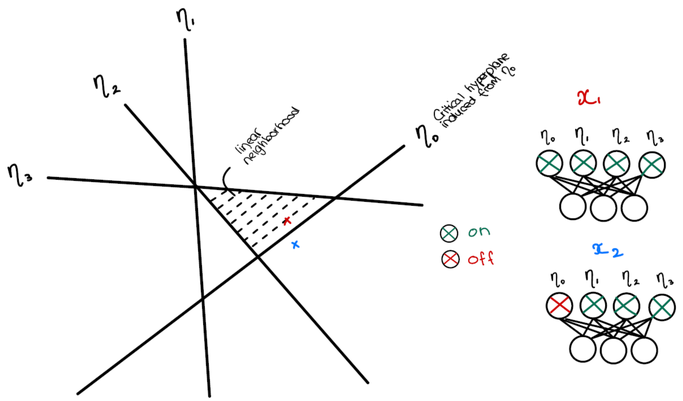
Beyond Slow Signs in High-fidelity Model Extraction

Hanna Foerster, Robert Mullins (Department of Computer Science, University of Cambridge)
Ilia Shumailov, Jamie Hayes (Google DeepMind)

UNIVERSITY OF CAMBRIDGE
Department of Computer Science and Technology

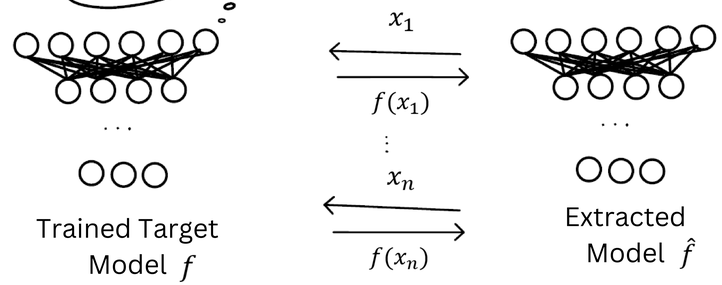
link to our github page

NEURAL INFORMATION PROCESSING SYSTEMS

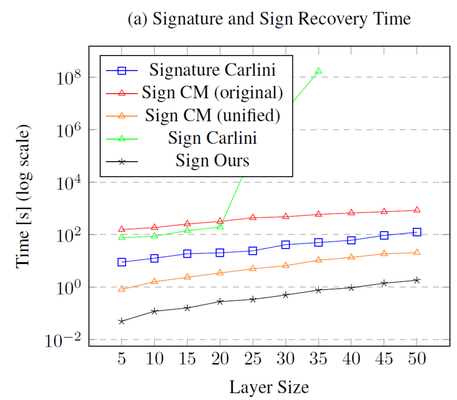


Cryptanalytical extraction, how does it work?

- Neurons contribute to the final output by activating or deactivating given an output
- Two part extraction:
  - Signature extraction (extraction of weights up to a multiplicative factor)
  - Sign extraction



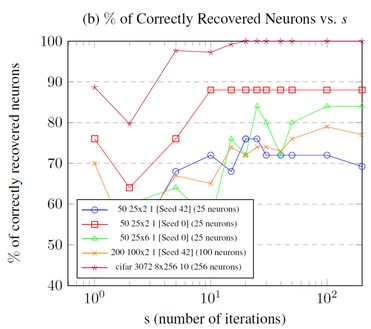
Beyond Slow Signs...



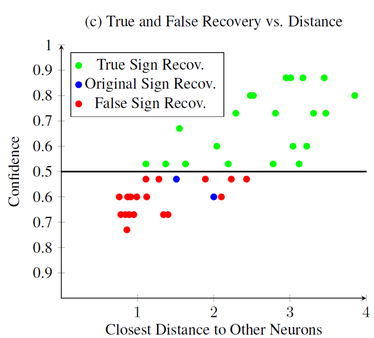
Sign Extraction SpeedUp
~Up to 16.4 times

Whole extraction SpeedUp:
~Up to 2.5 times

'Easy' and 'Hard' to Sign Extract Neurons



Percentage of correct sign recovery does not change with more iterations after s=15



If target neuron is closer to others, with high confidence we have wrong sign recovery

DNNs with fully connected layers and RELU activation functions can be extracted in polynomial time as long as they have only around 4 hidden layer with less than 256 neurons per layer, e.g., small models used in healthcare or in specialised areas controlling policies in nuclear fusion models can be stolen.

Discussion

Model Information			Signature [s]		Sign [s]		Queries	
Model (Training Seed)	Layer	Params	Mean	Var	Mean	Var	Mean	Var
784-8x2-1 (s1)	2	72	10.39	0.25	0.25	0.002	$5.13 \cdot 10^4$	$4.9 \cdot 10^8$
784-16x2-1 (s1)	2	272	7.22	8.85	0.60	0.005	$6.92 \cdot 10^4$	$9.3 \cdot 10^8$
784-32x2-1 (s1)	2	1056	22.58	31.59	2.07	0.61	$2.28 \cdot 10^5$	$3.7 \cdot 10^9$
784-64x2-1 (s1)	2	4096	135.32	$2.9 \cdot 10^3$	7.17	6.32	$9.03 \cdot 10^5$	$1.9 \cdot 10^{10}$
784-128x2-1 (s1)	2	16512	758.5	$1.5 \cdot 10^5$	30.46	8.02	$4.17 \cdot 10^6$	$1.1 \cdot 10^6$
784-128x2-1 (s2)	2	16512	1040.85	103.32	30.66	5.72	$4.35 \cdot 10^6$	$1.5 \cdot 10^6$
MNIST784-8x2-1 (s2)	2	72	12.75	9.17	0.26	0	49,730	$9.6 \cdot 10^5$
MNIST784-16x2-1 (s2)	2	272	19.15	37.03	0.67	0.01	$1.92 \cdot 10^5$	$4.6 \cdot 10^9$
MNIST784-32x2-1 (s2)	2	1056	98.10	1179.81	2.00	0.07	$7.7 \cdot 10^5$	$8.0 \cdot 10^{10}$
MNIST784-64x2-1 (s2)	2	4096	496.2	$1.5 \cdot 10^5$	6.32	0.32	$3.05 \cdot 10^6$	$1.1 \cdot 10^{13}$
MNIST784-64x2-1 (s1)	2	4096	4649.95	$1.6 \cdot 10^6$	6.85	1.79	$4.9 \cdot 10^6$	$2.8 \cdot 10^{13}$
MNIST784-16x8-1 (s2)	1	12560	$1 \cdot 10^4$	-	63.04	-	$5.38 \cdot 10^6$	-
MNIST784-16x8-1 (s2)	2	272	470.19	$3.4 \cdot 10^4$	0.67	0	$5.27 \cdot 10^5$	$1.2 \cdot 10^{10}$
MNIST784-16x8-1 (s2)	4	272	$> 36hrs$					
MNIST784-16x8-1 (s2)	8	272	$> 36hrs$					
MNIST784-16x8-1 (s2)	9	17	0.01	0	0	0	100	0
MNIST784-16x3-1 (s1)	2	272	1854.42	$2 \cdot 10^6$	0.96	0.15	$9.7 \cdot 10^6$	$5.2 \cdot 10^{13}$
MNIST784-16x3-1 (s1)	3	272	$6.9 \cdot 10^4$	-	0.54	-	$4.4 \cdot 10^7$	-

In green one can see how layer 2 extraction for the same number of neurons can vary with model depth. In blue one can see the variance of extracting two models trained similarly but on different randomness. In red one can see how deeper layers become increasingly hard to extract.