Unsupervised Cross-system Log Anomaly Detection via Domain Adaptation

Xiao Han and Shuhan Yuan

Utah State University

Introduction

LogTAD

Experiment and Analysis

Conclusion

What is Log Anomaly Detection

- System logs are widely used on online services to record the status of the system.
- Anomalous logs can be useful in maintaining and increasing reliability and stability.
- Log anomaly detection is aimed to detect a point of the anomalous event or an abnormal pattern of multi-status.

Why do we need LogTAD

- Data Imbalance
- Scarcity of Samples from a Newly Deployed System

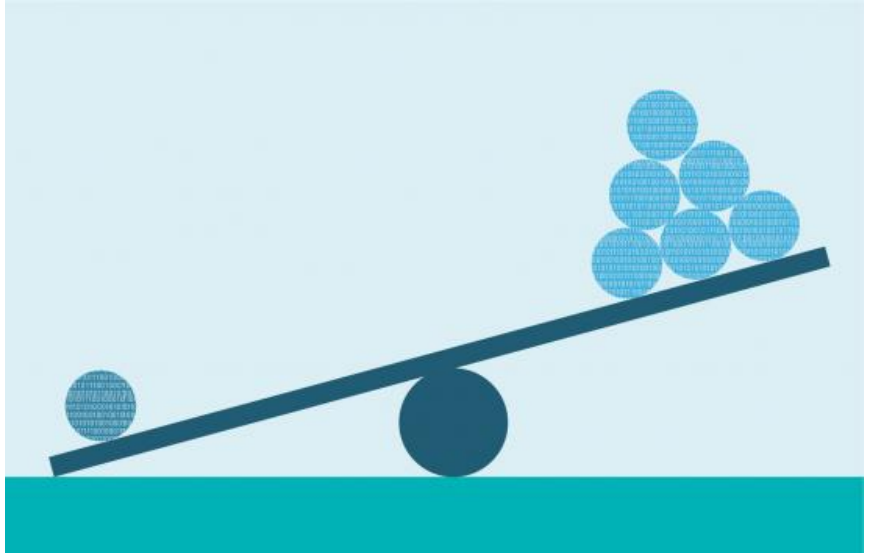


Image from: Predicting the improbable, Part 1: The imbalanced data problem. Datascience.aero. (2021, January 7). Retrieved September 16, 2021, from https://datascience.aero/predicting-improbable-part-1-imbalanced-data-problem/.

