MALWARE DETECTION WITH ~~MACHINE LEARNING~~ NEURAL NETWORKS

BRETT LUSKIN

DATA 606 – CAPSTONE SPRING 2020

GITHUB: HTTPS://GITHUB.COM/BLUSK44/CAPSTONE606

RECAP

- Network Intrusion Detection System (NIDS) using UNSW-NB15 Dataset
- No Domain Knowledge
- Benchmark using existing research

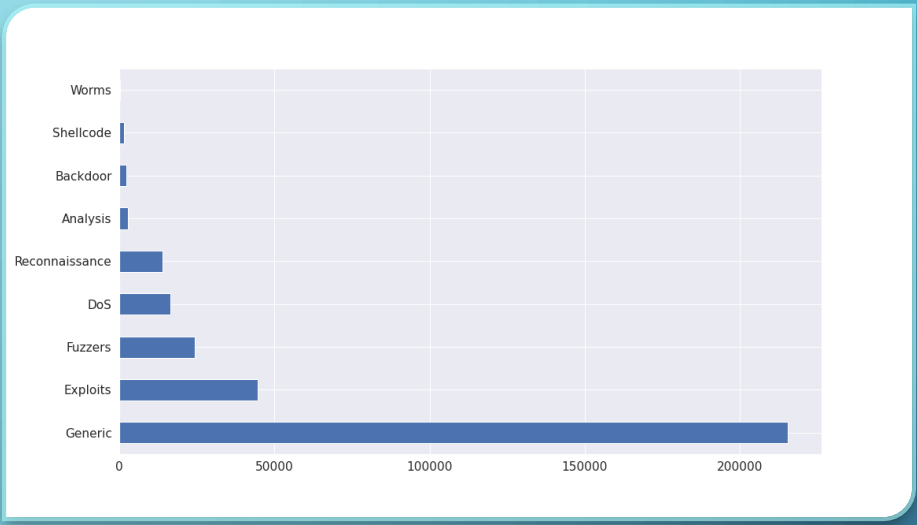
SIGNATURE VERSUS ANOMALY DETECTION

Signature detection: NIDS based on a database of existing known attacks

Anomaly detection: NIDS based on detecting unknown attacks using profile parameters

MACHINE LEARNING RESULTS

- Logistic Regression – Mediocre, linear method
- SVM (RBF kernel) – Coin flip, compute intensive
- PCA – Coin flip
- Random Forest – Strong
- ADABoost – Strong

