Transfer Learning Approach for Botnet Detection based on Recurrent Variational Autoencoder

Jeeyung Kim

Scientific Data Management Research Group
Computational Research Division
Lawrence Berkeley National Laboratory

Introduction

- **Botnet is one of the most significant threats to the cyber-security**
  - Bot masters hijack other machines, and command to act together to attack more machines
  - Attack types : DDos, Click-fraud, spamming, crypto-mining
  - Communication methods : Internet Relay Chat (IRC), peer-to-peer (P2P) and HTTP

➢ One of the task of cybersecurity research is to detect botnets

Introduction

- **Existing approaches: signature based and anomaly-based**
  - a) signature-based : detect botnets with a set of rules or signatures
  - b) anomaly-based methods : detect botnets based on a number of network traffic anomalies such as high network latency, high volumes of traffic and unusual system behavior (Zeidanloo et al. 2010)
    - Machine learning(ML) methods: Zhao et al. 2013, Venkatesh et al. 2012, Singh et al. 2014, Beigi et al. 2014, Stevanovic et al. 2014

Introduction

- **Supervised learning methods**
  - Promising results with a high degree of accuracy for detecting botnets (Du et al. 2019, Ongun et al. 2019, Singh et al 2014)
  - Assumes the provision of data labels to classify -> unavailable in practice.
- **Semi-supervised learning methods**
  - Straightforward to collect
  - The detection performance: generally much lower than supervised learning techniques
    - Autoencoders (AEs) (Dargenio et al. 2018)
    - Variational Autoencoder (VAEs) (An et al. 2015, Nguyen et al. 2019, Nicolau et al. 2018)
    - One-class support vector machines (OSVMs) (Nicolau et al. 2018)

Introduction

- **Transfer learning methods : utilize labeled data available in another domain (“source domain”) for the domain of interest(“target domain”)**
  - **Transfer learning – construct a learning model without the data-labeling effort via knowledge transfer (Pan et al. 2009)**
- **Transfer learning methods in anomaly detection**
  - **Andrews et al. 2016 ,Chalapathy et al. 2018, Ide et al. 2017, Xiao et al. 2015**
  - **Focus on text classification, speech recognition, image classification**
- **Transfer learning for botnet detection**
  - **Alothman et al. 2018, Bhodia et al. 2019, Jiang et al. 2019, Kumagai et al. 2019, Singla et al. 2019, Stevanovic et al. 2014**
  - **Depend on naive techniques**
    - **Calculating similarity or heuristic methods**
  - **Most of them require both normal and anomalous instances for source and target domains**

Contribution

- **Transfer learning framework which constructs a learning model without the label information in the target domain**
  - **Use Recurrent Variational Autoencoder (RVAE) model to obtain anomaly scores**
- **Detect potential botnets in the new network monitoring data set**
  - **With the knowledge transferred from the popular dataset, CTU-13, as the source domain**

Preliminary

- **Transfer Learning**
  - Classification or regression tasks in one domain of interest
  - Only have sufficient labeled data in different domains, where the latter data may follow a different data distribution (Pan et al. 2009)
  - Can be divided into three categories according to source/target domains label existence and the types of tasks
    - Inductive transfer learning
    - Transductive transfer learning
    - Unsupervised transfer learning
- **Recurrent Variational Autoencoder**
  - Combine seq2seq(RNN-to-RNN structure) with VAE
  - The methods to use RVAE as botnet detector in (Kim et al. 2020)

