Special Topics: Machine Learning (ML) for Networking

COL867

Holi, 2025

Traffic Classification

Tarun Mangla

Traffic Classification: Recap

- Categorize network traffic into different classes, typically application or traffic type or QoS category

- Potential approaches:
    - Port-based classification
    - Payload-based
    - Analyzing traffic characteristic using ML → $\boxed{\text{Supervised ML}}$
      → unsupervised ML

Paper: Class-of-Service Mapping for QoS.. [Roughan2004]

- Given an aggregate (server IP or port), predict its Class of Service Y_j

Aggregate-level Features → **Supervised ML** → Class of Service

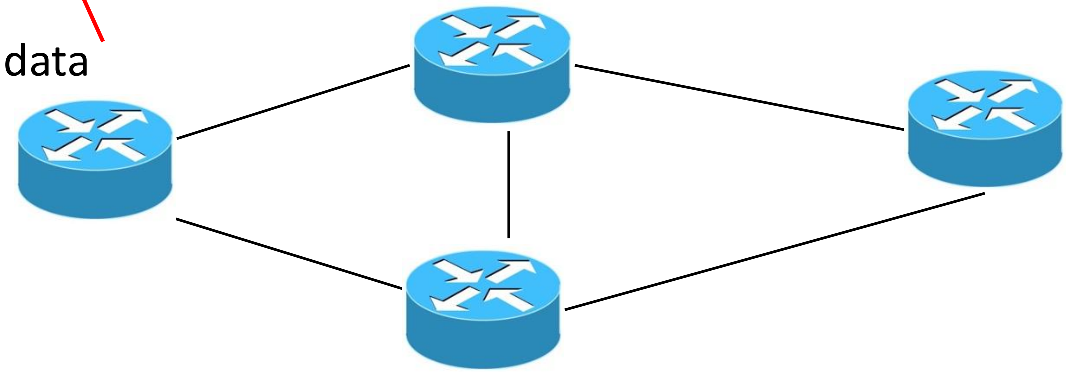
Quality of Service

Statistics on packet size, inter-arrival, and flow duration

k-Nearest Neighbor
Linear Discriminant Analysis

Interactive
Bulk data transfer
Streaming
Transactional

Packet-level data



(Online: f(Traffic) → QoS) → System cost

Inference Model

N/w traffic → Application → QoS

Server IP or Port → QoS

443 → Class1

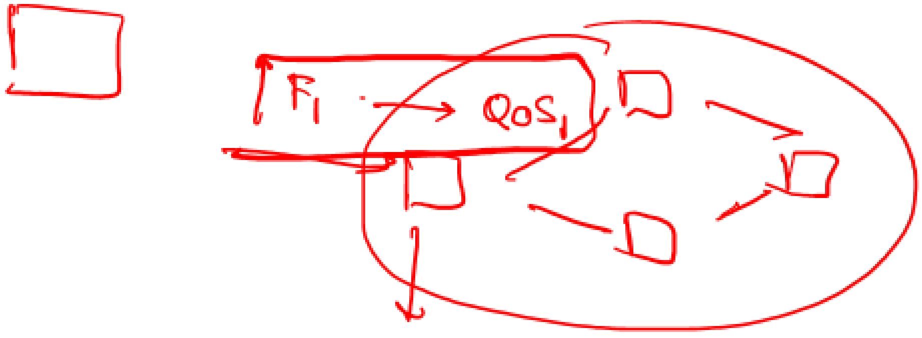
192.10.1.1 / Port → Streaming

- Server IP or port to class of service

- Remnant of the diffServ architecture

Customer 1

Application



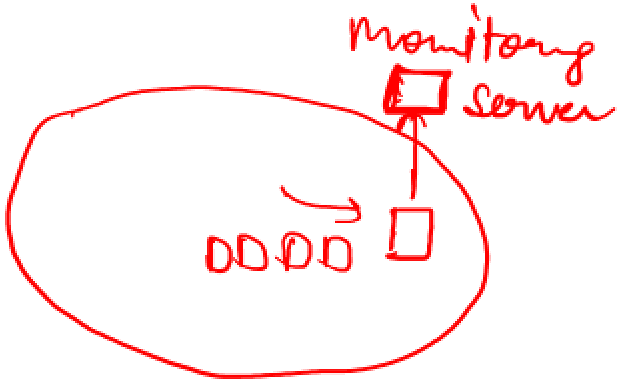
Edge router

Class of Service

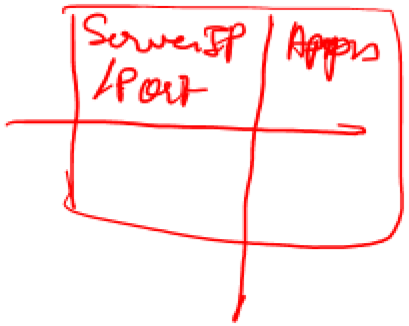
- Interactive: Telnet
- Bulk data transfer: SFTP
- Streaming:
- Transactional: DNS

