Recurrent Neural Network Language Models for Open Vocabulary Event-Level Cyber Anomaly Detection

Aaron Tuor[1], Ryan Baerwolf[2], Nicolas Knowles[2],
Brian Hutchinson[1,2], Nicole Nichols[1] and Robert Jasper[1]

[1]Pacific Northwest National Laboratory, Richland, WA
[2]Western Washington University, Bellingham, WA

Overview

Motivation: It is impossible to manually inspect all behavior on a network. Large organizations face many internal and external threats that can be mitigated or avoided entirely if immediately detected.

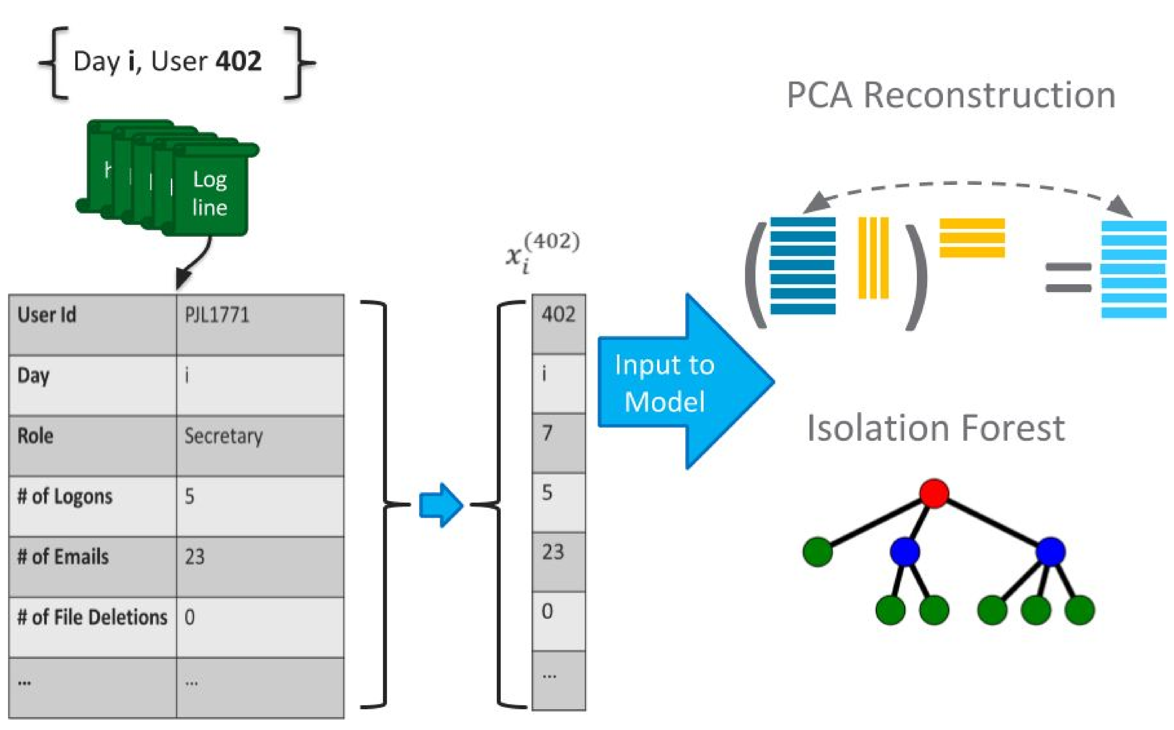
Goal: To effectively detect events of interest on a computer network and provide insight to human analysts.

Approach: Train RNN language models on log line sequences and flag improbable sequences as anomalous.

Background

Signature-based detection characterizes known attacks. Limited ability to address novel attacks.
Anomaly-based detection maintains a statistical baseline for normal behavior and flags events that deviate from the norm. Can have higher false positive rate.

