@PAD: ADVERSARIAL TRAINING OF POWER SYSTEMS AGAINST DENIAL OF SERVICE ATTACKS

Ali I Ozdagli, Carlos Barreto, and Xenofon Koutsoukos

Department of Electrical Engineering and Computer Science

Vanderbilt University

HotSoS 2020 Virtual Symposium: Sept. 22-24, 2020

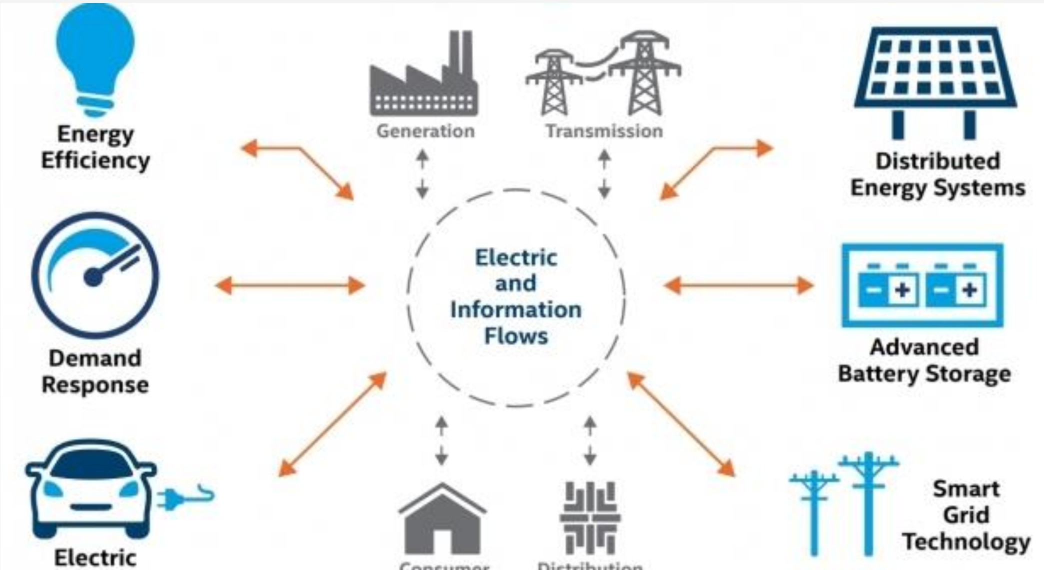
Session 1

LAYOUT

- Introduction
- Methodology
- Evaluation
- Conclusion and Future Research
- Acknowledgements

INTRODUCTION

- Smart Energy Grids (SEG) to become essential by 2030
- Control, monitoring, and telecommunication networks.
- Power systems: Previously isolated, currently accessible to general public.
- **Open to cyber-physical threats**



Taken from iotone.com

MOTIVATION

- Quality requirements for power systems
  - Monitoring and analysis of disturbances and faults
- Difficulty of human recognition for abnormal events for large systems
- Exploration of machine learning (ML) for discriminating power system disturbances [1]
- Failure of ML for discrimination in high-dimensional inputs

[1] Hink, Raymond C. Borges, Justin M. Beaver, Mark A. Buckner, Tommy Morris, Uttam Adhikari, and Shengyi Pan. "Machine learning for power system disturbance and cyber-attack discrimination." In *2014 7th International symposium on resilient control systems (ISRCS)*, pp. 1-8. IEEE, 2014.

HYPOTHESIS & OBJECTIVE

- Denial-of-Service Attacks
  - Attack on sensors (features)
  - Delay of data ➔ *Deletion* of feature

- Hypothesis

  Deletion of targeted features may cause misclassification [2]

- Objective

  i) Development of a DoS attack model to deceive neural network (NN) classifiers

  ii) Development a defense model against such DoS attacks

[2] Globerson, Amir, Choon-Hui Teo, Alexander Smola, and Sam Roweis. "An adversarial view of covariate shift and a minimax approach." In *Dataset shift in machine learning*. MIT Press, 2009.

