Steerable Deep Transfer Learning for Critical Infrastructure Attack Detection

Kayla Cummings† | **Mentor: Jason Laska**

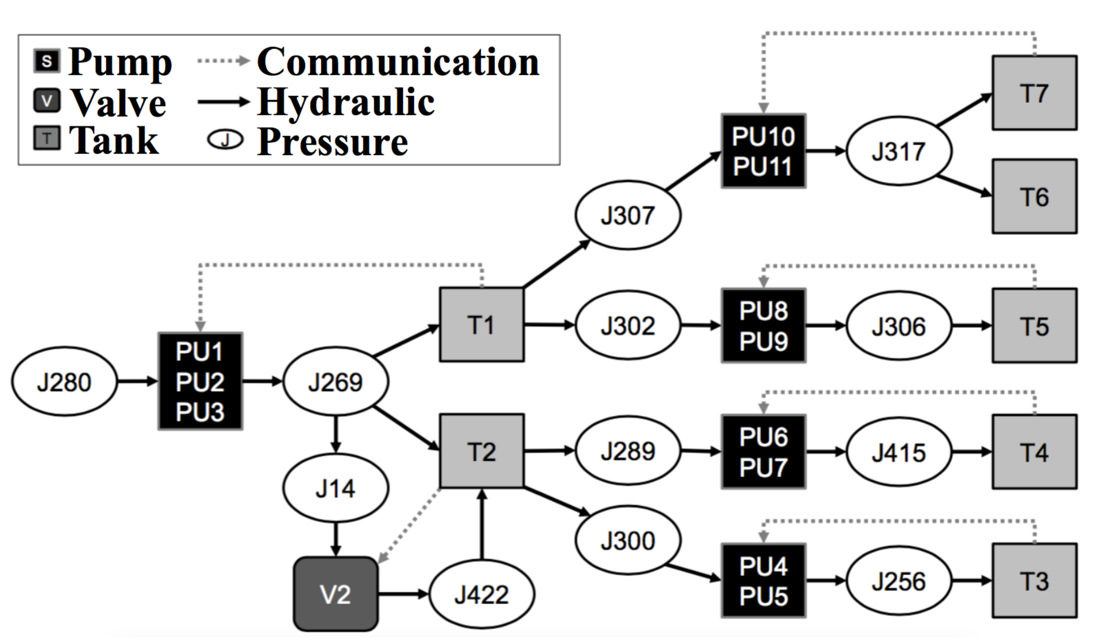
†kaylac@mit.edu

Problem

- **Motivation:** Computers regulate modern water distribution, increasing likelihood of cyber attacks [1].
- **Approach:** We use deep learning to flag cyber attacks in sensor data from water distribution systems.
- **Competition:** Recent literature favors unsupervised deep learning, which characterizes attacks as anomalous occurrences in clean sensor data.
- **Contribution:** We contribute a *steerable* transfer learning model that trains a deep neural network on customizable synthetic attacks.
- **Justification:** Our model gives the driver more agency in the attack detection process.
- **Benefit:** Success of our model and transition to deployment informs safe water distribution.

Data

- Raw data are 8761 hourly, simulated readings from 43 sensors in a real water distribution system [2].
- We remove 7 negligible-variance features, scale each feature within $[0, 1]$, and roll the data into windows.
- Hierarchy of water distribution system's hydraulic components shown below [3].