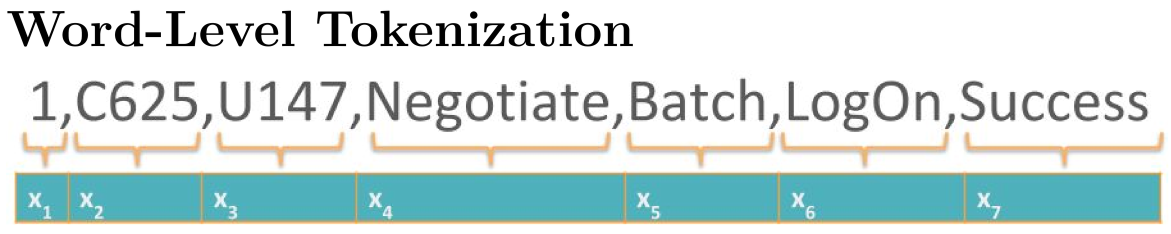
Typically, user statistics over a window of events are aggregated into a feature vector and then standard anomaly detection techniques are applied.



Approach

- Treat an individual event as a sequence of tokens.
- At each step in the sequence, predict the next token.
- The event's anomaly score is the negative log probability.
- Update the weights using cross entropy loss.

Word-Level Tokenization



Character-Level Tokenization

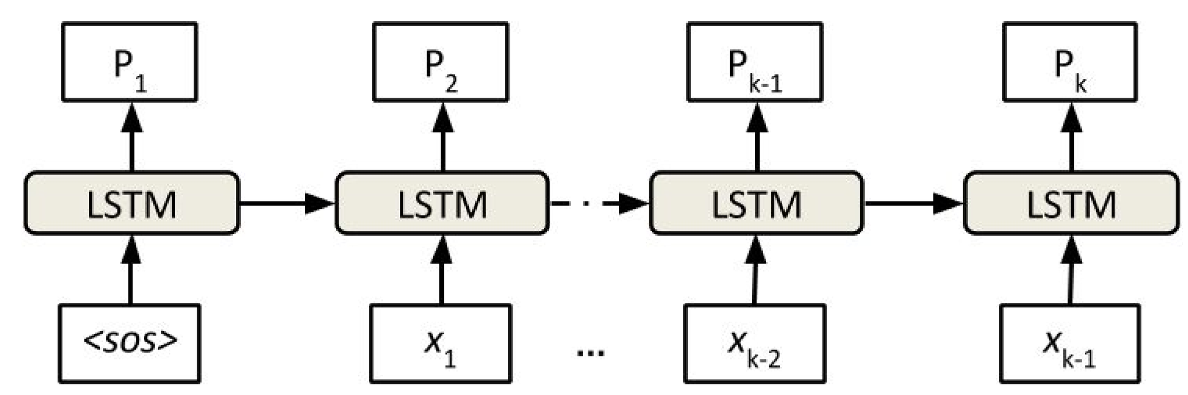
Variable length sequences - use recurrent neural networks.

