A Semi-supervised Stacked Autoencoder Approach for Network Traffic Classification

Ons Aouedi, Kandaraj Piamrat, Dhruvjyoti Bagadthey

HDR-Nets 2020 Workshop

The 28th IEEE International Conference on Network Protocols

October 10, 2020

Outline

UNIVERSITÉ DE NANTES

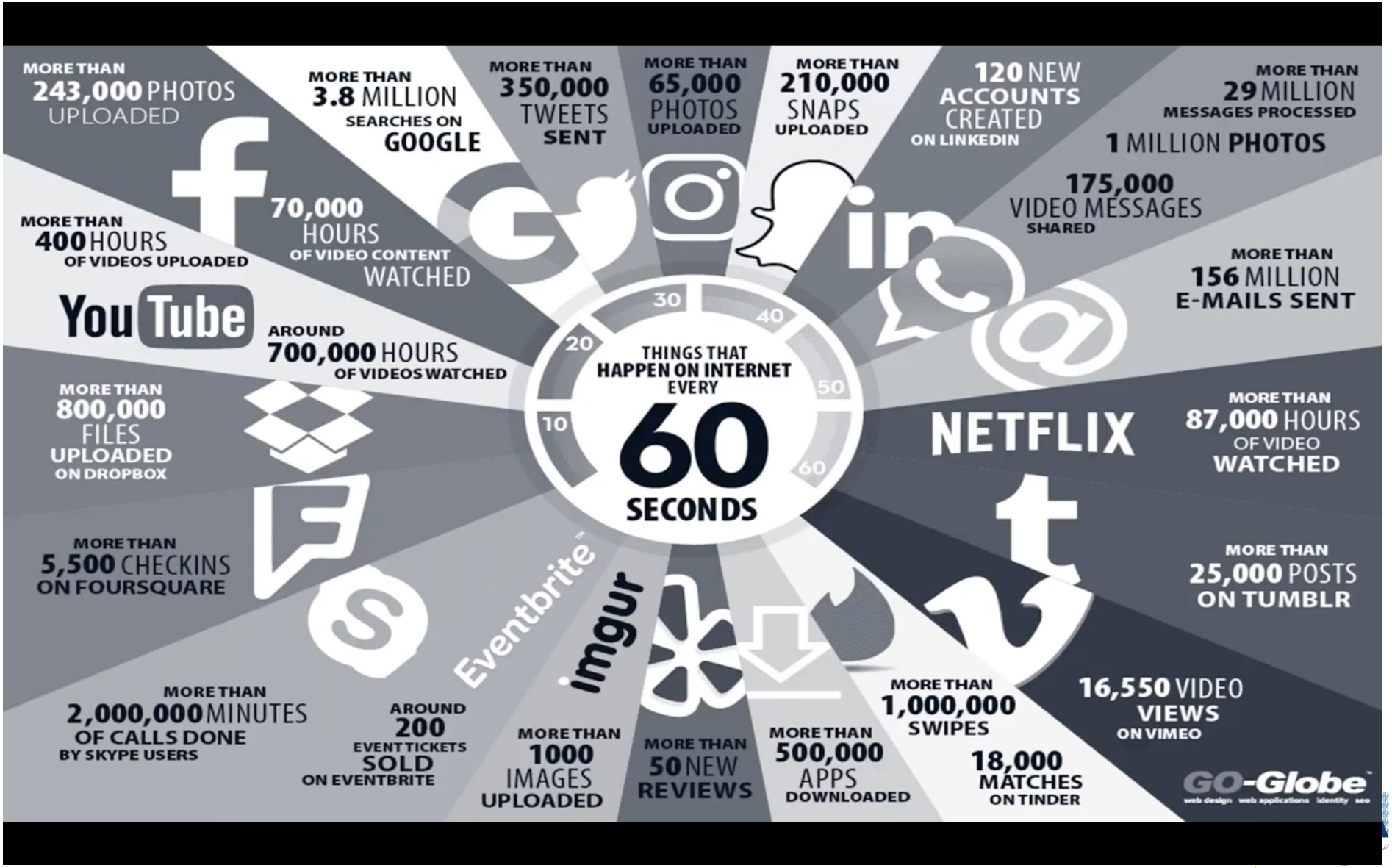
Outline

UNIVERSITÉ DE NANTES

Introduction

The Internet

Introduction



Introduction

- **Port based traffic classification:** is the most simple technique since an analysis of the packet header is used to identify only the port number and its correspondence to the well-known port numbers.

