Deep Learning Based Intrusion Detection System

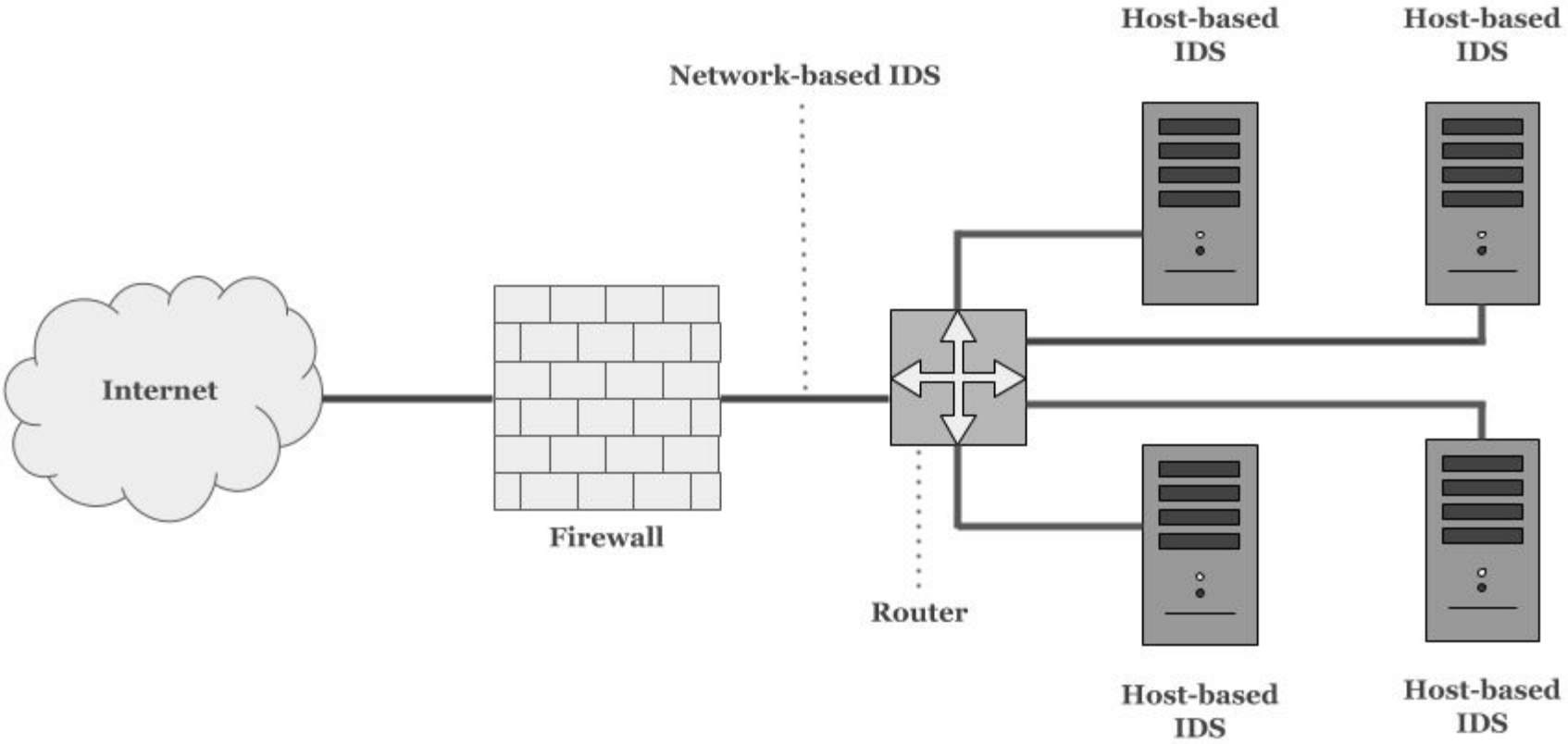
CDAC Workshop
Sriram S

Intrusion Detection System (IDS)

- Intrusion detection system deals with unauthorised access and threats to systems and information by any type of user or software.
- Intrusion can be **external** or **internal**.
- External intrusion is when an intruder tries to gain access to a protected internal network.
- Internal intrusion is when an insider with a motive tries to misuse, attack or steal information. This is also called **insider threat**.

Types of IDS

- Two major types of IDS are **network-based** IDS and **host-based** IDS.
- A network-based intrusion detection system (NIDS) is used to monitor and analyze network traffic to protect a system from network-based threats.
- A Host-based intrusion detection system (HIDS) is a system that monitors the system in which it is installed to detect both internal and external intrusion, misuse and responds by logging the activities and notifying the designated authority.


Host-based IDS
Host-based IDS
Network-based IDS
Internet
Firewall
Router
Host-based IDS
Host-based IDS

NIDS

- Signature based NIDS uses signatures which are extracted from previously known attacks.
- Signatures are manually generated and stored in the database whenever a new attack is identified.
- New attacks will not be detected by this system.
- Anomaly based NIDS models the normal behaviour of the network and raises alarm whenever it detects an anomalous behaviour.
- Hybrid NIDS uses the combination of the above two approaches.

Attack category	Description	NSL-KDD (10% of Data)	
		Train	Test
Normal	Normal connection records	67,343	9,710
DoS	Attacker aims at making network resources down	45,927	7,458
Probe	Obtaining detailed statistics of system and network configuration details	11,656	2,422
R2L	Illegal access from remote computer	995	2,887
U2R	Obtaining the root or superuser access on a particular computer	52	67
Total		**125973**	**311029**

Attack Types

- Denial of Service Attack (DoS): It is an attack in which the attacker makes some computing or memory resource too busy or too full to handle legitimate requests, or denies legitimate users access to a machine.
- User to Root Attack (U2R): It is a class of exploit in which the attacker starts out with access to a normal user account on the system and is able to exploit some vulnerability to gain root access to the system.

Attack Types

- Remote to Local Attack (R2L): It occurs when an attacker who has the ability to send packets to a machine over a network but who does not have an account on that machine exploits some vulnerability to gain local access as a user of that machine.
- Probing Attack: It is an attempt to gather information about a network of computers for the apparent purpose of circumventing its security controls.

Adversarial Attacks

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$

$\text{sign}(\nabla_{\boldsymbol{x}} \text{L}(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$=$

$\boldsymbol{x} +$
$\epsilon \text{sign}(\nabla_{\boldsymbol{x}} \text{L}(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Fast Gradient Sign Method (FGSM)

- Fast gradient sign method is a simple method of generating adversarial samples. In FSGM, a small perturbation is calculated in the direction of the gradient:

$$p = \epsilon sign(\nabla_x L(\theta, x, y))$$

- where p is the perturbation, $\in$ is a small constant, ∇xL(θ, x, y) is the gradient of loss function L which is used for training the model, x is the input to the model and y is the class of input x. This perturbation *p* is added to the input data to generate adversarial samples.

Jacobian-based Saliency Map Attack (JSMA)

- Jacobian-based Saliency Map Attack uses the concept of saliency maps to generate adversarial samples.
- A saliency map gives insights about the features of the input data that are most likely to create a change of targeted class.
- In other words, saliency maps rates each feature how influential it is for causing the model to predict a target class.
- JSMA causes the model to misclassify the resulting adversarial sample as a specified erroneous target class by modifying the high-saliency features.

Adversarial Robustness Toolbox (ART) Library

- Python library for adversarial attacks and defenses.
- Link : https://github.com/IBM/adversarial-robustness-toolbox