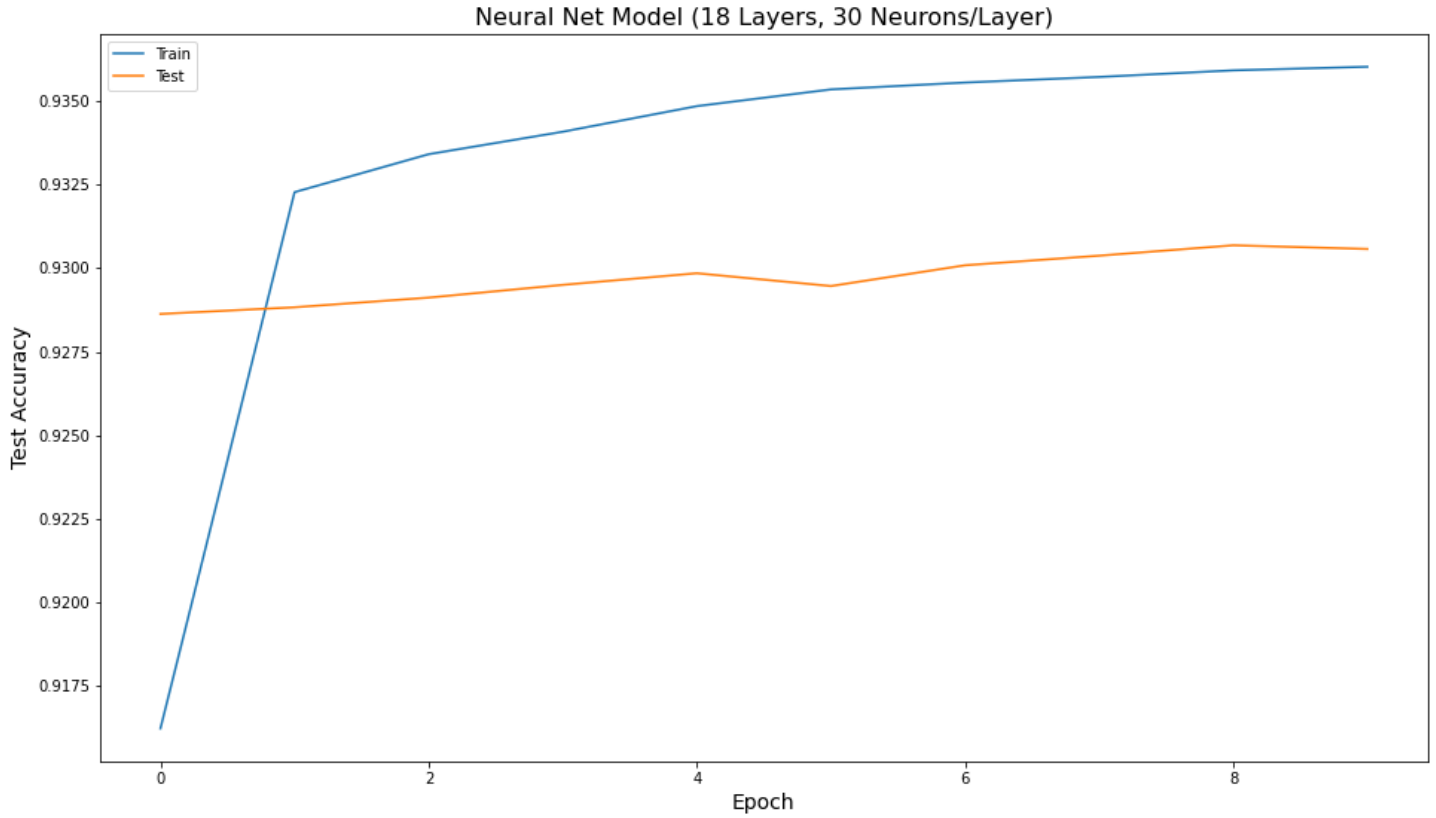


RANDOM FOREST RETROSPECTIVE

NEURAL NETWORK

- L. Zhiqiang, G. Mohi-Ud-Din, L. Bing, L. Jianchao, Z. Ye and L. Zhijun, "Modeling Network Intrusion Detection System Using Feed-Forward Neural Network Using UNSW-NB15 Dataset," 2019 IEEE 7th International Conference on Smart Energy Grid Engineering (SEGE), Oshawa, ON, Canada, 2019, pp. 299-303.
- Ten hidden layers, 10 neurons each, Stochastic Gradient Descent, 10 epochs
- Activation function? Batch normalization?

PREVIOUS MODEL



UNSW-NB15 DATASET REVISITED

ORIGINAL

- No training and test set
- 2.5 million records
- 13% attack data

PREPARED

- Training and Test set
- 250,000 records
- 68% attack data

DATA REVISITED

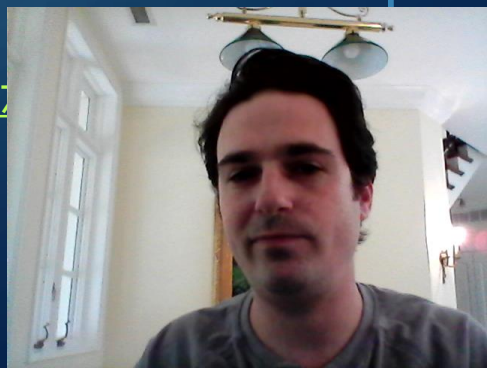
DATA QUALITY ISSUES:

- DUPLICATE LABELS
- NORMALIZATION
- CLASS IMBALANCE

NORMALIZATION

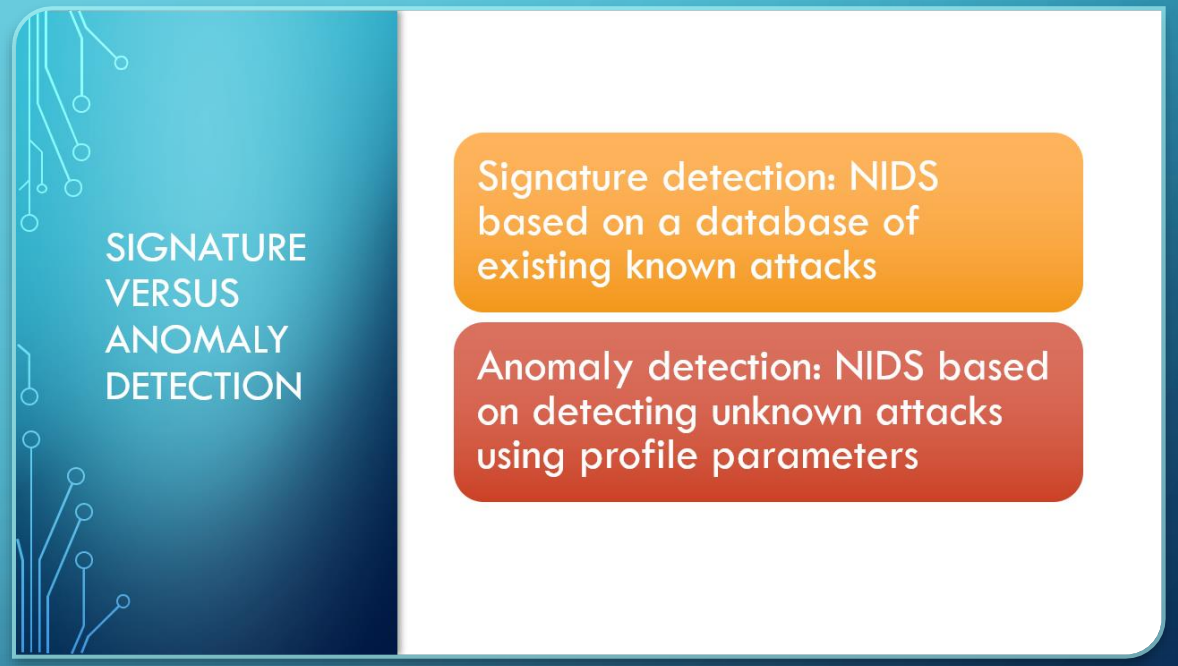
- Not necessary but can be more efficient and lead to better predictor
- Learning can be harder with all positive or all negative values
  - Normalizing with a mean around zero gives you positive and negative values
  - Allows weights to be adjusted independently during Gradient Descent
- Scale (magnitude) of inputs become the same

https://towardsdatascience.com/why-data-should-be-normalized-before-training-a-neural-network-c626b7