ASSUMPTIONS

- White-box attack: Access to the control system/sensor readings
- Adversary resources: attack on limited number of sensors
- RELU activated neural network
- Guided adversary: attack on *abnormal* events
- Neither data nor attack is time-correlated

METHODOLOGY
ATTACK MODEL

- Find features to delete to maximize prediction error

$$\alpha_i^{\max} = \arg\max \ [1 - y_i \ \hat{y}_i]_+$$
$$= \arg\max \ [1 - y_i \ F(x_i \circ (1 - \alpha_i))]_+$$
$$s.t. \quad \alpha_i \in \{0, 1\}^d$$
$$\sum_{j=1}^{d} \alpha_{ij} \leq K$$

If the adversary does not cause any misprediction, then the error is zero

$F(x)$: discriminator neural network

$x_i \in \mathbb{R}^d$: input $\quad$ $y_i \in \{-1, +1\}$: true label $\quad$ $\hat{y}_i \in \{-1, +1\}$: predicted label

$\alpha_i = [\alpha_{i1}, \cdots, \alpha_{ij}]$: features to be deleted $\quad$ K: attacker budget

SOLVING FOR ATTACK MODEL

- For linear classifiers, the optimization problem presented is a convex mixed-integer LP (MILP)
  - NP-Hard, solved heuristically

$$\begin{aligned} \alpha_i^{\max} &= \arg\max \quad [1 - y_i\ \hat{y}_i]_+ \\ &= \arg\max \quad [1 - y_i\ F(x_i \circ (1 - \alpha_i))]_+ \end{aligned}$$

- For NN with RELU activation, the solution space is not convex MILP
  - Still solvable by computationally exhaustive nonlinear programming (NILP) approaches
- Relaxation: NN with RELU holds piece-wise linearity characteristics
  - Reconstruction of NN as a set of logic formulas
  - Utilization of Disjunctive Normal Form (DNF) [3]
  - NN can be written as a MILP using DNF

[3] Katz, Guy, Clark Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer. "Reluplex: An efficient SMT solver for verifying deep neural networks." In International Conference on Computer Aided Verification, pp. 97-117. Springer, Cham, 2017.

DNF RELAXATION

- Example for single layer NN:

NN

$$\hat{y} = \text{RELU}(w\ x) \\ = max(0, w\ x)$$

DNF

$$(\hat{y} == w\ x\ \wedge\ y > 0) \\ \vee(\hat{y} == 0\ \wedge\ w\ x \leq 0)$$

$$\alpha_{i,1} = \arg\max\ [1 - y_i\ \hat{y}_i]_+$$

$$s.t.\quad \alpha_i \in \{0,1\}^d$$

$$\sum\nolimits_{j=1}^{d} \alpha_{ij} \leq K$$

$$\hat{y}_i == w\ x_i \circ (1 - \alpha_i)$$

$$\hat{y}_i > 0$$

MILP for the first DNF

- For all clauses:

$$\alpha_i^{\max} = \underset{\alpha_{i,1},\ldots,\alpha_{i,k}}{\arg\max}\ [1 - y_i\ F(x_i \circ (1 - \alpha_i))]_+$$

Ideal Optimal Solution

- Limitation: 2^k clauses for k neurons
- Further relaxation: No need to maximize error among all clauses
  - We only need one clause that will cause mislabeling

HOTSOS 2020

FINAL ATTACK MODEL

Input: (x_i, y_i), w, $F(x)$

Output: α_i

Generate DNF clauses for the given weights of the network

foreach *DNF clause set* **do**

- Assign clause components as constraints to Equation 2
- Solve Equation 2 with new constraints
- **if** *Problem is infeasible* **then**
  - **continue** with the next clause set
- **else**
  - Obtain α_i
- Predict the label $\rightarrow \hat{y}_i = F(x_i \circ (1 - \alpha_i))$
- **if** $\hat{y}_i$ == *normal* **then**
  - `/* there is a successfully attack! */`
  - **continue** with the next input (x_{i+1}, y_{i+1})

if $\hat{y}_i$ == *normal for all DNF clause sets* **then**

- `/* there is no successfully attack! */`
- $\alpha_i = 0$

continue with the next input (x_{i+1}, y_{i+1})