① Running mean

Feature Extraction



- Four categories of features:
  - Packet-level: packet size [Mean / Variance]
  - Flow-level: flow volume, # of packets
  - Intra-flow features: inter-arrival times, latency
  - Multi-flow: aggregate multiple connections (# connections, mean size per connection)



- Features extracted in a streaming manner ↦ algorithms

average:

→ Var(X) : $E(X - \bar{X})^2$

$$\rightarrow \quad \bar{X}_{j+1} = \frac{1}{j+1}X_{j+1} + \frac{j}{j+1}\bar{X}_j,$$

Median / Quantile : Approx algorithm

Training

- Data collection
  - Public traffic traces
  - Collected from within the ISP network
  - Server logs for a specific application
  - Within enterprise network – collected over two different time intervals

T_1 T_2

generalizability

- Data labeling
  - Port numbers
  - Application payload

Classification Accuracy

$$\frac{\text{Incorrect classification}}{\text{Total samples in test}}$$

algorithm	error rate		
	4 class	3 class	7 class
LDA	5.6 %	3.4 %	10.9 %
1-NN	7.9 %	3.4 %	12.6 %
3-NN	5.1 %	2.5 %	9.4 %
5-NN	5.6 %	2.5 %	9.9 %
7-NN	5.6 %	2.8 %	9.7 %
15-NN	6.2 %	3.4 %	11.4 %

→ Application Type

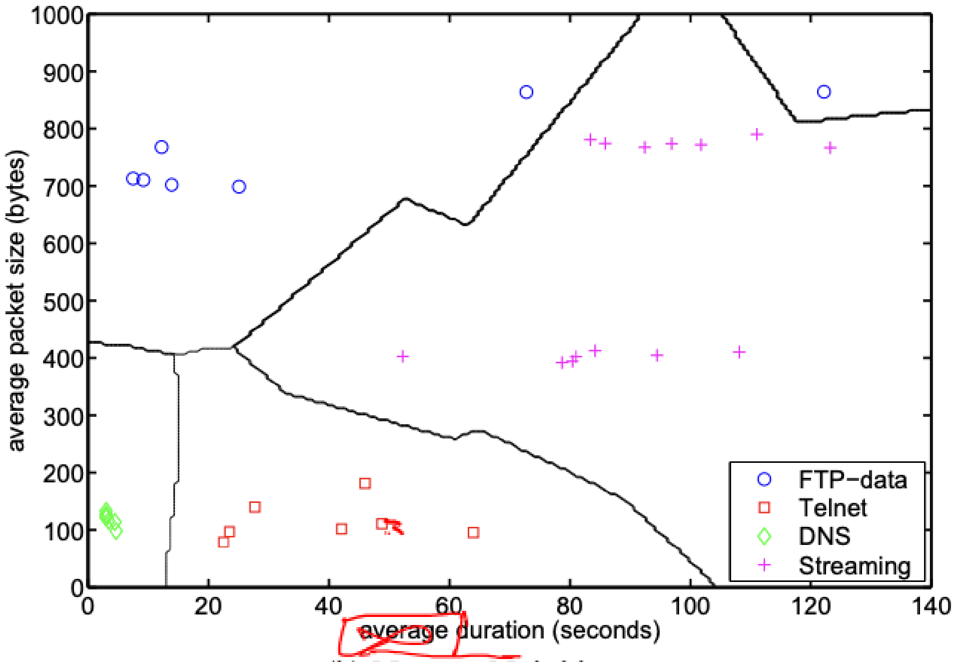
↳ DNS

↳ HTTPS

Which are the most important features?