Related Works

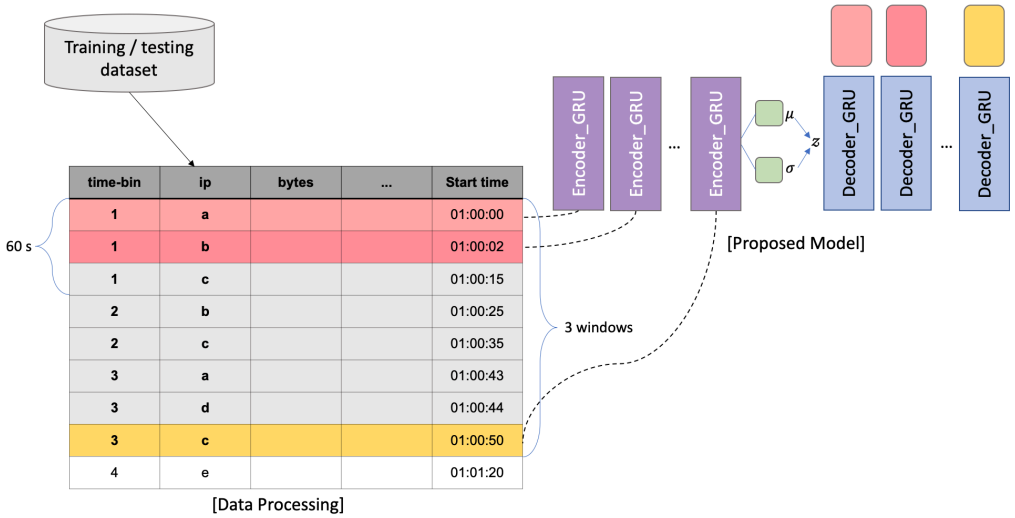
- **Network IDS methods**
  - Daya et al. 2020, Binkley el al. 2006, Gu et al. 2008, Paxson et al. 1999, Roesch et al. 1999, Zeidanloo et al. 2010
    - Use statistical deviations or rules to detect botnet
    - Cannot detect new botnets
  - Zeek : popular network IDS, which is a monitoring system for detecting network intruders in real-time
    - Zeek is not for detecting botnet
- **ML methods**
  - VAE/AE
    - Dargenio et al. 2018, Kim et al. 2020, Nguyen et al. 2019, Nicolau et al. 2018
    - The methods overlook sequential characteristics within network traffic
  - RNN
    - Kim et al. 2020, Ongun et al. 2019, Sinha et al. 2019, Torres et al. 2016
    - The method cannot be applied to the online anomaly detection system
  - Others Random Forest, Neural Network
    - Du et al. 2019, Ongun et al. 2019, Venkatesh et al. 2012
  - Require fully labeled dataset which is hard to obtain due to lack of labeled data on changing network traffic.

Related Works

- **Transfer learning on botnet detection**
  - Alothman 2018, Bhodia et al. 2019, Jiang et al. 2019, Kumagai et al. 2019, Singla et al. 2019, Taheri et al. 2018
  - Most depends on naive techniques such as calculating similarity
    - requires high computation cost
  - Clustering & naïve rule methods
    - Jiang et al. 2019
  - Neural Network
    - Bhodia et al. 2019, Singla et al. 2019, Taheri et al. 2018
    - Requires labeled dataset for both source and target domains contrary to the proposed method not requiring labeled dataset for a target domain.

Proposed Model

- **Anomaly Detection Method**
  - Use RVAE as an anomaly detector
  - Input : pre-processed flow-based features
  - Output : reconstructed input
  - Training / evaluation method
    - Train the model with only normal instances
      - Reconstruction errors of anomalous samples: larger than that of the normal samples
    - Collect each reconstruction loss, then estimate distribution in the validation phase
      - Represents collected reconstruction errors from normal and anomalous instances, respectively.
    - Get two likelihoods for each instance from normal and anomalous distributions in the testing phase
      - The network traffic flow data can be classified by comparing the two values.



RVAE [Kim et al. 2020]

Proposed Model

- **The process of transfer learning**
  1. Follow the procedure of transfer anomaly detection method (Kumagai et al. 2019)
  2. Further develop the method to be trained without label information on the target domain
     - Hard to obtain labeled data of network traffic data

➢ Two cases of training data on botnet detection: labeled dataset on the target domain (*with_label*) and unlabeled dataset on the target domain (*without_label*).

- The normal and anomalous instances in a source domain are used for training RVAE in the both methods
- After updating parameters of RVAE with the source domain samples, update parameters of RVAE with the target domain samples

Proposed Model

- **The objective function of the source domain (Kumagai et al. 2019) :**

$$s_{\phi,\theta}(x|z) = \sum_{n=1}^{N} (1 - x_n) log(1 - \tilde{x}_n) + x_n log\tilde{x}_n$$

$$L_s(\theta, \phi|z) = \frac{1}{N_s^-} \sum_{n=1}^{N_s^-} s_{\phi,\theta}(x_s^-|z) - \frac{\lambda}{N_s^- N_s^+} \sum_{n,m=1}^{N_s^-, N_s^+} f(s_{\phi,\theta}(x_s^+|z) - s_{\phi,\theta}(x_s^-|z))$$

$$\mathbb{L}_s(\phi, \theta) = \mathbb{E}_{q_\phi(z|x)}[L_s(\theta, \phi|z)] + \beta D_{KL}(q_\phi(z|X_s^-)|p(z))$$

