Hardware Acceleration of Lattice Networks

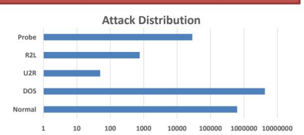
Matthew Feldman, Tushar Swamy { mattfel | tswamy }@stanford.edu

Introduction

- Problem: Detecting malicious packets in high speed datacenter networks is nearly impossible. We propose to solve this with hardware acceleration.
- Lattices are a new low latency building block for neural networks. They use interpolated, n-dimensional look-up tables to transform data.
- FPGAs are reprogrammable digital circuit devices that have recently grown in popularity due to their low power and ability to parallelize computation.
- Our work: A side by side comparison of a lattice network and DNN running on an FPGA and CPU.

Dataset and Features

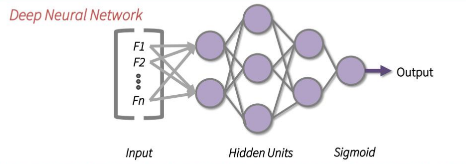
- Data: KDD Cup Dataset.
  - ~750MB of normal network data
  - 5e6 packets are split as 2% train, 2% dev, 96% train to test generalization of lattices.
  - 41 features per row ranging from header information to flow level information in the form of strings, ints, and floats. Each row corresponds to a packet.

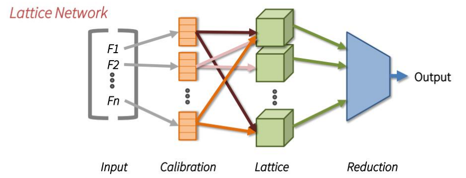


Theory

- *Interpolation on k vertices:* $\phi_k(x) = \prod_{d=0}^{D-1} x[d]^{v_k[d]}(1 - x[d])^{1-v_k[d]}$
- *Applying lattice parameters:* $f(x) = \theta^T \phi(x)$
- *Objective Function:* $\theta = argmin_\theta \sum_{i=1}^{n} l(y_i, \theta^T\phi(x_i)) + R(\theta)$
- *Input Calibration:* $c(x[d]; a, b) = \sum_{k=1}^{K} \alpha[k] ReLU\,(x - a[k]) + b[1]$
- *Torsion Regularization* $R(\theta) = \sum_{d=1}^{D} \sum_{d'=1, d'\neq d}^{D} \sum_{r,s,t,u} ((\theta_r - \theta_s) - (\theta_t - \theta_u))^2$
- *Optimizer and Loss:* *ADAM and Squared Error*

Models

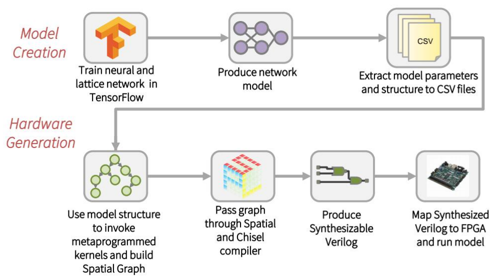




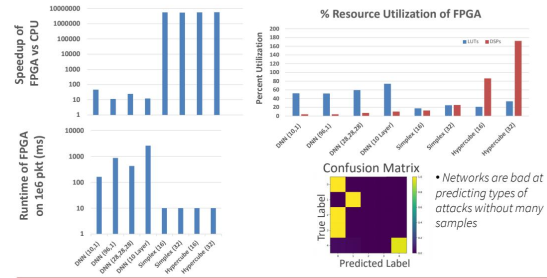
Model	Train Accuracy	Test Accuracy	Recall	False Negatives
DNN (10,1)	0.9999	0.9990	0.9999	4202
DNN (96,1)	0.9999	0.9990	0.9999	2844
DNN (28,28,28,1)	0.9999	0.9991	0.9998	3266
DNN (10 Layer)*	0.9999	0.9937	0.9999	3041
Simplex (16)	0.9988	0.7685	0.9999	253
Simplex (32)	0.9970	0.7699	0.9935	6090
Hypercube (16)	0.9998	0.7687	0.9996	377
Hypercube (32)	0.9998	0.7686	0.9990	879

*(128,128,64,64,64,64,64,64,32,32,1)

Model to Hardware Generation



Evaluation



Discussion

- Tradeoff: Network admins need to choose their models based on the speed of the datacenter as well as tolerance of malicious packets. Neither DNNs nor lattices are clear winners.
- FPGAs can accelerate DNNs up to *46x* and lattices up to *5.5Mx*
- Lattices are well suited to hardware because of their reliance on lookup tables and reduction trees.
- Overfitting is still an issue for lattices even with torsion. There is not much literature on dealing with bias/variance for lattices.
- Dataset is not well rounded which may have contributed to DNN's success. We need more real world tests.

Future Work

- Explore combinations of DNN units and lattices
- Experiment with lattice structures like embedded tiny lattices
- Investigate the effect of reduced precision on lattices

References

- Gupta, Maya, et al. "Monotonic calibrated interpolated look-up tables." The Journal of Machine Learning Research 17.1 (2016): 3790-3836.
- Mane, Vrushali D., and S. N. Pawar. "Anomaly based ids using backpropagation neural network." International Journal of Computer Applications 136.10 (2016): 29-34.
- Koeplinger, David, et al. "Spatial: a language and compiler for application accelerators." Proceedings of the 39th ACM SIGPLAN Conference on Programming Language Design and Implementation. ACM, 2018. https://spatial.stanford.edu
- KDD Dataset: https://archive.ics.uci.edu/ml/datasets/kdd+cup+1999+data