Introduction

LogTAD

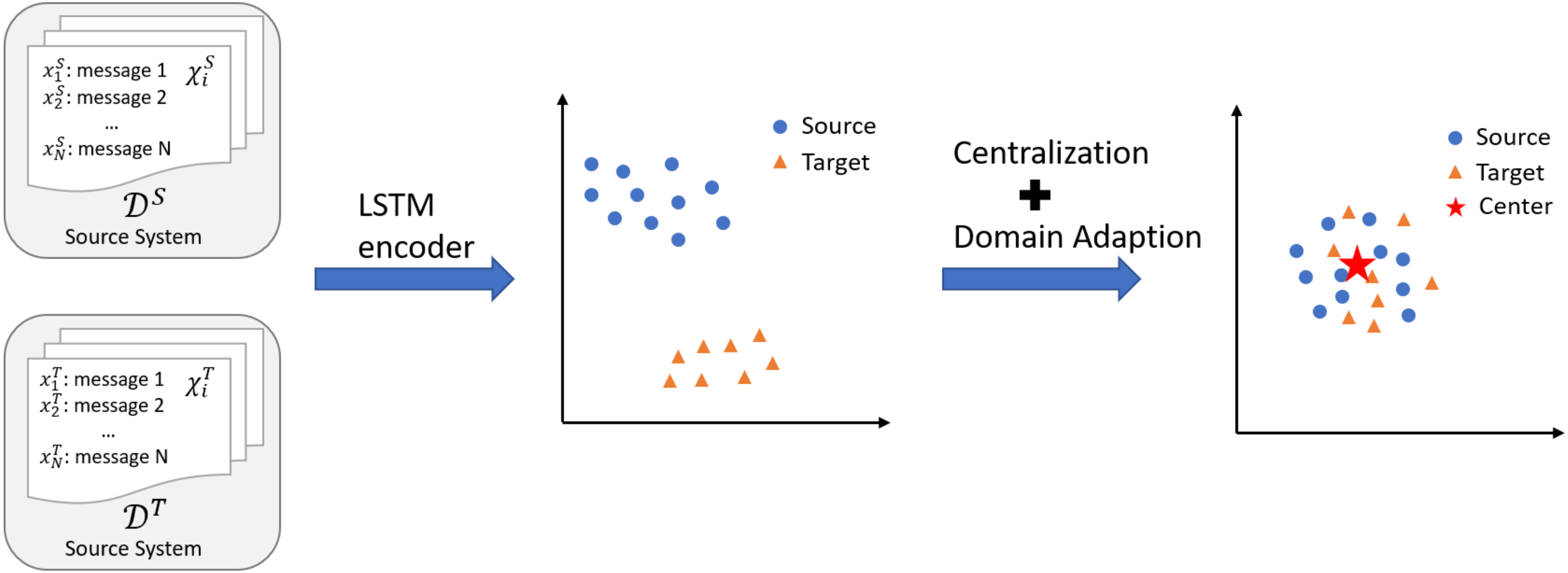
Experiment and Analysis

Conclusion

Problem Statement

- Prerequisites: A dataset $\mathcal{D}$ consisting of normal log sequences from the source system $\mathcal{D}^S$ and a small number of normal log sequences from the target system $\mathcal{D}^T$.
- Goal: Building an unsupervised and transferable log anomaly detection model to detect the anomalous log sequences from both source and target system.

Workflow of LogTAD



Log Sequence Centralization

- Encodes the log messages in a sequence to a sequence representation,

$$\mathbf{h}_n = LSTM(\mathbf{x}_n, \mathbf{h}_{n-1}),$$
$$\mathbf{v} = \mathbf{h}_N.$$

- Inspired by the DeepSVDD that the normal log sequences should be in a hypersphere and close to the center in the embedding space,

$$\mathbf{c} = Mean(\mathbf{v}_i^{\epsilon}), \text{ where } \epsilon \in \{\mathrm{S}, \mathrm{T}\}.$$

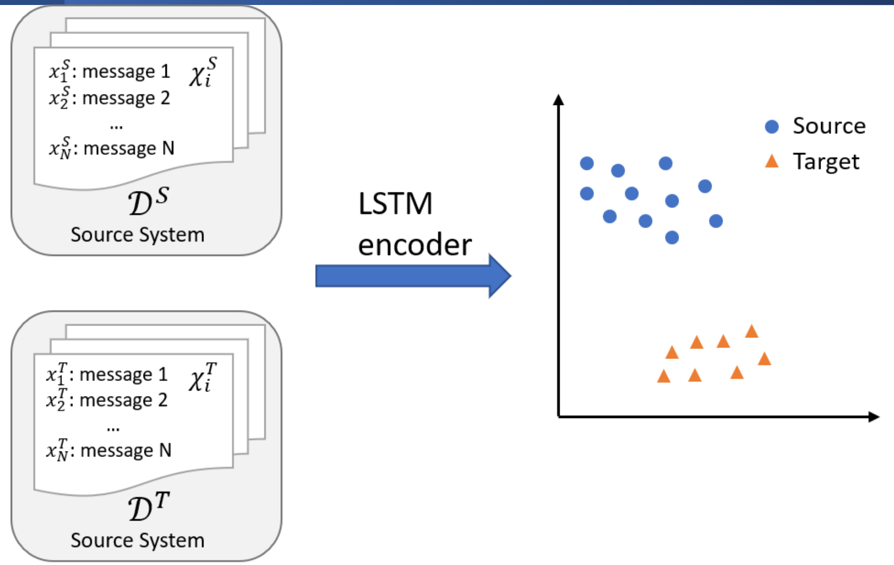
- To make the representation of normal log sequences close to the center $\mathbf{c}$, we develop the following objective function,

$$\mathcal{L}_{en} = \sum_{\epsilon \in \{\mathrm{S},\mathrm{T}\}} \sum_{i=1}^{M_\epsilon} || \mathbf{v}_i^{\epsilon} - \mathbf{c} ||^2,$$