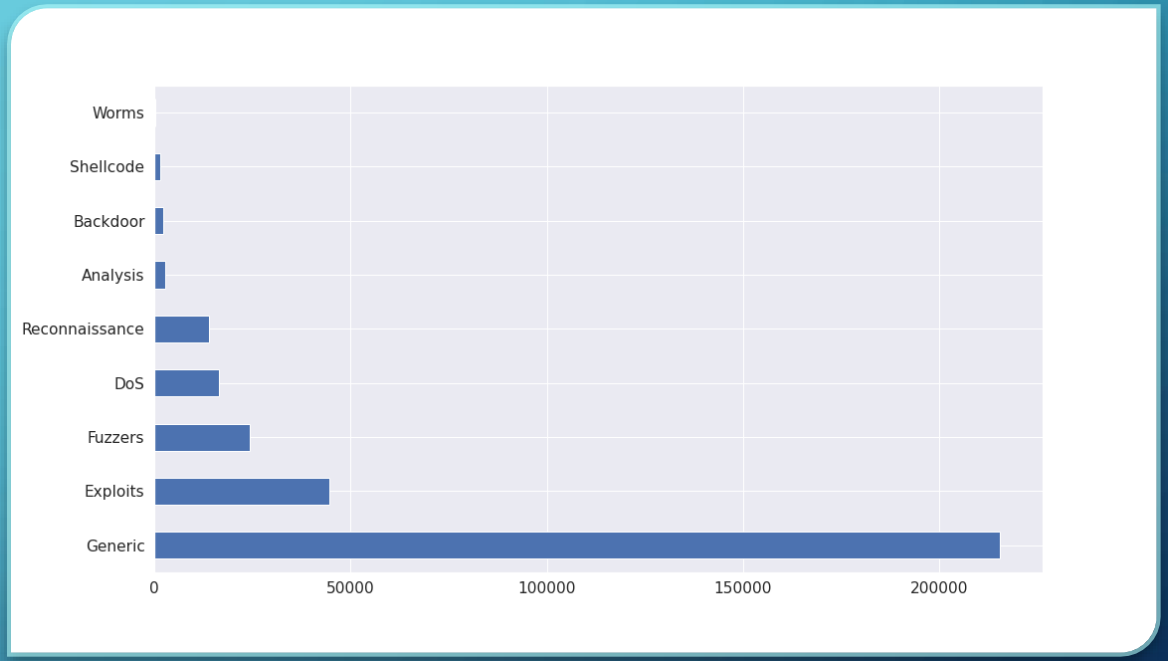
CLASS IMBALANCE

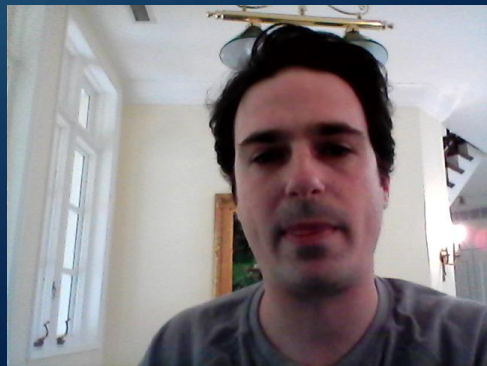
- Full dataset only has 13% of values as attacks
- Mode collapse
  - Network “learns” to predict the classes with the highest counts
- Stratified train, test, validation



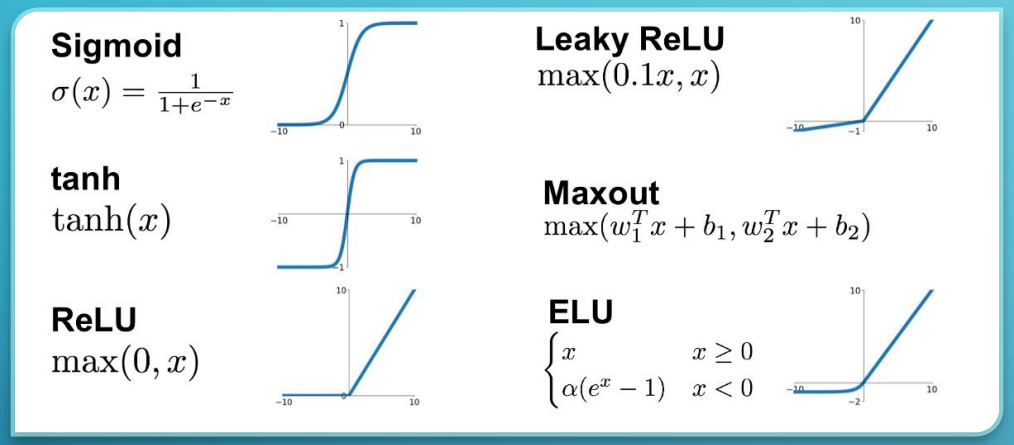
CLASS IMBALANCE

- Full dataset only has 13% of values as attacks
- Mode collapse
  - Network “learns” to predict the classes with the highest counts
- Stratified train, test, validation