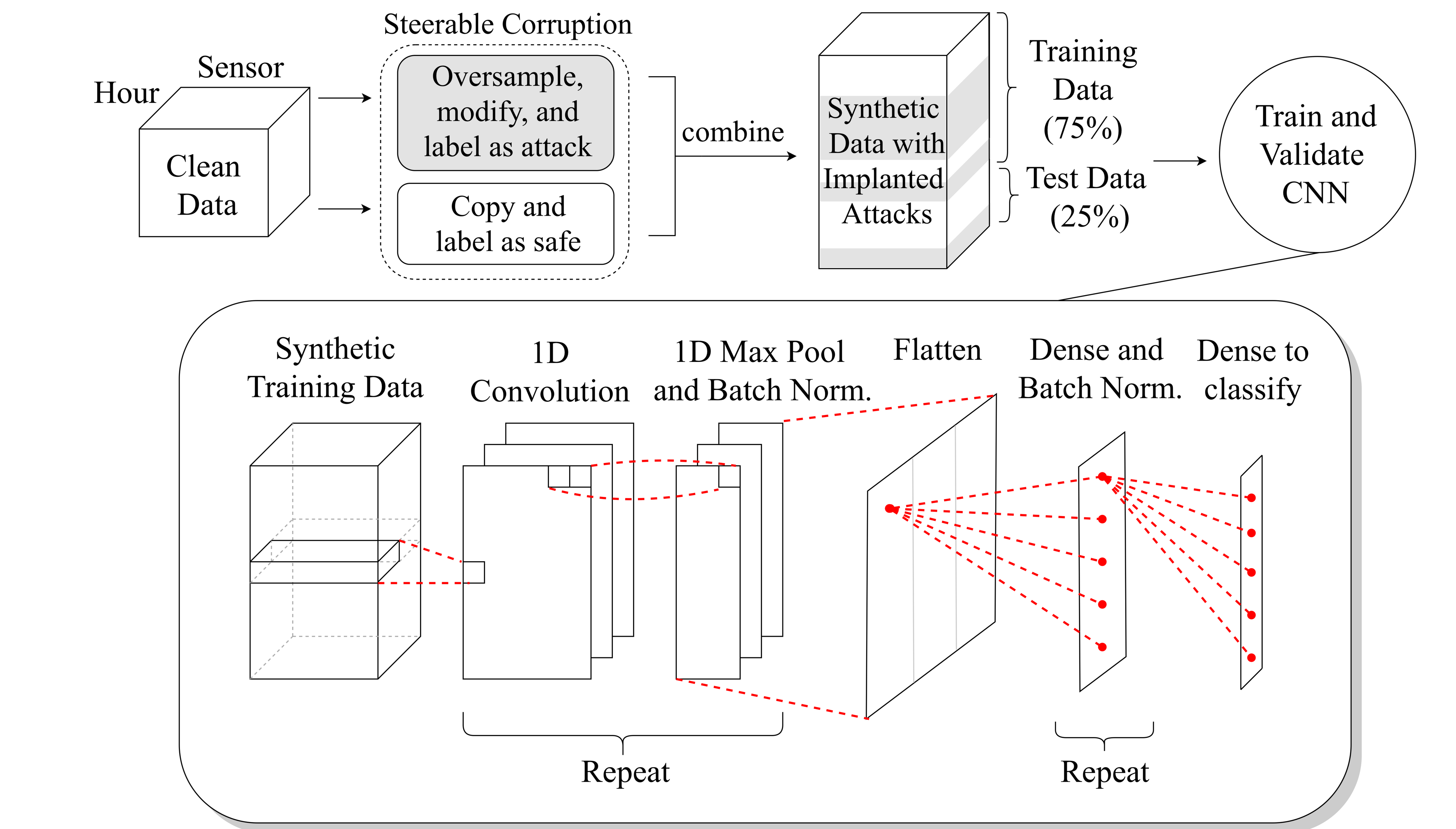


Attack data. We generate synthetic data as follows:

- Oversample clean windows and modify each observation with one attack by swapping random contiguous sub-ranges of 2 sensors' readings.
- Include one unmodified copy of pre-processed data for examples of safe observations.

Model

- We coded the convolutional neural network (CNN) in SciPy [4], Keras [5], and TensorFlow [6].
- Convolutional layers learn temporal patterns and dense layers recognize global relationships.
- Other layers regularize parameters, prevent over-fitting, and reduce computational intensiveness.



Results

Domain-agnostic model detects synthetic attacks?

