

UNIVERSITY OF TECHNOLOGY SYDNEY  
Faculty of Engineering and Information Technology

**Classifying Encrypted WiFi Traffic Using Deep  
Learning Methods**

by

**Ying Li**  
Supervisor: **Dr. Christy Jie Liang**

THESIS FOR  
**Doctor of Philosophy**

Sydney, Australia

2022

# Certificate of Original Authorship

I, Ying Li declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Computer Science, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:  
Signature removed prior to publication.

Date: 1/12/2022

# ABSTRACT

## Classifying Encrypted WiFi Traffic Using Deep Learning Methods

by

Ying Li

Supervisor: Dr. Christy Jie Liang

In this thesis, the goal is to classify encrypted WiFi traffic using deep learning methods.

1) Firstly, we investigate the possibility of making useful inferences from passively observed WiFi traffic that is encrypted at both the transport layer as well as the MAC layer. This is more challenging in comparison to making predictions from the IP layer traffic due to the lack of any meta information. We identify content from encrypted network traffic flows using video streaming as an example because videos are highly popular on the Internet and are frequently misused in many ways including the distribution of fake news, hate speech, and radical and propaganda content. Besides, in network protection and situational awareness applications, there is a strong need to identify whether certain known videos are being watched, either by certain individuals or in a certain area. In the first work, we create a video wireless traffic dataset that contains 10 YouTube videos collected at the WiFi layer. And we demonstrate the possibility of identifying video content using different deep learning models. We do experiments on this traffic dataset and show that our model can achieve a good performance. Besides, we evaluate the longevity of our classifier by making predictions two weeks apart. The results of this work will be further elaborated in Chapter Three.

2) Secondly, not limited to video streaming, other types of traffics(e.g. web and

audio streaming) need to do classification due to the purpose of service management. However, only a limited amount of work has looked into the possibility of building a generic traffic classifier that can handle different classes of traffic. we show that encrypted WiFi traffic fingerprinting can be generalized and applies to many common internet traffic types such as web, video streaming, and audio streaming. In this work, we expand our video wireless traffic dataset to a general wireless traffic dataset that includes web, video streaming, and audio streaming. And we propose a novel hierarchical classifier that can make coarse-grained predictions (e.g. web, video, or audio) as well as fine granular predictions (e.g. content providers/platforms and exact content). Moreover, this approach allows us to estimate network usage characteristics for the purpose of service management in large networks and also identify unknown service providers for different traffic classes. This is explained in detail in Chapter Four.

3) Finally, we investigate how to generate WiFi traffic samples by category automatically. A high-quality, high-volume dataset is very important for the deep learning-based classifier. Specific to the network domain, the classifier is sensitive to the dataset. For example, the network environment of an individual and an enterprise is different in terms of network transfer speed and network configures. Besides, data collection is time-consuming. Therefore, a generator that can generate samples by category automatically is needed. There are many existing generative models. But, the labeled data is required when they generate samples by category. In this work, we propose two novel generative models, namely infinite Gaussian mixture auto-encoder(IGMVAE) and the infinite mixture of infinite Gaussian mixture auto-encoder ( $I^2GMVAE$ ). IGMVAE is a variant of variational auto-encoder(VAE) with an infinite Gaussian Mixture model (IGMM) as the prior distribution of the latent variables.  $I^2GMVAE$  is a variant of VAE with the infinite mixture of infinite Gaussian Mixture model ( $I^2GMM$ ) as the prior distribution of the latent variables. They are explained in detail in Chapter Five.

## Acknowledgements

I would like to thank the following people. They give me great help and support during my Ph.D. journey.

First and foremost I wish to thank my supervisor, Dr. Christy Jie Liang. It is my great luck to meet her as my supervisor. Christy is very professional and kind. She not only inspired me academically but also help me a lot in life. Her good personality will affect me in my future life.

I would like to thank my co-supervisor, Prof. Richard Yida Xu. Richard is not only a professional scholar but also a person who loves to share knowledge. He gives us weekly courses on machine learning and computer technology. The most amazing thing is that he can let us learn a lot of complicated knowledge in an easy-to-understand way. His hardworking and kindness will benefit me for life.

I would also like to thank my co-supervisor, Prof. Massimo Piccardi. He helped me a lot. Thanks so much for your patience and kindness.

I also want to thank Dr. Junyu Xuan. Junyu is a very professional and responsible scholar. We discussed the innovations, model details, and experiments together. It is my pleasure to have a such good partner in the past few years.

I would also like to thank my research partners, including Suranga Seneviratne, Guillaume Jourjon, Adriel Cheng, Darren Webb, Kanchana Thilakarathna, and David B. Smith. We share knowledge and discuss experiment details every week. Thank you for giving me an unforgettable and rewarding project experience.

I am grateful to all lab mates, including Yi Huang, Shuai Jiang, Wanming Huang, Haodong Chang, Caoyuan Li, Xuan Liang, Jason Traish, Ziyue Zhang, Wei Huang,

Chen Deng, Congzhentao Huang, Chris Markos, Yunce Zhao, etc. We shared knowledge and hiked together on weekend. Because of those mates, I had a pleasant and meaningful time in Sydney.