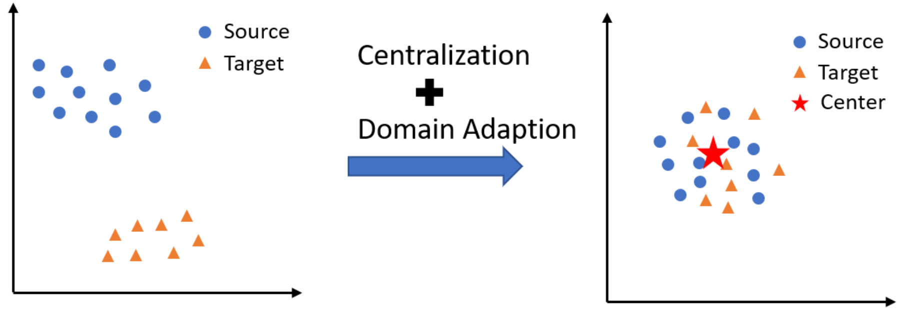
where M_ϵ is the number of samples from the specific domain.

System-agnostic Representation

- Although we adopt one shared LSTM model to map log sequences into a hypersphere, the representations of log sequences from different systems can be still located in different regions.



- Hence, we propose an adversarial training method for cross-system data mapping.
- In specific, we formulate the adversarial training with a discriminator D and a shared LSTM as a generator G.



System-agnostic Representation via Domain Adversarial Training

- Discriminator D is used to distinguish whether the representations of log sequences are from the source or target system,

$$D(\mathbf{v}^{\epsilon}) = \sigma(\mathbf{w}^T \mathbf{v}^{\epsilon} + b),$$

  where $\sigma(\bullet)$ indicates the logistic function, $\mathbf{w}$ and b are the trainable parameters.

- The shared generator G is trained to make representations of log sequences,

$$\mathbf{v}^{\epsilon} = G(\chi^{\epsilon}).$$

System-agnostic Representation via Domain Adversarial Training

- With the adversarial training objective function,

$$\mathcal{L}_{adv} = \min_{G} \max_{D} (\mathbb{E}_{\chi^{s} \sim P_{source}}[\log D(G(\chi^{S}))] + \mathbb{E}_{\chi^{T} \sim P_{target}}[\log(1 - D(G(\chi^{T}))]),$$

  our goal is to mix the distributions of source and target log sequences.

- Final objective function for LogTAD,

$$\mathcal{L} = \mathcal{L}_{en} + \lambda \mathcal{L}_{adv}.$$

Cross-system Log Anomaly Detection

- For a log sequence χ^{ϵ},

$$\hat{y}_{\chi^{\epsilon}} = \begin{cases} anomalous, & if\ ||G(\chi^{\epsilon}) - c||^2 > \gamma^{\epsilon} \\ normal, & else \end{cases}$$

where $\epsilon \in \{\mathrm{S}, \mathrm{T}\}$ and γ^{ϵ} can be derived from a small validation set.

- Introduction
- LogTAD
- Experiment and Analysis
- Conclusion

Datasets

- Statistics of the Datasets

Dataset	# of Logs	# of Log Sequences	
		Normal	Anomalous
BGL	1,212,150	265,583	37,450
TB	3,737,209	565,817	368,481

- Statistics of Shared Words Across Systems

	BGL Normal	BGL Anomalous	TB Normal	TB Anomalous
BGL Normal	664	133	254	25
BGL Anomalous	133	195	99	16
TB Normal	254	99	1753	49
TB Anomalous	25	16	49	54

Baselines

- Unsupervised Log Anomaly Detection Approaches
  - PCA
  - LogCluster
  - DeepLog
  - DeepSVDD
- Supervised Transfer Learning Approach for Log Anomaly Detection
  - LogTransfer