- Candidate features: average packet size, flow duration, bytes per flow, packet per flow, and root mean square packet size
- **Most important features:** Average packet size and flow duration

Real Media
& FTP

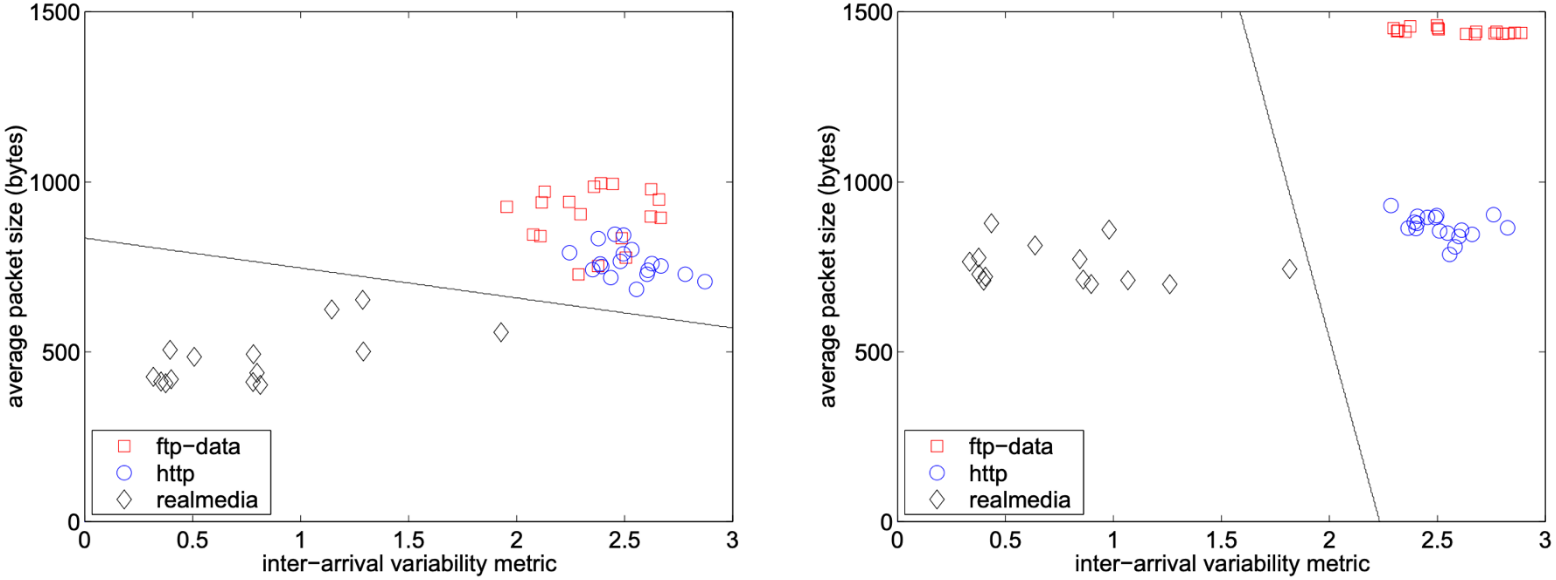


(b) Nearest Neighbor.

Separate FTP and Realmedia using inter-arrival Variability Metrics

Data

ACK



(a) Including ACK streams.

(b) Excluding ACK streams.

(1) Data cleaning →

Application adaptation

From 2004 → 2024: What has changed

FROM POV TRAFFIC CLASSIFICATION

- For good
  - Flexible and scalable network monitoring
  - Advancement in ML techniques
  - Compute capabilities
- For bad
  - Diversity of applications (e.g., IoT traffic)
  - More encryption VPN / TOR → The onion router
  - Scale

Deep Packet: A Novel Approach … [Lotfollahi18]

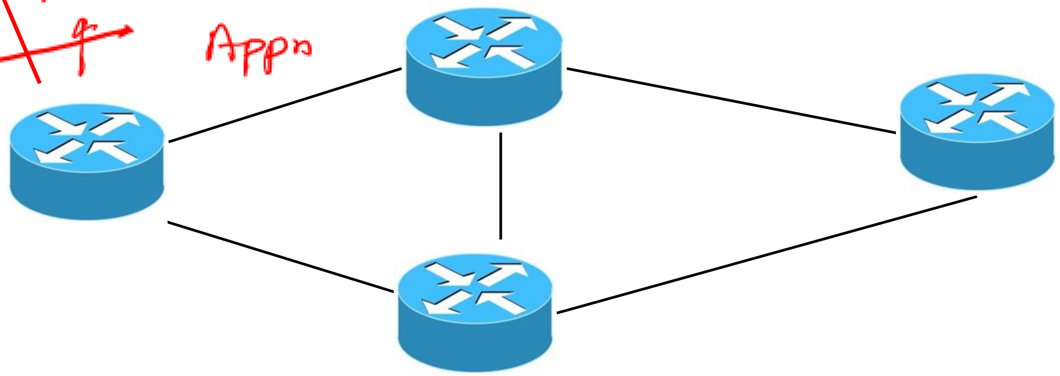
- Given an ~~aggregate~~ packet, predict its ~~Class of Service~~ application