Predicted Class \ True Class	ATTACK	SAFE
ATTACK	0.99	0.11
SAFE	0.01	0.89

Domain-agnostic synthetic test set

Domain-agnostic model detects real attacks?

Predicted Class \ True Class	ATTACK	SAFE
ATTACK	0.97	0.79
SAFE	0.03	0.21

Realistic test set [2]

Domain-specific model detects real attacks?

Predicted Class \ True Class	ATTACK	SAFE
ATTACK	0.94	0.36
SAFE	0.06	0.64

Realistic test set [2]

Iterative model detects real attacks?

Predicted Class \ True Class	ATTACK	SAFE
ATTACK	0.39	0.03
SAFE	0.61	0.97

Realistic test set [2]

- **Best model:** 3 convolutional layers with 10 filters of length 5 and 2 dense layers of dimension 30.
- Networks with highest test accuracy had many convolutional layers with few filters (144 tests).
- Model classifies synthetic attacks with 95% accuracy and realistic attacks [2] with high recall.
- Domain-specific ablation: swapping only among like sensors reduces false positives.
- By training on domain-specific synthetic data, and then a realistic training set [2], false positives diminish to negligible levels at expense of false negatives. **Results are more accurate than training on realistic training set alone** (confusion matrix not shown).

Conclusions

- Our model accurately classifies implanted attacks.
- Domain-agnostic and domain-specific models have high recall.
- False positives drastically decrease with only one simple domain-specific assumption.
- Iterative training with synthetic and realistic attack data reduces overfitting and suppresses alarms at the expense of false negatives.
- Naïve steering has potential to facilitate customizable attack detection.

Next steps.

- Perform more extensive ablation study.
- Characterize synthetic attacks that steer model toward detecting specific real attacks.
- Replicate results with datasets from other domains.

Acknowledgements

Many thanks to Jason Laska for his valued research guidance and to Robert Bridges for poster feedback. This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internship program. Managed by UT Battelle, LLC under Contract No. DE-AC05-00OR22725 for the U.S. Department of Energy. This material is based upon work supported in part by the National Science Foundation Graduate Research Fellowship under Grant No. 1122374. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

[1] Hill, M. "Water Treatment Plant Hit by Cyber-attack." *Infosecurity Magazine,* March 2016.

[2] Taormina, R. *et al.* The Battle Of The Attack Detection Algorithms: Disclosing Cyber Attacks on Water Distribution Networks. *Jnl. of Water Resources Planning and Mgmt.*, 2018.

[3] Chandy, S. *et al.*, Cyberattack detection using deep generative models with variational inference, *Jnl. Water Resources Planning and Mgmt.*, 2018.

[4] Jones E. *et al.* SciPy: Open Source Scientific Tools for Python, 2001. `https://www.scipy.org/`.

[5] Chollet, F. *et al.* Keras, 2015. `https://keras.io`

[6] Abadi, M. *et al.* TensorFlow: Large-scale ML on heterogeneous systems, 2015. `https://tensorflow.org`

[7] Ioffe, S. and C.Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, CoRR, vol. abs/1502.03167, 2015.

Steerable Deep Learning for Critical Infrastructure Attack Detection

Kayla S. Cummings (Massachusetts Institute of Technology, Cambridge, MA 02139)

Jason A. Laska (Oak Ridge National Laboratory, Oak Ridge, TN 37831)

Central control systems automatically regulate water distribution networks based on hydraulic sensor readings. Adversaries can hack this emergent cyber layer to alter readings and disrupt safe water distribution. We furnish a flexible model that uses transfer learning and customizable data augmentation to detect attacks. We augment clean sensor data with domain-agnostic "attacks", with which we train a 1D-convolutional neural network. Our model detects real attacks with high recall and a high false positive rate. By making one simple domain-specific assumption on synthetic attacks, we drastically reduce false positives by more than half with minimal sacrifice. When training on real attacks, a warm start on domain-specific synthetic attacks yields higher accuracy than no warm start. Our work supports naïve steering as a new attack detection paradigm. Success and transition to deployment would inform national water infrastructure security.