Thanks to the CA panel Prof. Guandong Xu and Prof. Andrew Zhang.

Sincerely, I want to thank my parents, my sister, my husband, and my child. Because of your support and encouragement, I can concentrate on my research and move forward.

Thanks again to everyone who has helped me!

# List of Publications

## Conference Papers

- C-1. **Ying Li**, Yi Huang, Suranga Seneviratne, Kanchana Thilakarathna, Adriel Cheng, Guillaume Jourjon, Darren Webb and Richard Yi Da Xu, “Deep Content: Unveiling Video Streaming Content from Encrypted WiFi Traffic”, *17th Int. Symp. on Network Computing and Applications – NCA 2018*. (**CORE A conference**)

## Journal Papers

- J-1. **Ying Li**, Yi Huang, Suranga Seneviratne, Kanchana Thilakarathna, Adriel Cheng, Guillaume Jourjon, Darren Webb, David B. Smith and Richard Yi Da Xu, “From Traffic Classes to Content: A Hierarchical Approach for Encrypted Traffic Classification”, *Computer Networks* (**CORE B journal**)
- J-2. **Ying Li**, Junyu Xuan, Yi Huang, Christy Liang and Richard Yida Xu, “Infinite Gaussian Mixture Autoencoders for Data Generation”, *Transactions on Image Processing* (**ready to submit**)
- J-3. Yi Huang, **Ying Li**, Timothy Heyes, Guillaume Jourjon, Adriel Cheng, Suranga Seneviratne, Kanchana Thilakarathna, Darren Webb and Richard Yi Da Xu, “Probability Based Task Adaptive Siamese Open-Set Recognition for Encrypted Network Traffic With Bidirectional Dropout Data Augmentation”, *Pattern Recognition Letters* (**CORE B journal**)
- J-4. Yi Huang, **Ying Li**, Guillaume Jourjon, Suranga Seneviratne, Kanchana Thilakarathna, Adriel Cheng, Darren Webb and Richard Yi Da Xu, “CRAAE:

Calibrated Reconstruction Based Adversarial AutoEncoder Model for Novelty Detection”, *Pattern Recognition Letters* (**Under Review, CORE B**)



# Contents

Certificate	ii
Abstract	iii
Acknowledgments	v
List of Publications	vii
List of Figures	xiii
List of Tables	xv
Abbreviation	1
<b>1 Introduction</b>	<b>2</b>
1.1 Background and Motivation . . . . .	2
1.1.1 Encrypted WiFi Traffic Analysis . . . . .	2
1.1.2 Deep Learning based Traffic Classification Technology . . . . .	3
1.1.3 Hierarchical Architecture for Traffic Classification . . . . .	3
1.1.4 Data Augmentation Technology . . . . .	4
1.2 Research Objectives . . . . .	4
1.3 Research Contributions . . . . .	7
1.4 Thesis Structure . . . . .	8
<b>2 Literature Review</b>	<b>11</b>
2.1 Network Traffic Classifier . . . . .	11

2.1.1	Network Traffic Classifier Foundation . . . . .	11
2.1.2	Traffic Classification on HTTPS Communications . . . . .	14
2.1.3	Traffic Classification in WiFi and Physical Layers . . . . .	16
2.1.4	Hierarchical Traffic Classifier . . . . .	17
2.1.5	Data Augmentation Methods for Network Traffic Classification	18
2.2	Related Deep Learning Techniques . . . . .	19
2.2.1	Deep Neural Networks . . . . .	19
2.2.2	Deep Generative Models . . . . .	21
2.2.3	Bayesian Nonparametric Models . . . . .	24

### **3 Classifying Encrypted WiFi Videos Using Deep Learning Models** **27**

3.1	Introduction . . . . .	27
3.2	Method . . . . .	28
3.2.1	DASH Streaming . . . . .	28
3.2.2	Preprocessing & Feature Engineering . . . . .	29
3.2.3	Classifier Architectures . . . . .	31
3.3	Experiments and Results . . . . .	33
3.3.1	Dataset and Evaluation Metric . . . . .	33
3.3.2	Implementation Details . . . . .	35
3.3.3	Performance . . . . .	36
3.3.4	Performance Analysis . . . . .	39
3.4	Summary . . . . .	42

<b>4</b>	<b>Classifying Encrypted WiFi Traffic Using A Hierarchical Classifier</b>	<b>43</b>
4.1	Introduction . . . . .	43
4.2	Method . . . . .	44
4.2.1	Streaming and Other Time Sensitive Traffic . . . . .	44
4.2.2	Architecture . . . . .	45
4.2.3	Training Process . . . . .	47
4.3	Experiments and Results . . . . .	47
4.3.1	Dataset . . . . .	47
4.3.2	Evaluation Metrics . . . . .	49
4.3.3	Implementation Details . . . . .	52
4.3.4	Results . . . . .	52
4.3.5	Result Analysis . . . . .	53
4.4	Summary . . . . .	57
<b>5</b>	<b>Generating Samples by Category Using Bayesian Non-parametric Autoencoders</b>	<b>58</b>
5.1	Introduction . . . . .	58
5.2	IGMVAE . . . . .	60
5.2.1	Method . . . . .	60
5.2.2	Inference Process . . . . .	61
5.2.3	Architecture . . . . .	65
5.2.4	Training and Testing . . . . .	66