Packet → Deep Learning → Application

Netflix
Facebook
Gmail
…

ML model

Appn



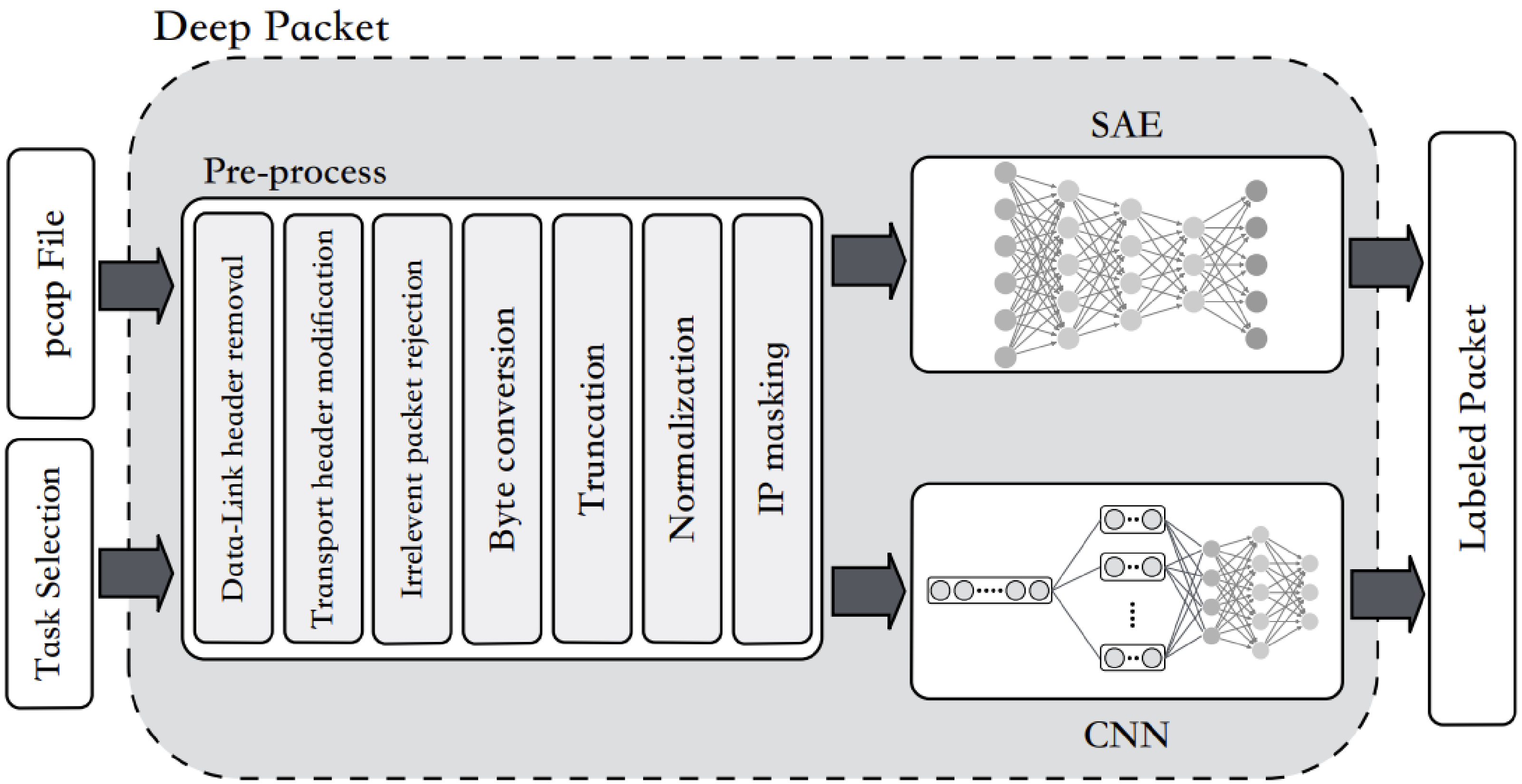
Packet-level data

Motivation: Feature engineering → sub-optimal (expensive, time-consuming, prone to errors)