Applications can use dynamic port number or ports associated with other protocols to hide from network security tools.

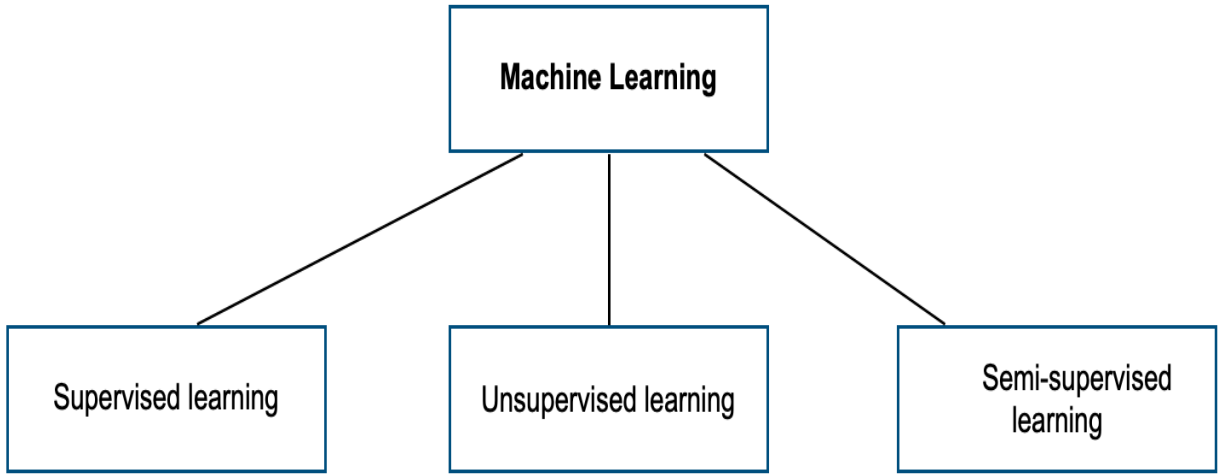
- **Deep Packet Inspection (DPI):** has been proposed to inspect the payload of the packets searching for patterns that identify the application.

It checks all packets data, which consumes a lot of CPU resources and can cause a scalability problem.

UNIVERSITÉ DE NANTES

Introduction

- **ML-based traffic classification**
  Many research works have already used ML methods in network application classification in order to avoid the limitation of DPI and port-based traffic classification.



UNIVERSITÉ DE NANTES

Motivation

- There may exist a large number of unknown traffic within the dataset.
- As new applications emerge every day, it is not possible to have all the flow labeled in a real-time manner.

Source IP in network format	Source port number	Destination IP in network format	Destination port number	Web service detected nDPI (e.g., Facebook, WhatsApp, Google, etc
192.168.128.3 6%		172.16.255.200 27%		Google
192.168.122.52 2%		172.16.255.183 5%		DNS
Other (2506126) 93%	0 65.5k	Other (1815175) 67%	0 65.5k	Other (1715096)
192.168.121.2	49209	172.16.141.247	2222	Unknown
192.168.121.2	49210	172.16.141.247	2222	Unknown
192.168.121.3	50052	172.16.141.250	2222	Unknown
192.168.121.3	50052	172.16.141.250	2222	Unknown
192.168.121.3	50052	172.16.141.250	2222	Unknown
192.168.121.3	50053	172.16.141.250	2222	Unknown
192.168.121.3	50053	172.16.141.250	2222	Unknown
192.168.121.3	50053	172.16.141.250	2222	Unknown
192.168.121.3	50071	172.16.141.250	2222	Unknown
192.168.121.3	50071	172.16.141.250	2222	Unknown
192.168.121.3	50072	172.16.141.250	2222	Unknown
192.168.121.3	50072	172.16.141.250	2222	Unknown

NIVERSITÉ DE NANTES

Motivation

- Semi-supervised learning is a combination of supervised and unsupervised approaches and is used when the dataset consists of input-output pairs but the outputs values are not known for certain observations.