Experimental Results Compared with Unsupervised Approaches

BGL -> TB				
Method	Source		Target	
	F1	AUC	F1	AUC
PCA w/o TB	0.642	0.816	0.558	0.504
LogCulster w/o TB	0.713	0.829	0.559	0.504
DeepLog w/o TB	0.578	0.867	0.556	0.500
DeepSVDD w/o TB	0.566	0.789	0.577	0.646
LogTAD	0.926	0.964	0.758	0.804

TB -> BGL				
Method	Source		Target	
	F1	AUC	F1	AUC
PCA w/o BGL	0.760	0.779	0.229	0.658
LogCulster w/o BGL	0.724	0.716	0.223	0.500
DeepLog w/o BGL	0.660	0.677	0.223	0.500
DeepSVDD w/o BGL	0.794	0.808	0.195	0.497
LogTAD	0.788	0.797	0.845	0.909

Experimental Results Compared with Unsupervised Approaches Cont.

BGL -> TB				
Method	Source		Target	
	F1	AUC	F1	AUC
PCA w/ TB	0.322	0.587	0.731	0.776
LogCulster w/ TB	0.530	0.746	0.677	0.716
DeepLog w/ TB	0.662	0.854	0.590	0.619
DeepSVDD w/ TB	0.499	0.725	0.567	0.616
LogTAD	0.926	0.964	0.758	0.804

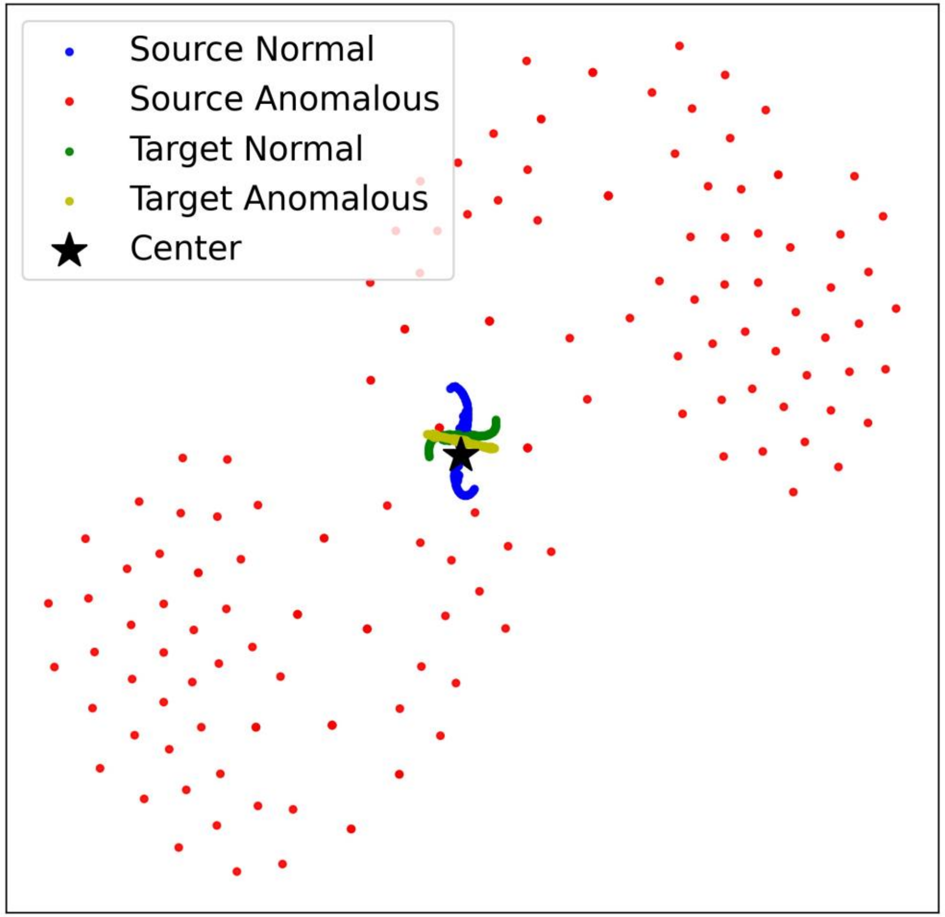
TB -> BGL				
Method	Source		Target	
	F1	AUC	F1	AUC
PCA w/ BGL	0.789	0.798	0.577	0.773
LogCulster w/ BGL	0.708	0.688	0.697	0.886
DeepLog w/ BGL	0.687	0.701	0.527	0.843
DeepSVDD w/ BGL	0.660	0.699	0.196	0.537
LogTAD	0.788	0.797	0.845	0.909