5.3	I <sup>2</sup> GMVAE . . . . .	66
5.3.1	Method . . . . .	67
5.3.2	Inference Process . . . . .	69
5.3.3	Architecture . . . . .	71
5.3.4	Training and Testing . . . . .	72
5.4	Experiments and Results . . . . .	73
5.4.1	Dataset and Evaluation Metrics . . . . .	73
5.4.2	Implementation Details . . . . .	74
5.4.3	Results . . . . .	74
5.4.4	Result Analysis . . . . .	78
5.5	Summary . . . . .	78
<b>6</b>	<b>Conclusions and Future Work</b>	<b>80</b>
6.1	Conclusions . . . . .	80
6.2	Future Work . . . . .	81

# List of Figures

1.1	Thesis structure . . . . .	10
2.1	Literature review section structure . . . . .	11
2.2	Illustration of stick-breaking construction . . . . .	26
3.1	I/O graphs of different traffic flows for the same video . . . . .	29
3.2	I/O graphs of a single run for 10 different videos . . . . .	32
3.3	Architecture of different models . . . . .	34
3.4	Data collection setup . . . . .	35
3.5	Accuracy of various neural network models . . . . .	38
3.6	Classification performance confusion matrix for models with F1 (number of packets (data) on down-link) . . . . .	40
3.7	T-SNE embedding of the last layer . . . . .	41
4.1	Hierarchical classifier architecture . . . . .	46
4.2	Classification performance confusion matrix for hierarchical models with F6-combination . . . . .	54
4.3	I/O graphs of the same Stan video on different runs . . . . .	55

4.4	I/O graphs of different HTTP traffic types on different content providers. . . . .	55
5.1	Probabilistic graphical models for the IGMVAE(left) and I <sup>2</sup> GMVAE(right). . . . .	61
5.2	The neural network architecture of IGMVAE. . . . .	66
5.3	The neural network architecture of I <sup>2</sup> GMVAE. . . . .	72
5.4	Cluster number decision curve of IGMVAE on YouTube. . . . .	75
5.5	Generated YouTube videos I/O graph of IGMVAE. . . . .	76
5.6	Cluster number decision curve of IGMVAE for the incremental experiment (The blue color line is based on video 0 to 6. The orange color line is based on video 0 to 9). . . . .	77
5.7	Visualisation of the latent variables on YouTube: (a) GMVAE learns the latent variables with cluster number 5. (b) GMVAE learns the latent variables with cluster number 10. (C) GMVAE learns the latent variables with cluster number 25. (d) the latent variables of IGMVAE(our). . . . .	77

# List of Tables

3.1	Feature selection from wireless traffic data . . . . .	30
3.2	Feature evaluation for CNN model . . . . .	36
3.3	Feature evaluation for LSTM model . . . . .	37
3.4	Feature evaluation for MLP model . . . . .	38
3.5	MLP model accuracy with new dataset two weeks later . . . . .	39
4.1	Hierarchical model results on dataset 1 (in percent) . . . . .	50
4.2	Hierarchical model results on dataset 2 (in percent) . . . . .	51
5.1	IGMVAE and I <sup>2</sup> GMVAE results (on YouTube) . . . . .	75

## Abbreviation

TLS - Transport Layer Security  
WPA2 - WiFi Protected Access 2  
VAE - Variational Auto-encoder  
IGMVAE - Infinite Gaussian Mixture Auto-encoder  
I<sup>2</sup>GMVAE - Infinite Mixture of infinite Gaussian Mixture Auto-encoder  
IGMM - Infinite Gaussian Mixture Model  
I<sup>2</sup>GMM - Infinite Mixture of Infinite Gaussian Mixture Model  
HTTPS - Hypertext Transfer Protocol Secure  
ML - machine learning  
DL - Deep learning  
MLP - Multi-Layer Perceptron  
CNN - Convolutional Neural Network  
RNN - Recurrent Neural Network  
LSTM - Long Short-term Memory  
ABC - Australian Broadcasting Corporation  
SMH - Sydney Morning Herald  
GANs - Generative Adversarial Nets  
TCP -Transmission Control Protocol  
P2P - Peer-to-peer Internet  
AAE - Adversarial Auto-encoder  
DASH - Dynamic Adaptive Streaming over HTTP  
HAS - HTTP based Adaptive Streaming  
GMM - Gaussian Mixture Model  
GMVAE - Gaussian Mixture Auto-encoder  
BNP - Bayesian Nonparametric  
ELOB - Evidence Lower Bound