$\Longrightarrow$ **Reflects the situation of most of the network datasets.**

UNIVERSITÉ DE NANTES

Our contribution

- Takes advantage of **both labeled and unlabeled data** to implement a classification task. Making use of unlabeled data is of significance for the network-traffic classification.
- **Extracts robust features automatically** without the need for an expert to extract features manually.

UNIVERSITÉ DE NANTES

Outline

UNIVERSITÉ DE NANTES

Semi-supervised traffic classification

- Developed a semi-supervised classification method for traffic classification. It consists of **unsupervised feature extraction** task and **supervised classification task.**
- Both unlabeled and labeled data have been used to extract more valuable information and make a better classification.

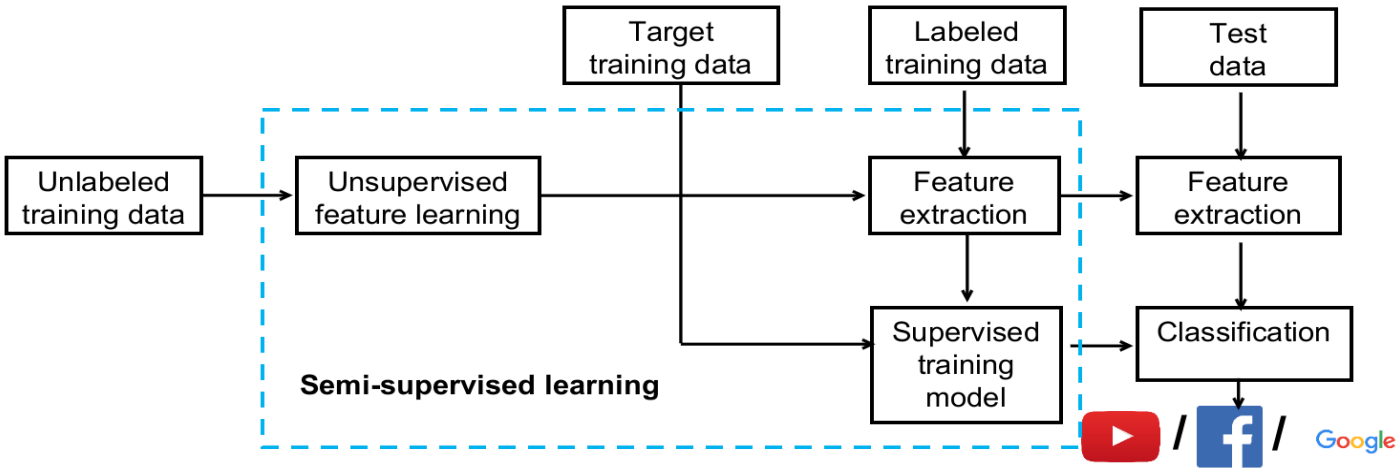


Figure: Structure of the semi-supervised network traffic classification model

UNIVERSITÉ DE NANTES

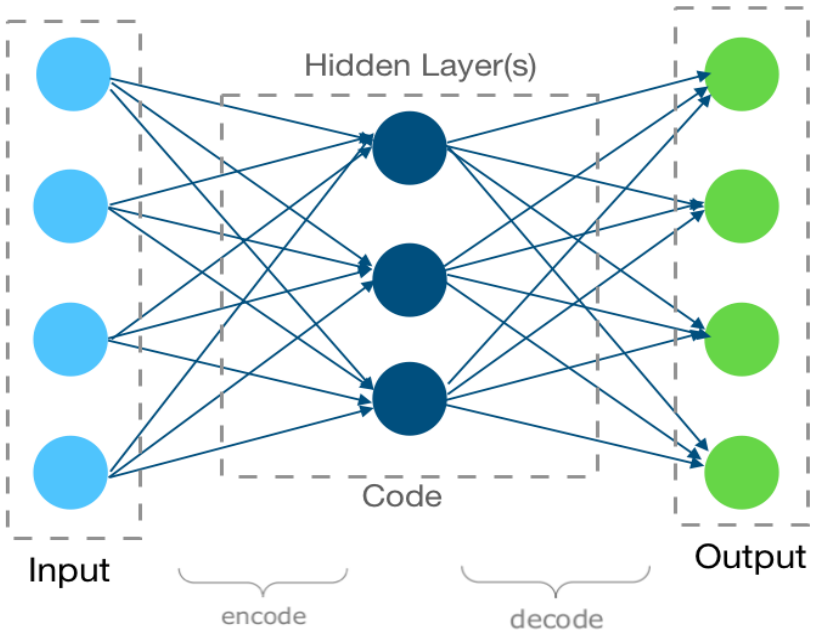
Semi-supervised traffic classification

- SSAE as semi-supervised classification method for traffic classification.