System Overview

Models

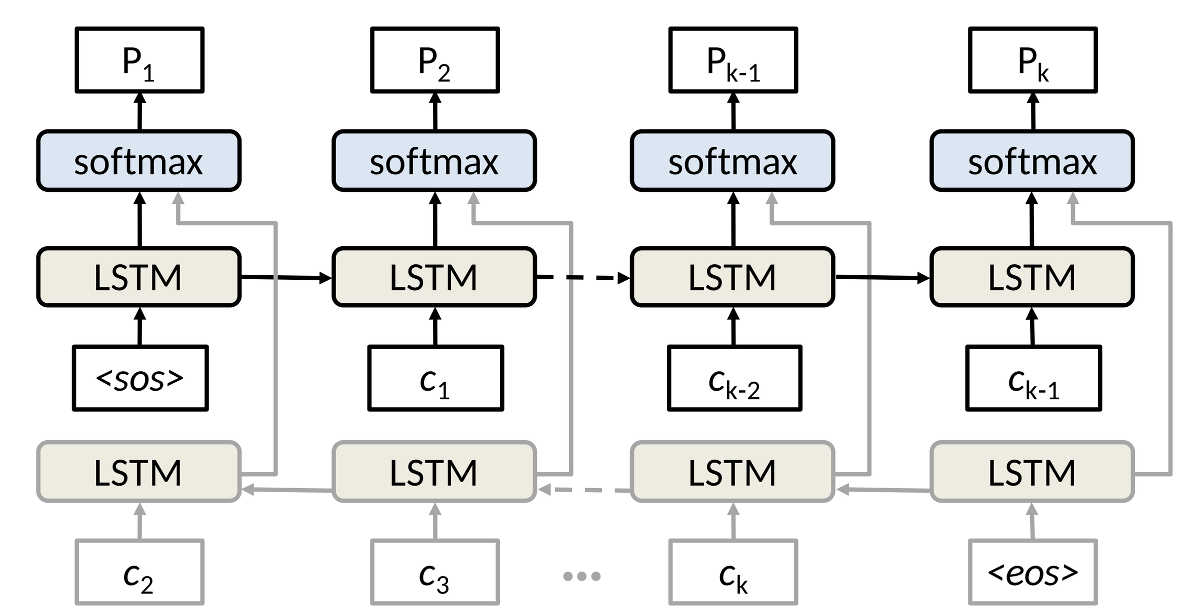
Anomaly score: $-\Sigma_i^k \log p_i$

Event Model (EM)



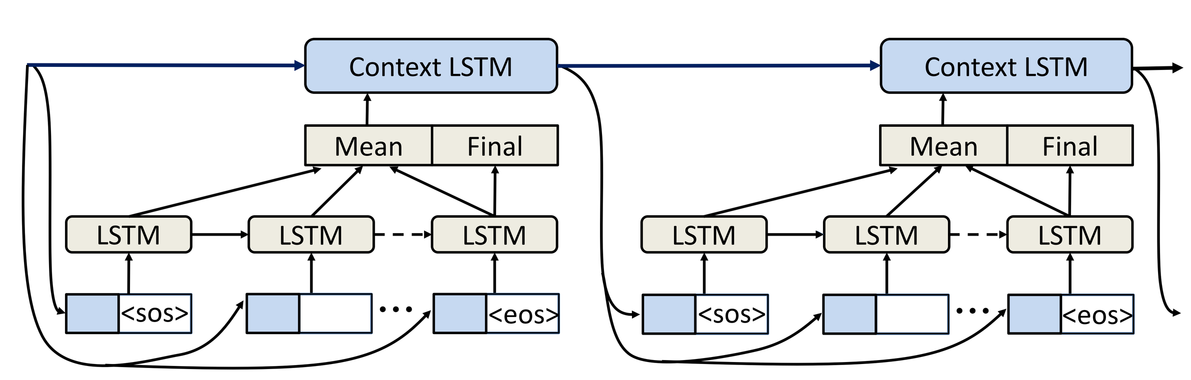
- Standard long short-term memory RNN
- $p_i = P(x_i | x_1, \ldots, x_{i-1})$

Bidirectional Event Model (BEM)



- Bidirectional LSTM RNN
- $p_i = P(x_i | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$

Tiered Event Model (TEM)



- Lower tier is EM or BEM.
- Upper tier model dynamics over user sessions.
- Context is provided to each lower level model.
- Hidden states of lower tier are averaged, concatenated to final cell state, fed into upper tier.

Experimental Setup

Metrics

Area under ROC curve.

Baselines

- Isolation Forest
- PCA

Data

- Los Alamos National Laboratory dataset
- Real world de-identified data from the LANL network.
- Over 1,051,430,459 events (749 red team events)
- Train and dev : days 1-12
- Test : days 13-30

Training

- Red team labels are used to evaluate performance only. Training is done unsupervised for all models.
- Tuned with random search on train/dev.