- **Notation used**
  - X_s^+: a set of anomalous instances in a source domain
  - X_s^-: a set of normal instances in a source domain
  - X_t^+: a set of anomalous instances in a target domain
  - X_t^-: a set of normal instances in a target domain
  - D : the number of features
  - F_θ: Encoder, G_ϕ : Decoder
  - N_s^+, N_s^- : the number of instances of anomalous and normal on the source domain
  - z : the latent variable

Proposed Model

- **The process of transfer learning**
  - **The proposed method can be categorized into two based on whether the labeled dataset on the target domain is necessary or not.**
    - **Transfer learning with the unlabeled dataset on the target domain is different from the method with the method using the labeled data set on the target domain regarding that it uses entire instances in the target domain for training.**
    - **Only normal instances in the target domain are used for training on with label method**
    - **Different objective function of the target domain of the two methods.**
    - **In the source domain, the objective functions on both methods are equal to each other**

Proposed Model

Algorithm 1: The Procedure of Training Transfer Anomaly Detection *with_label* Method

Input: instances of source domain $x_s^{-,+} \in X_s^{-,+}$ and instances of target domain $x_t^- \in X_t^-$
Output: G_θ, F_ϕ
Procedure
  for *the number of epochs* **do**
    Sample minibatches from X_s^+, X_s^- and X_t^- $(B_{x_s^-} \subset X_s^-, B_{x_s^+} \subset X_s^+, B_{x_t^-} \subset X_t^-)$
    for $B^a_{x_s^-}, B^c_{x_s^-}, B^e_{x_t^-}, (a = 1, ..., A, c = 1, ..., C, e = 1, ..., E)$ **do**
      forall $x_s^- \in B^a_{x_s^-}$ **do**
        $\tilde{x}_s^- = G_\theta(F_\phi(x_s^-))$
        forall $x_s^+ \in B^c_{x_s^-}$ **do**
          $\tilde{x}_s^+ = G_\theta(F_\phi(x_s^+))$
        end
        Update the Encoder and the Decoder by descending its stochastic gradient:
        $\nabla_{\theta,\phi}(L_s(\phi, \theta))$
      end
      forall $x_t^- \in B^e_{x_t^-}$ **do**
        $\tilde{x}_t^- = G_\theta(F_\phi(x_t^-))$
      end
      Update the Encoder and the Decoder by descending its stochastic gradient:
      $\nabla_{\theta,\phi}(L_t(\phi, \theta))$
    end
  end
end

1. Using label information in a target domain (*with_label*)

- Use only normal instances for training on a target domain
- The objective function for the target domain :

$$\mathbb{L}_t(\phi, \theta) = \mathbb{E}_{q_\phi(z|x)}\left[\frac{1}{N_t^-}\sum_{n=1}^{N_t^-} s_{\phi,\theta}(x_t^-|z)\right] + \beta D_{KL}(q_\phi(z|X_t^-)|p(z))$$

Proposed Model

2. Not using label information in a target domain (*without_label*)

- Use the entire instances of the dataset for the first several epochs during training on the target domain.
  - After E epochs, we collect instances which show lower reconstruction errors in each mini-batch.
    - The instances with lower reconstruction errors -> possibly to be normal.
- Normal instance selection process
  a) Sort the instances by the size of reconstruction errors every minibatch.
  b) Select an instance of the bottom r% of reconstruction errors in minibatch and add the portion of instances to the next minibatch training samples.

$$\mathbb{L}_t(\phi,\theta) = \mathbb{E}_{q_\phi(z|x)}\left[\frac{w_{x_t}}{M_t}\sum_{n=1}^{M_t} s_{\phi,\theta}(x_t|z)\right] + \beta D_{KL}(q_\phi(z|X_t)|p(z))$$

➢ Train the anomaly detector effectively on the target domain without label information via the selecting samples method

Experiments

- **Evaluation metrics : AUROC, TPR, FPR, TNR, FNR**
- **Evaluation datasets**
  - Existing studies used the same dataset for a target domain and source domains
  - Our objective is detecting suspicious botnet connections on the new network monitoring dataset
  - Source domain : CTU-13 dataset (scenario 1,2 and 9) – botnet Neris
    - Data collected from Zeek
  - Target domain : a network monitoring data set from a large research institute (dataset K)
    - Data collected from Zeek