- Worse-case scenario:
  - Go through all clauses
  - Find no solution
  - $O(2^k)$ vs $O(K(d-K)!)$
- Further relaxation:
  - Limit number of clauses

$$
\begin{aligned}
\alpha_i^{\max} &= \arg\max \ [1 - y_i\ \hat{y}_i]_+ \\
&= \arg\max \ [1 - y_i\ F(x_i \circ (1 - \alpha_i))]_+ \\
s.t.\quad & \alpha_i \in \{0, 1\}^d \\
& \sum_{j=1}^{d} \alpha_{ij} \le K
\end{aligned}
$$

Eq. 2

VANDERBILT UNIVERSITY

METHODOLOGY
DEFENSE MODEL

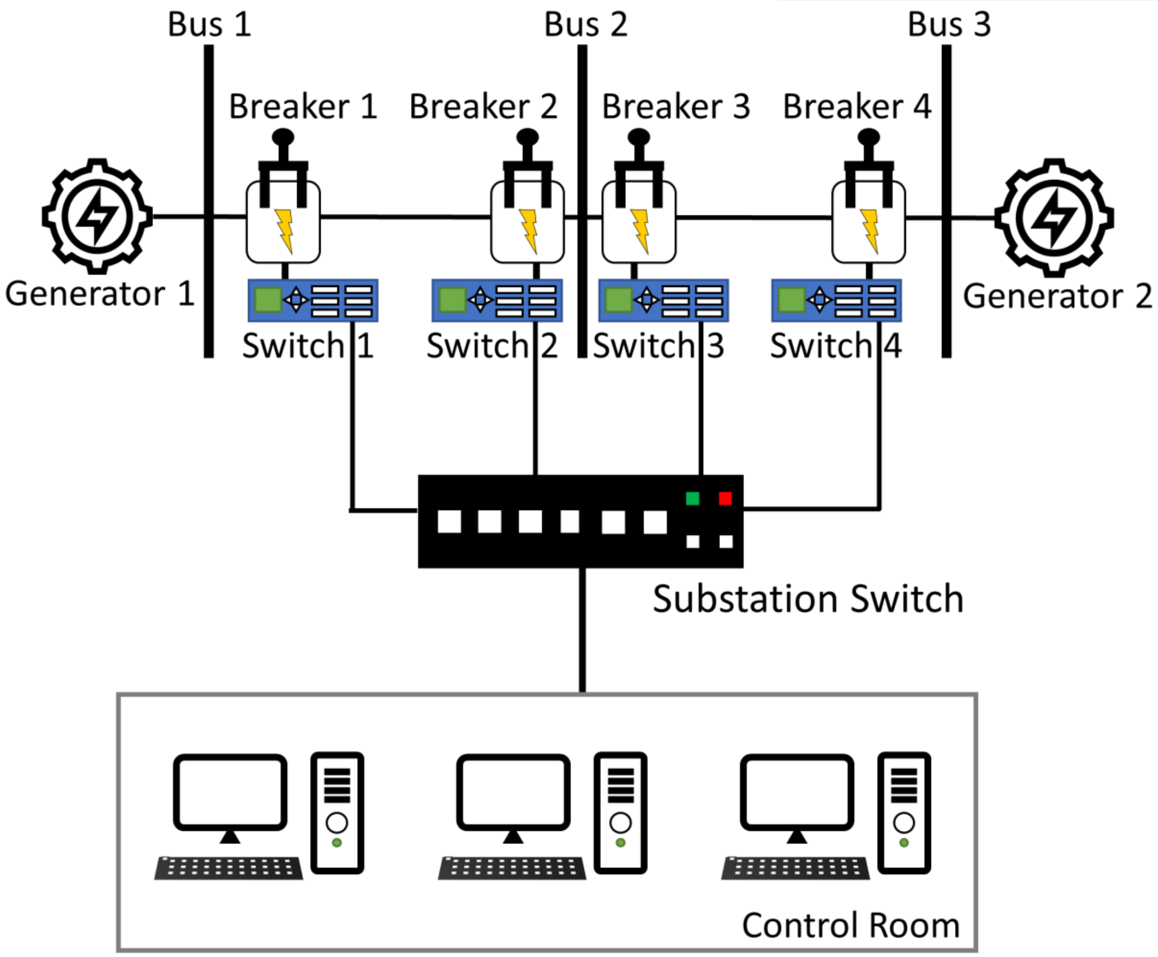
- MiniMax Problem
  - Minimization of average maximum prediction error over the entire dataset