Data Pre-processing

→ Applications → Netflix
→ Youtube

→ Application class → Streaming
→ WWW



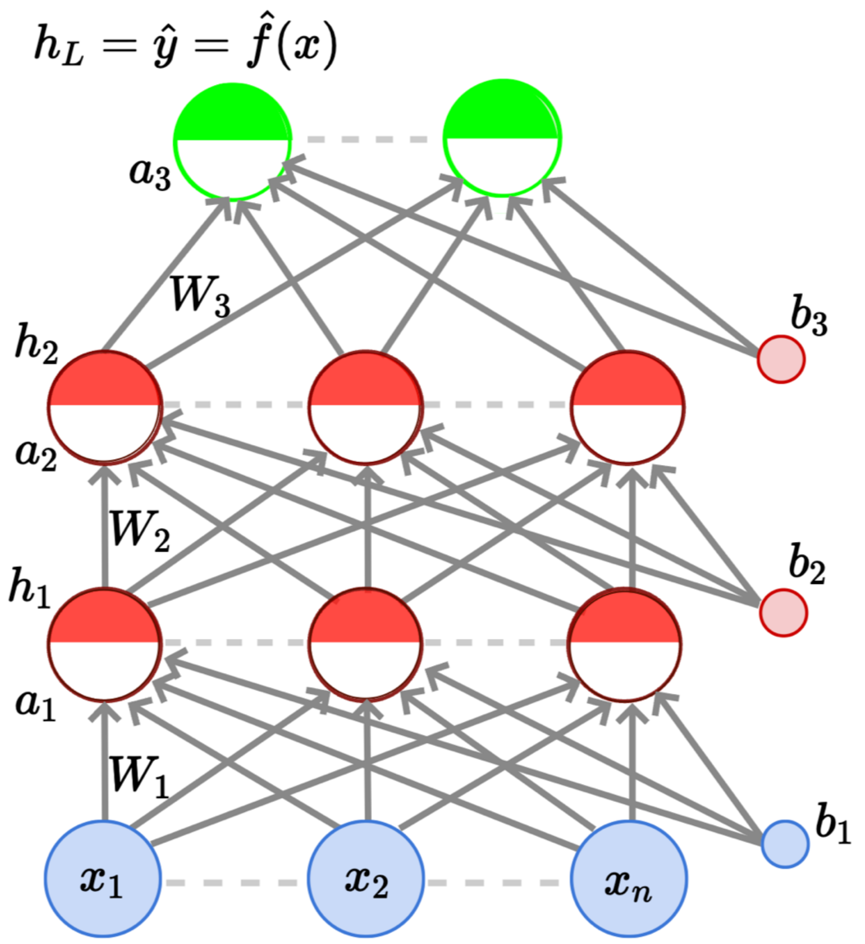
Deep Learning Models Considered

- Autoencoder
- Convolutional Neural Networks

Deep learning MIT:

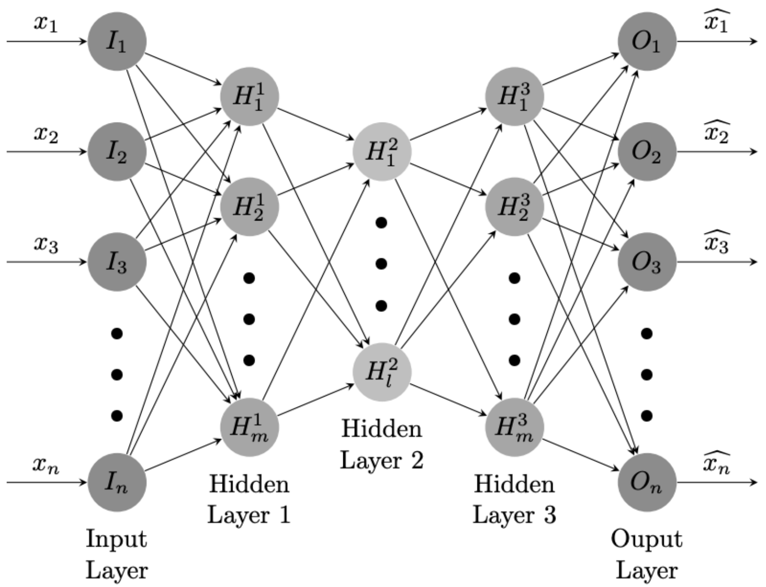
Artificial Neural Networks: Multi-layer Perceptron