Experiments

- **Labeling method**
  - Both CTU-13 Zeek data and the dataset K do not have label
    - New labeling method is required
  - *weird.log* has no correlation with botnet label in the original CTU-13
  - Most connections with the indication of *irc_line_too_short* and *irc_invalid_line* are given by Neris
    - Neris accounts for 84% / 82% of connections with the indication of *irc_line_too_short* / *irc_invalid_line* among data from 13 scenarios.
  - ➢ Use the indication information from *weird.log*, and label host IP address with *irc_line_too_short* and *irc_invalid_line* as *malicious*

Experiments

- **Data Preprocessing**
  - Use the aggregated flows statistics (Kim et al. 2020)
- **Comparison methods**
  - Propose method : *with_label, without_label*
  - Baseline : RVAE
    - Semi-supervised anomaly detection method (Kim et al. 2020)

Results and Discussion

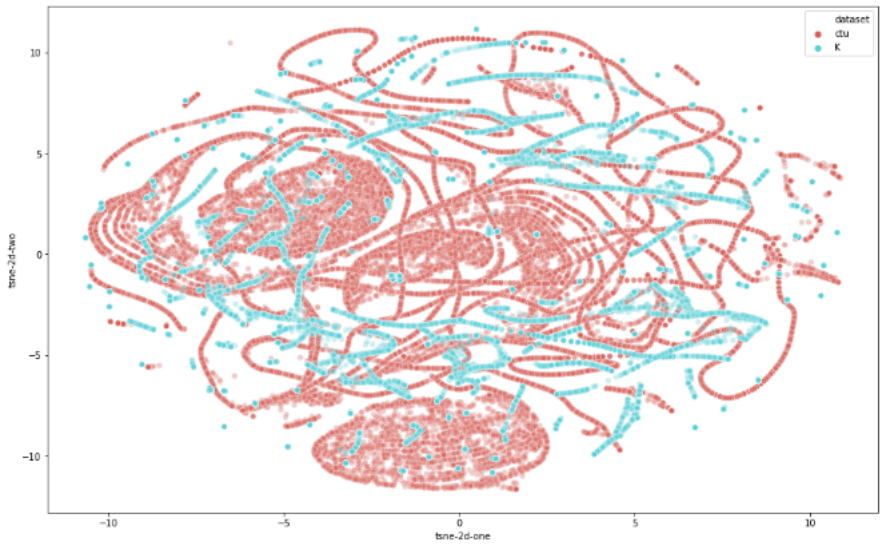


Figure 1: T-SNE plot over the source and the target domains

- For transfer learning, two domains should be related and share common characteristics.
- The source domain dataset and the target domain dataset are not generated in the same environment.
  - The source domain dataset
    - Made in the environment where attacks of botnet are controlled.
  - The target domain dataset
    - Collected using a Zeek server connected to the switch between the Internet and the local network.
- The two distributions cannot be completely overlapping, but at the same time, the two distributions should not be completely separated
  - Both data share common characteristics generated from Zeek
  - Transfer learning can be applied to the datasets

Results and Discussion

Table 1: Average of each metric over the target domain (dataset K), r_s is 0.1 in the *without_label* method

Model	TNR	TPR	FNR	FPR	AUROC
RVAE	**0.685**	0.811	0.189	**0.309**	0.779
with_label	0.652	**0.915**	**0.084**	0.371	**0.810**
without_label	0.634	0.850	0.150	0.365	0.764

- **Our proposed method *with_label* outperform**
  - **TPR (detection rate) of *with_label* method is 0.915 while TPR of RVAE method is 0.811**
  - ***with_label* method shows higher AUROC than the RVAE**
- **Even *without_label* method which does not use label information on the target domain shows higher performance than RVAE on TPR and FNR metrics.**
- **Proposed method detects suspicious botnet better on the target domain**
  - **Using transferred knowledge which is obtained on the related domain (source) can provide useful information for the target domain lack of training data.**

Conclusion

- **Transfer learning framework: an effective botnet detection strategy**
  - Useful for network security applications because security challenges such as botnets are constantly evolving
- **Train neural network on labeled data from CTU-13 and apply the network for anomaly detection on a fresh set of network monitoring data.**
  - Test shows that transfer learning could reliably identify anomalies.
- **For future studies,**
  - Propose more systematic method beyond empirical ways to improve $without\_label$ method
  - Improve performance of the anomaly detector in FPR measure as it shows weak performance relatively