Experimental Results Compared with Supervised Approach

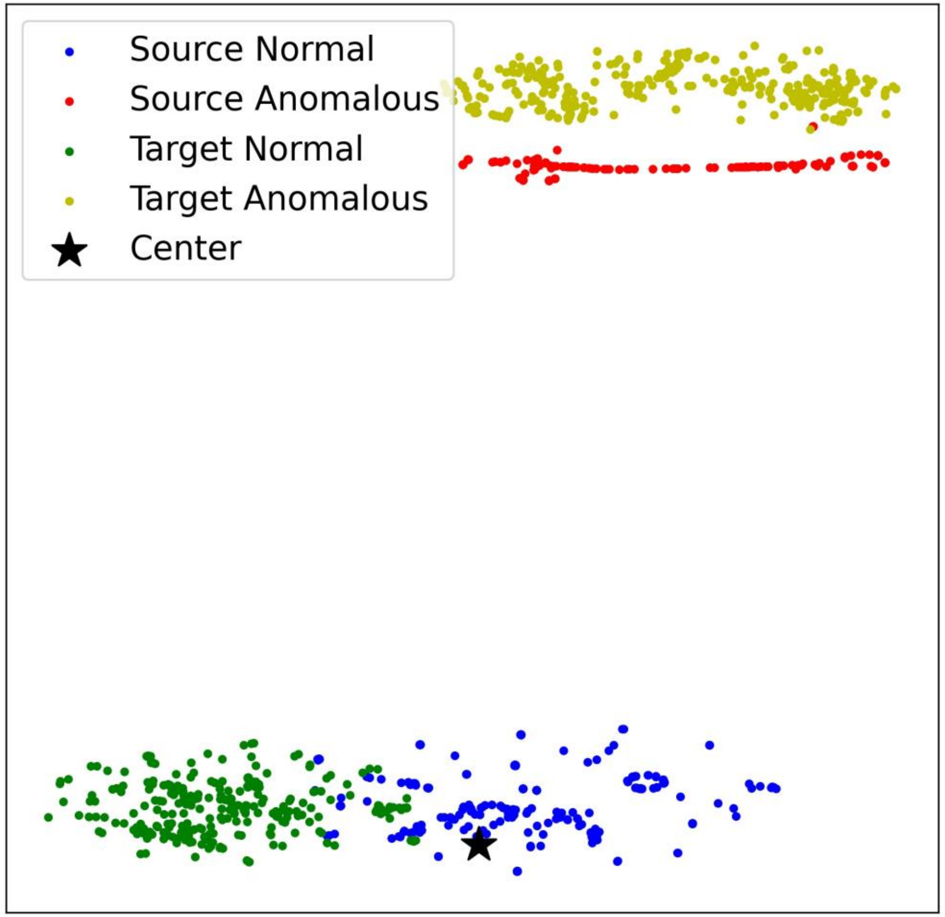
BGL -> TB				
Method	Source		Target	
	F1	AUC	F1	AUC
LogTransfer	0.971	0.972	0.792	0.828
LogTAD	0.926	0.964	0.758	0.804

TB -> BGL				
Method	Source		Target	
	F1	AUC	F1	AUC
LogTransfer	0.995	0.995	0.788	0.833
LogTAD	0.788	0.797	0.845	0.909

Log Sequences Visualization



Without domain adaption



With domain adaption

- Introduction
- LogTAD
- Experiment and Analysis
- Conclusion

Summary

- We propose an unsupervised cross-system log anomaly detection framework.
- LogTAD utilizes the domain adversarial adaption to make the log data from different systems follow similar distributions.
- LogTAD can detect anomalies in different systems with large distances to the center.
- The experiment results show the effectiveness of our framework.

Thank You for Your Attention!

This work was supported in part by NSF 2103829.