- Multi-dimensional input features
- Apply weights and pass them through a neuron with non-linear activation functions
- The weights are derived during training using the backpropagation method
- Deep Neural Networks (DNNs): Similar to MLP but higher number of hidden layers



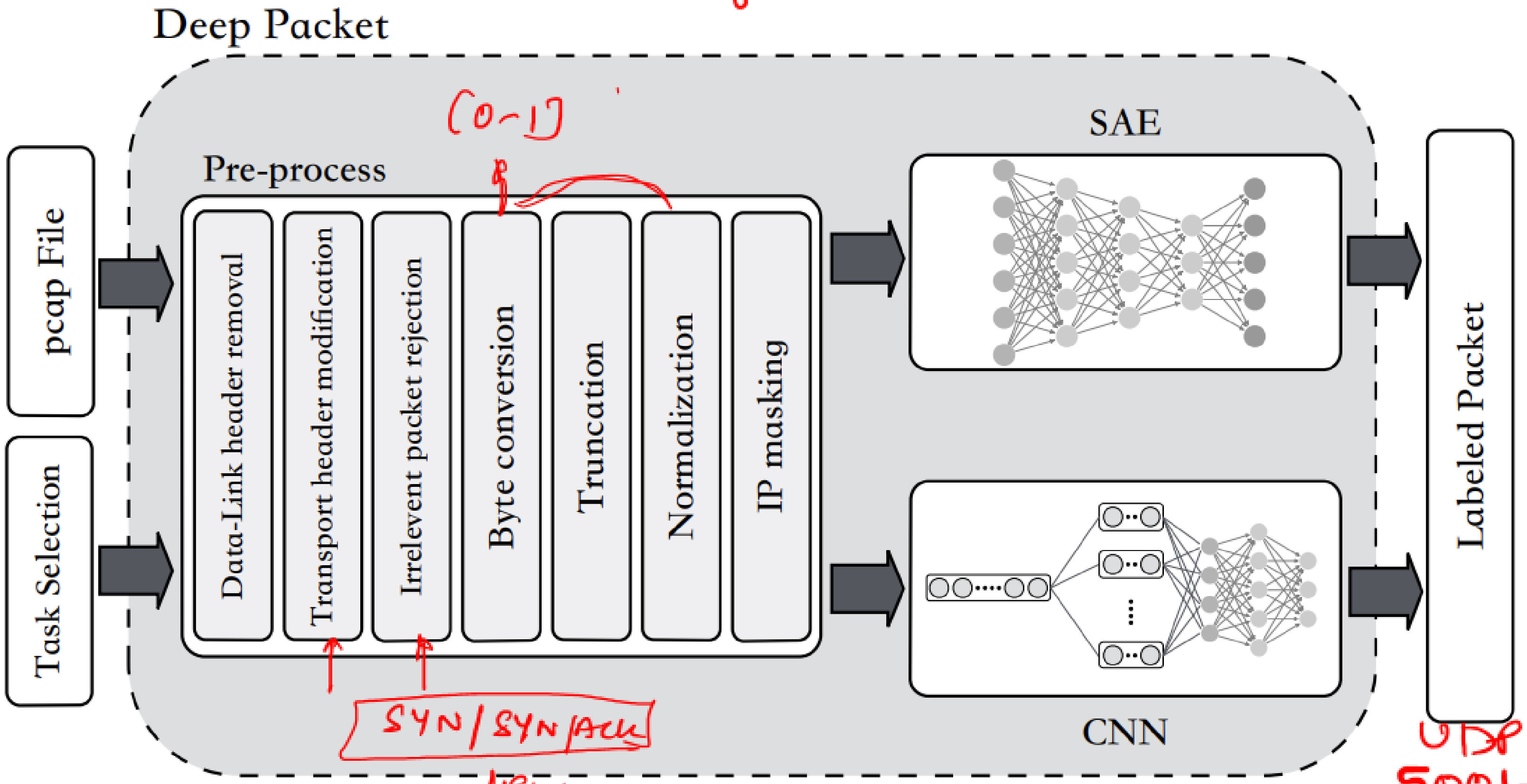
Autoencoder

$(\hat{x} - x)^2$



Autoencoder

Data Pre-processing



S1 byte
ETH | IP | TRANS | APPN
X
PAYLOAD
Server/DST IP

S byte
14XX bytes
S2 bytes

UDP : 8 bytes
TCP : 20 bytes

[0-1]