- To improve the performance of classification and lead to learn more robust and informative features with minimal risk of over-fitting we integrate **dropout** and **denoising code hyper-parameters** into our model.

UNIVERSITÉ DE NANTES

Semi-supervised traffic classification

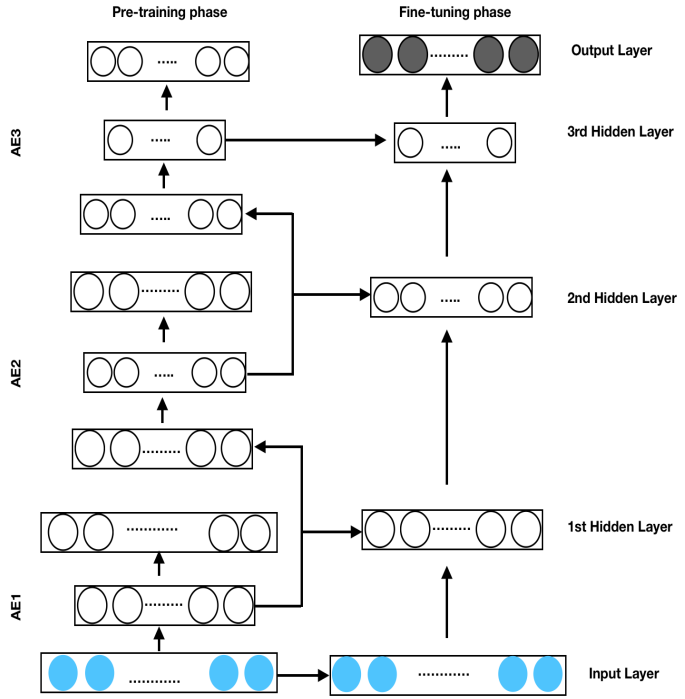
AutoEncoder is an unsupervised learning algorithm and can be divided into three parts (encoder, code, and decoder blocks). More specifically, the encoder obtains the input and converts it into an abstraction, which is generally known as a code, then the input can be reconstructed from the code layer through the decoder. It uses non-linear hidden layers to perform dimensionality reduction.



UNIVERSITÉ DE NANTES

Why SSAE?

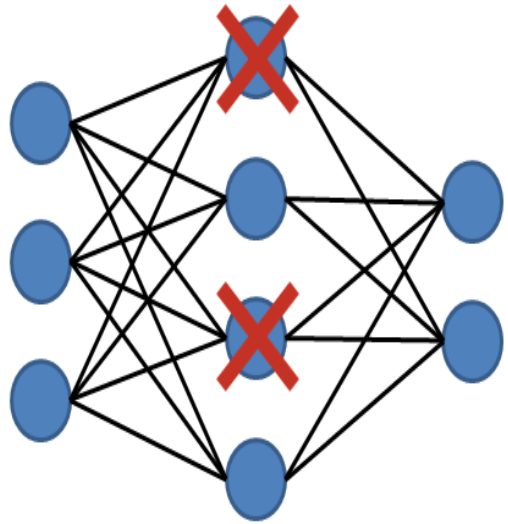
- The **layer-wise pre-training** helps the deep neural network models to yield much better local initialization than a random initialization.
- The **global fine-tuning** process optimizes the parameters of the entire model, which greatly improves the classification task.
- The **sparse constraint** on the hidden layers can help to capture high-level representations of the data.