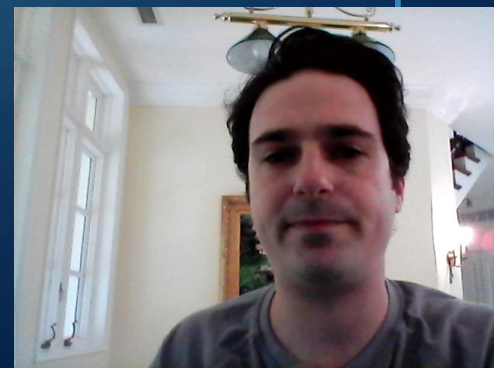


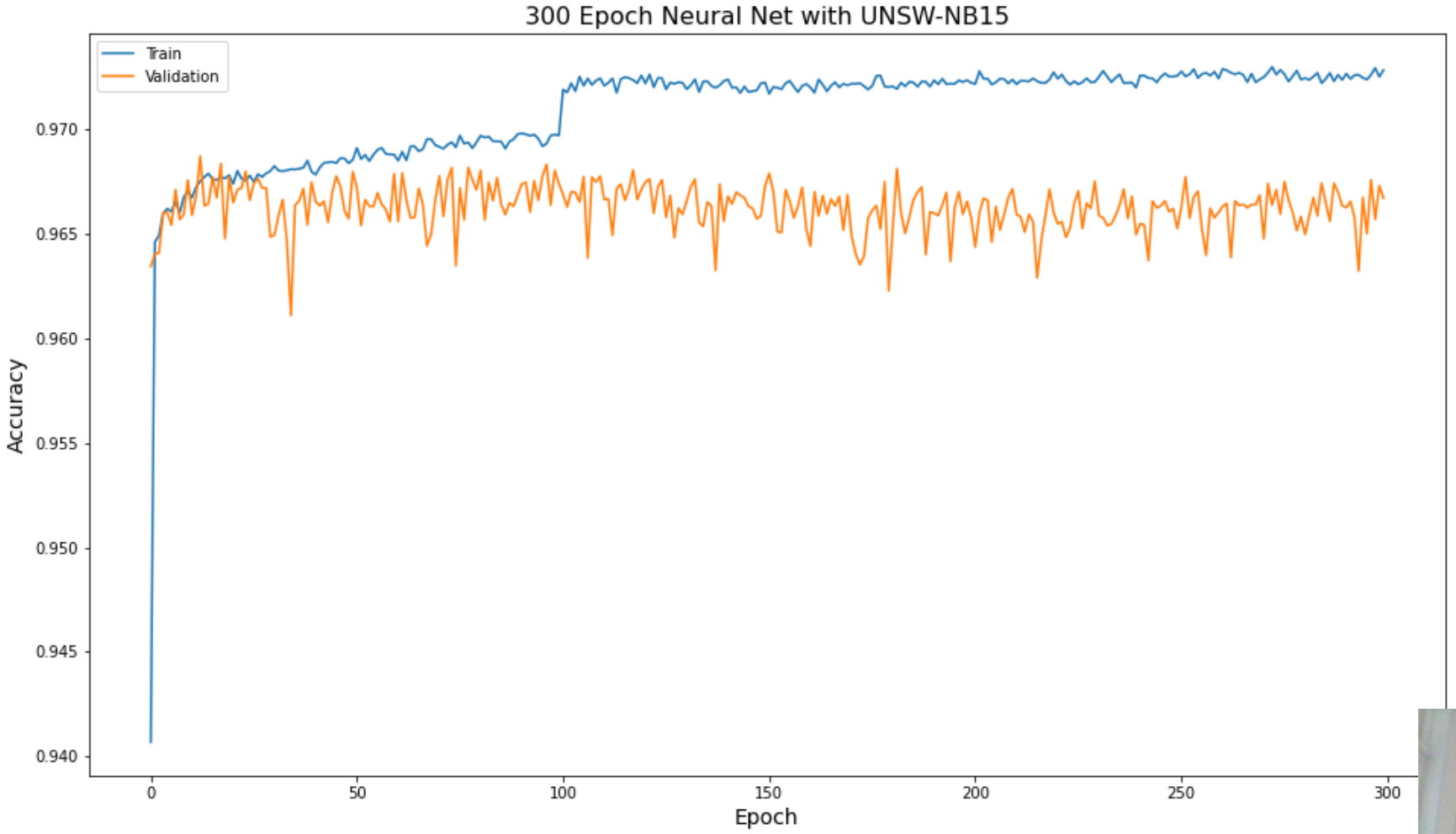
NETWORK DESIGN



CODE AVAILABLE ON
GITHUB: https://github.com/blusk44/Capstone606

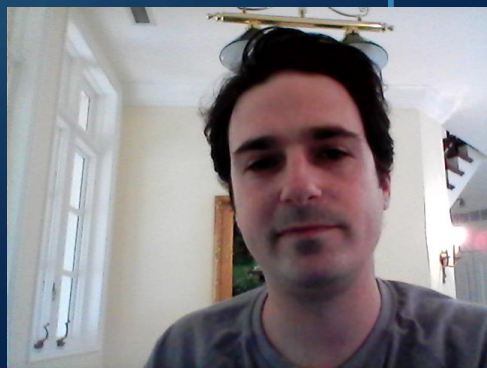
- 5 Linear layers
- Batch size 128
- Batch Normalization
- LeakyReLU activation
- 15 Neurons/layer
- AdamW optimizer




300 Epoch Neural Net with UNSW-NB15
Train
Validation
Accuracy
0.970
0.965
0.960
0.955
0.950
0.945
0.940
0
50
100
150
200
250
300
Epoch

FUTURE WORK IN UNSW-NB15

- Try to improve data quality
- Adjust model parameters
- Semi-supervised or unsupervised learning



LESSONS LEARNED

- Establish best practices
  - Long training time? Save the model in intervals
- Become an expert in boring
  - If Neural Networks are sports cars, data quality is the engine
- Working is easier for parents of young children when there is no pandemic

REFERENCES

https://towardsdatascience.com/why-data-should-be-normalized-before-training-a-neural-network-c626b7f66c7d

https://towardsdatascience.com/pytorch-tabular-multiclass-classification-9f8211a123ab

K. Hassine, A. Erbad and R. Hamila, "Important Complexity Reduction of Random Forest in Multi-Classification Problem," 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), Tangier, Morocco, 2019, pp. 226-231.

T. Janarthanan and S. Zargari, "Feature selection in UNSW-NB15 and KDDCUP'99 datasets," 2017 IEEE 26th International Symposium on Industrial Electronics (ISIE), Edinburgh, 2017, pp. 1881-1886.

D. Jing and H. Chen, "SVM Based Network Intrusion Detection for the UNSW-NB15 Dataset," 2019 IEEE 13th International Conference on ASIC (ASICON), Chongqing, China, 2019, pp. 1-4.

L. Zhiqiang, G. Mohi-Ud-Din, L. Bing, L. Jianchao, Z. Ye and L. Zhijun, "Modeling Network Intrusion Detection System Using Feed-Forward Neural Network Using UNSW-NB15 Dataset," 2019 IEEE 7th International Conference on Smart Energy Grid Engineering (SEGE), Oshawa, ON, Canada, 2019, pp. 299-303.

UNSW-NB15 Dataset: https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/

My github: https://github.com/blusk44/Capstone606/tree/master/Delivery-3