SYN/SYN/Ack
/FIN
② DNS Traffic

IP | TCP | Appln
IP | UDP | Appn
UDP | | App

UDP
500 byte
IP | UDP | 12 byte | 500 b | 0 | 0 | 0 | 0 | 0
S byte

Training

- Use Tensorflow for training at the backend
- Use early stopping and dropout techniques to avoid overfitting
- Train on ISCX VPN-nonVPN dataset
  - Labeled application traffic
  - VPN and non-VPN traffic as well as Tor traffic

Results

Recall → $\frac{TP}{TP+FN}$

Precision → $\frac{TP}{TP+FP}$

F1-score

1. Traffic is encrypted
2. Model-specific difference
3. Per-packet classification

Application	CNN			SAE		
	Rc	Pr	F_1	Rc	Pr	F_1
AIM chat	0.76	0.87	0.81	0.64	0.76	0.70
Email	0.82	0.97	0.89	0.99	0.94	0.97
Facebook	0.95	0.96	0.96	0.95	0.94	0.95
FTPS	1.00	1.00	1.00	0.77	0.97	0.86
Gmail	0.95	0.97	0.96	0.94	0.93	0.94
Hangouts	0.98	0.96	0.97	0.99	0.94	0.97
ICQ	0.80	0.72	0.76	0.69	0.69	0.69
Netflix	1.00	1.00	1.00	1.00	0.98	0.99
SCP	0.99	0.97	0.98	1.00	1.00	1.00
SFTP	1.00	1.00	1.00	0.96	0.70	0.81
Skype	0.99	0.94	0.97	0.93	0.95	0.94
Spotify	0.98	0.98	0.98	0.98	0.98	0.98
Torrent	1.00	1.00	1.00	0.99	0.99	0.99
Tor	1.00	1.00	1.00	1.00	1.00	1.00
VoipBuster	1.00	0.99	0.99	0.99	0.99	0.99
Vimeo	0.99	0.99	0.99	0.98	0.99	0.98
YouTube	0.99	0.99	0.99	0.98	0.99	0.99
Wtd. Average	**0.98**	**0.98**	**0.98**	**0.96**	**0.95**	**0.95**

Comparison with Other Papers

Paper	Task	Metric	Results	Alg.
Deep Packet	Application	Accuracy	0.98	CNN
Yamansavascilar et al. (2017)	Identification		0.94	k-NN
Deep Packet	Traffic	Precision	0.93	CNN
Gil et al. (2016)	Characterization		0.90	C4.5

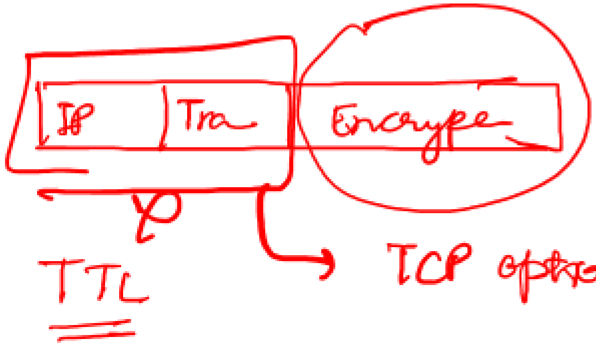
Why does DeepPacket work?

- DeepPacket does not inspect for keywords, **how does it work**?
- Ideal encryption scheme → produces patternless data
- **But, all schemes use (different) pseudo-random generators**
- Leads to patterns in the data

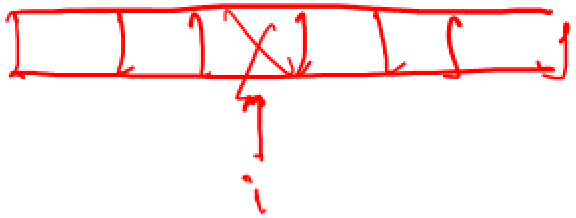
If there are
(1) Patterns in the data

- Is that really true?

Ablation study



TCP options present or not



Difference between the two studies?

- Manual feature extraction
- Explainability/Generalizability?
  - Is the DL model doing shortcut learning?
- Scalability concerns