$$\min_{w} \max_{\alpha_1,\ldots,\alpha_n} \frac{1}{n} \sum_{i=1}^{n} [1 - y_i \ F(x_i \circ (1 - \alpha_i))]_+$$

- One-shot training strategy [4]:
  - Train baseline NN with a dataset
  - Generate adversarial example dataset using baseline
  - Train a new NN with adversarial example dataset

[4] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).

EVALUATION



- Two categories
  - Normal event
  - Abnormal events
- 128 features
- ~4000 events for training
- ~1000 events for testing
- Ratio of normal events to abnormal events is ~28%

Hink, Raymond C. Borges, Justin M. Beaver, Mark A. Buckner, Tommy Morris, Uttam Adhikari, and Shengyi Pan. "Machine learning for power system disturbance and cyber-attack discrimination." In *2014 7th International symposium on resilient control systems (ISRCS)*, pp. 1-8. IEEE, 2014.

EFFECTIVENESS OF ATTACK

- Baseline NN model
  - Single hidden layer (5 neurons)
  - RELU for hidden layers
- Number of clauses, $2^5 = 32$
  - Clause modeled with CVXPY and Gurobi
- Attack model
  - Budget ($K = \{1,3,6\}$) corresponding to $\{1\%, 2.5\%, 5\%\}$ of all features

Dataset	Accuracy in Percentage		
Original Training Dataset	87.47		
Original Testing Dataset	83.23		
	K = 1	K = 3	K = 6
Modified Testing Dataset	31.08	16.29	12.77

EFFECTIVENESS OF DEFENSE

- Adversarial data generation with budget ($K = \{1,3,6\}$) for training
- Generalization over original training and testing data
- Attack on the defense model

Dataset	Accuracy in Percentage K = 1	K = 3	K = 6
Adversarial Training Dataset	86.12	86.70	88.06
Original Training Dataset	85.14	85.32	86.58
Original Testing Dataset	81.89	82.69	80.78
Modified Testing Dataset	39.23	26.05	19.51

Baseline Model: 31.08 16.29 12.77

Some improvements

SUMMARY & CONCLUSION

- DoS attack model is very powerful
  - Faults and attacks could be obscured
- NN with RELU can be modeled as piece-wise MILP
  - Features-to-delete can be found effectively
- Minimax approach as a defense mechanism
  - One-shot training improves the robustness against attacks to some degree

FUTURE RESEARCH

- More reliable defense models
- Multiple categories
- Black-box models
- MILP for more complex networks (convolutional)

ACKNOWLEDGEMENT

- Science of Security Program
- IBM Graduate Fellowship

HOTSOS 2020

THANK YOU!

For follow-up questions: Ali.I.Ozdagli@vanderbilt.edu

Xenofon.Koutsoukos@vanderbilt.edu

VANDERBILT UNIVERSITY