Results and Analysis

Day-Level Detection

Model	Tokenization	AUC
PCA	------------------	0.754
Isolation Forest	-----------------	0.763
EM	Word	0.794
BEM	Word	0.811
T-EM	Word	0.803
T-BEM	Word	0.838
EM	Character	0.754
BEM	Character	0.846
T-EM	Character	0.809
T-BEM	Character	**0.854**

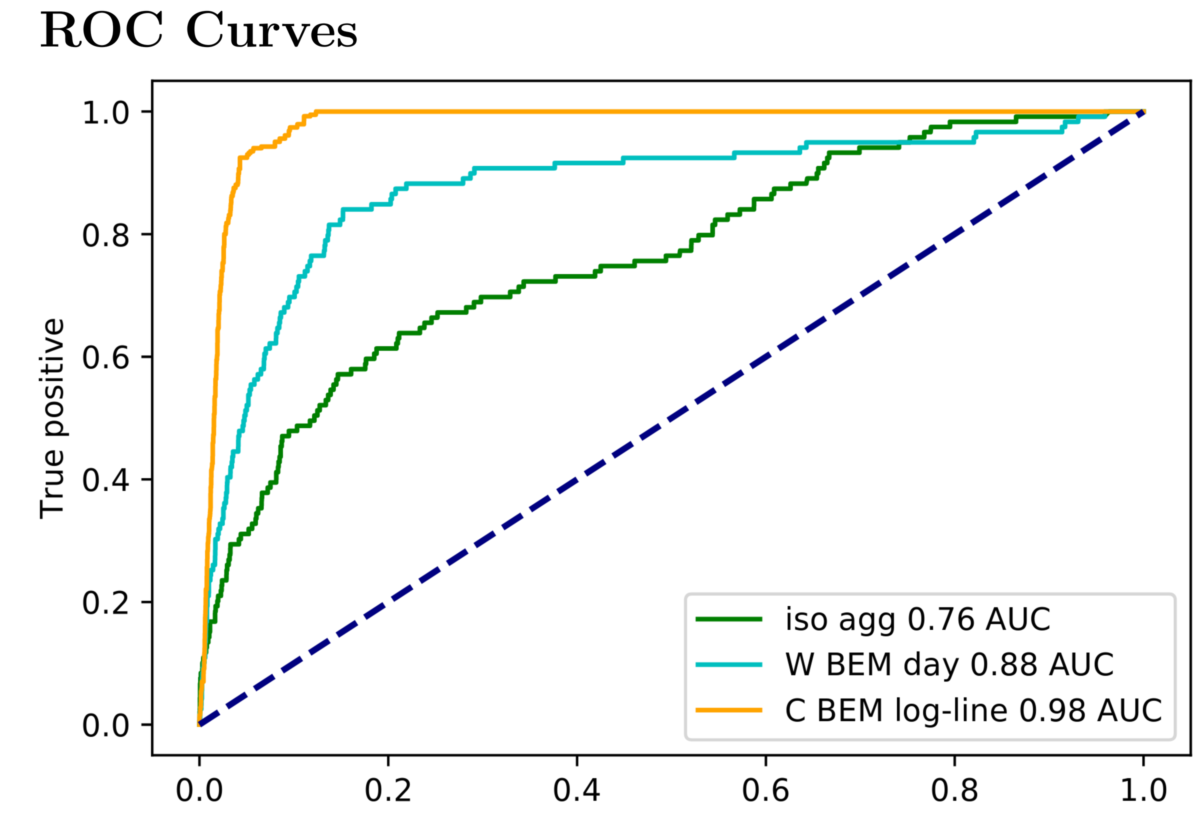
- Language models outperform baselines.
- Bidirectional and tiered modeling improve performance.

Event-Level Detection

Model	Word	Character
EM	0.932	0.935
BEM	0.895	**0.979**
T-EM	**0.948**	0.927
T-BEM	0.902	0.969

- Significantly better performance than day-level.
- Tiered not always needed (sufficient context within line).
- Best results overall with character BEM.

ROC Curves



Conclusion and Future Work

Event-level modeling

- gives better performance,
- avoids the need for feature engineering,
- is agnostic to log format.

Future Work

- Multi-source data streams.
- Interpretability
- Robustness to data poisoning