Cyber Security: Bayesian Neural Networks for Anomaly Detection with Uncertainty Estimation and Interpretability

Background

Anomaly detection is the process of identifying data points, events, or observations that do not conform to an expected pattern or behaviour [1]. In the realm of banking cyber security, this process is critical for identifying potential threats and mitigating risks associated with cyber-attacks. Traditionally, rule-based systems and **conventional statistical methods** have been deployed to monitor network traffic and system logs; however, these techniques often struggle to adapt to evolving threat landscapes and tend to produce high false positive rates. More recently, **deep learning models such as Autoencoders (AEs) and Variational Autoencoders (VAEs)** [2][3] have shown promise by learning the underlying distribution of normal behaviour and flagging deviations. Despite their effectiveness, these models typically lack the capability to provide robust uncertainty estimates—an essential feature for prioritizing security alerts and guiding manual investigations.

Project Overview

This project will focus on the development of **Bayesian Neural Networks (BNNs)** [4] for anomaly detection in cyber security, **emphasizing the estimation of uncertainty and model interpretability** [4]. By treating network weights as probability distributions rather than fixed parameters, BNNs inherently provide confidence intervals for each prediction. These confidence measures allow security teams to rank alerts based on risk, thus streamlining the investigation process. The project will employ the **UNSW-NB15 Network Intrusion Dataset** [5][6] —a comprehensive dataset featuring heterogeneous network traffic that includes both normal and malicious activities—to evaluate the proposed approach.

Research Questions

Given the above context, please address the following three research questions:

1. Can we build a **Bayesian Neural Network** to detect anomalies in cyber security data while providing reliable uncertainty estimation?
2. How do **Bayesian Neural Networks**, with integrated uncertainty quantification, compare to traditional anomaly detection techniques like **Local Outlier Factor (LOF) and DBSCAN** and/or how against advanced Deep Learning approaches

such as **Autoencoders (AEs) and Variational Autoencoders (VAEs)** in terms of detection accuracy?

3. How can interpretability frameworks (e.g., **LIME or SHAP**) [7][8] be used to ensure that the models' predictions are both actionable and secure?

Methodology

1. **Dataset -** Utilize the UNSW-NB15 Network Intrusion Dataset.
2. **Model Development –** Design and implement a BNN architecture for anomaly detection.
3. **Model Benchmarking –** compare your model's results against other statistical or ML/DL anomaly detection approaches.
4. **Evaluation Metrics -** assess model performance to measure the effectiveness of anomaly detection and evaluate the quality of uncertainty estimates.
5. **Interpretability Analysis -** apply interpretability frameworks to identify which features are most influential in the model's decision-making process.

References

[1] [Online]. Available: https://cumulocity.com/guides/machine-learning/anomaly-detection/

[2] I. G. a. Y. B. a. A. Courville, "Deep Learning", [Online]. Available: http://www.deeplearningbook.org.

[3] M. W. Diederik P Kingma, "Auto-Encoding Variational Bayes", [Online]. Available: https://arxiv.org/abs/1312.6114.

[4] Lingxue Zhu, Nikolay Laptev , "Deep and Confident Prediction for Time Series at Uber", [Online]. Available: https://arxiv.org/abs/1709.01907

[5] [Online]. Available: https://research.unsw.edu.au/projects/unsw-nb15-dataset.

[6] Z. Z. a. G. Serpen, "UNSW-NB15 Computer Security Dataset: Analysis through", [Online]. Available: https://arxiv.org/ftp/arxiv/papers/2101/2101.05067.pdf.

[7] [Online]. Available: https://github.com/marcotcr/lime.

[8] [Online]. Available: https://shap.readthedocs.io/en/latest/index.html.