UNIVERSITÉ DE NANTES

Dropout hyper-parameters

- It is a technique that aims to help a neural network model to **learn more robust features and reduces the interdependent learning** among the neurons.
- It removes units (i.e., neurons) from the network, along with all its incoming and outgoing connections. The choice of which units to drop is random.



UNIVERSITÉ DE NANTES

Denoising autoencoder

- It was proposed to improve the **robustness of feature representation.**
- It is trained to **reconstruct a clean input from a corrupted version** of it in order to extract more relevant features.
- This corruption of the data is done by first corrupting the initial input X to get a partially destroyed version X'.

UNIVERSITÉ DE NANTES

Outline

UNIVERSITÉ DE NANTES

Dataset

It was collected in a network section from Universidad Del Cauca, Popayàn, Colombia. It was constructed by performing packet captures at different hours, during the morning and afternoon over six days in 2017. However, we have used only the traffic collected from one day, which is 09/05/2017.

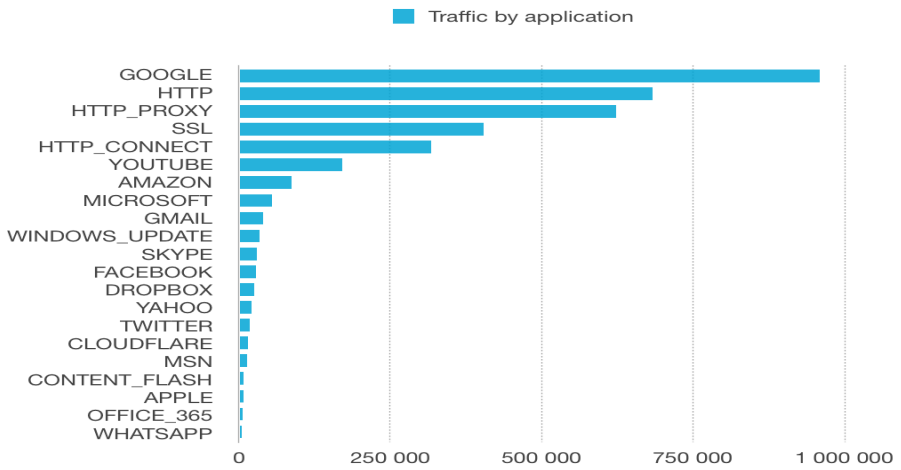
Number of features	**87**
Number of instances	**404,528**
Label	**Name of application**
Application	**54**
Labeled data	**283,186 (70%)**
Unlabeled data	**121,342 (30%)**

UNIVERSITÉ DE NANTES

Dataset

Detail | Compact | Column — 7 of 87 columns

Source.Port The source port number	Destination.IP The destination IP address.	Destination.Port The destination port number.	Timestamp The instant the packet was captured stored in the next date format: Dd/mm/yyyy HH:MM:SS	Flow.Duration The total duration of the flow	ProtocolName This attribute is the objective class of the dataset. It holds the application name following the code
0 — 65.5k	10.200.7.8 9% 10.200.7.7 9% Other (2939141) 82%	0 — 65.5k	41915 unique values	1 — 120m	GOOGLE 27% HTTP 19% Other (1934452) 54%
52680	10.200.7.5	3128	26/04/201711:11:18	405528	YOUTUBE
44858	216.58.202.225	443	26/04/201711:11:18	346773	YOUTUBE
52677	10.200.7.5	3128	26/04/201711:11:18	407686	YOUTUBE
3128	192.168.42.31	52680	26/04/201711:11:18	12	HTTP
443	10.200.7.195	44857	26/04/201711:11:18	216013	YOUTUBE
3128	192.168.42.31	52680	26/04/201711:11:18	479	HTTP
3128	192.168.42.31	52677	26/04/201711:11:18	47	HTTP
3128	192.168.42.31	52677	26/04/201711:11:18	97	HTTP
52681	10.200.7.5	3128	26/04/201711:11:18	407276	YOUTUBE



