid prototyping is valuable
- Challenges faced?
  - Integration testing takes time
- What went well?
  - Work was entertaining