UNIVERSITÉ DE NANTES

Model architecture

- In the experiment, we separate the labeled data into training **(80%)**, validation **(10%)**, and testing **(10%).**

SSAE Model	#Hidden layers	Number of neurons					Test accuracy
		L1	L2	L3	L4	L5	
SSAE 1	2	100	50	-	-	-	79.4%
SSAE 2	2	100	100	-	-	-	83%
SSAE 3	2	100	200	-	-	-	84.2%
SSAE 4	2	100	400	-	-	-	82%
SSAE 5	2	70	50	-	-	-	77.4%
SSAE 6	2	70	30	-	-	-	76.8%
SSAE 7	3	100	200	50	-	-	85.4%
SSAE 8	3	100	200	100	-	-	85.7%
SSAE 9	3	100	200	200	-	-	85.3%
SSAE 10	3	200	400	60	-	-	85.2%
SSAE 11	3	100	200	400	-	-	85.8%
SSAE 12	4	100	200	400	100	-	86.3%
SSAE 13	**4**	**100**	**200**	**400**	**50**	-	**86.89%**
SSAE 14	5	100	200	400	50	30	84.8%

UNIVERSITÉ DE NANTES

The effect of dropout & denoising rate

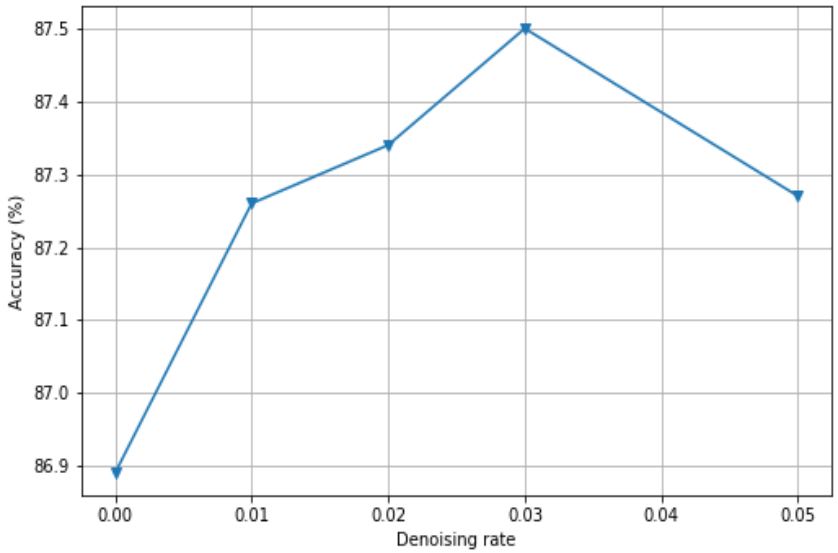


Figure: Effect of denoising coding

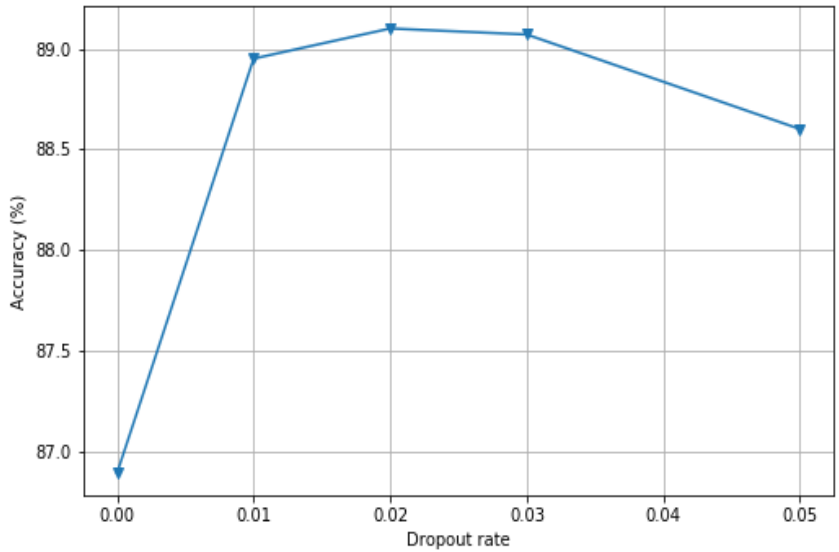


Figure: Effect of dropout

UNIVERSITÉ DE NANTES

The effect of dropout & denoising rate

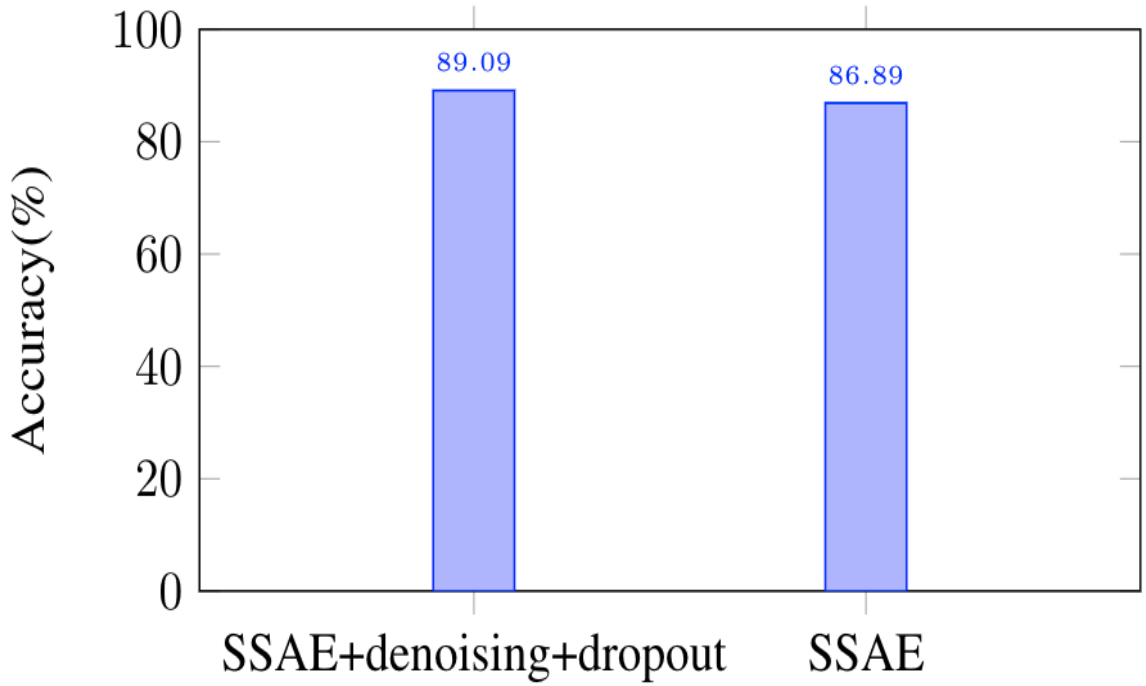


Figure: Accuracy of our model with/without enforcement (dropout/denoising).

UNIVERSITÉ DE NANTES

Comparison of ML classification results

Model	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
SSAE+RF	87.13	88.54	87.13	87.49
SSAE+SVM	55	63.22	55	56.79
SSAE+DT	84.37	86.60	84.37	85.13
Our model	**89.09**	**89.51**	**88.35**	**89.05**

UNIVERSITÉ DE NANTES

Outline

UNIVERSITÉ DE NANTES

Conclusion

- We have used supervised and unsupervised learning for network traffic classification.

UNIVERSITÉ DE NANTES

Conclusion

- We have used supervised and unsupervised learning for network traffic classification.
- To improve the performance of the feature extracted through our model and to avoid over-fitting, we injected dropout and denoising code hyper-parameters.

UNIVERSITÉ DE NANTES

Conclusion

- We have used supervised and unsupervised learning for network traffic classification.
- To improve the performance of the feature extracted through our model and to avoid over-fitting, we injected dropout and denoising code hyper-parameters.
- For future works, we plan to use a much larger amount of unlabeled data to verify its impact on the classification performance.

UNIVERSITÉ DE NANTES

THANK YOU FOR YOUR ATTENTION!

UNIVERSITÉ DE NANTES