Computer Science and Technology
Research Students' Symposium

19-21 May 2021

Programme

Friday 21 May

9:15AM–9:30AM · Stage
Dhruv Makwana
Translation Validation for Optimised Binaries

9:30AM–9:45AM · Stage
Chi Ian Tang
Improving Human Activity Recognition through Self-training with Unlabelled Data

9:45AM–10:00AM · Stage
Smita Vijaya Kumar
Packing Compute Resources using Decentralized Schedulers for Effective Cloud Scale Scheduling

10:00AM–10:15AM · Stage
Vadim Safronov
Towards distributed and protocol-independent IoT interoperation in smart spaces

10:15AM–10:30AM · Stage
Jack Hughes
Using machine learning and big data approaches to understand the role of gaming within cybercrime pathways on underground discussion platforms

10:45AM–11:15AM · Stage
Tea break

11:15AM–11:30AM · Stage
Felix Opolka
Bayesian Link Prediction with Deep Graph Convolutional Gaussian Processes

11:30AM–11:45AM · Stage
Ryan Kortvelesy
A Modular Graph Neural Network Framework

11:45AM–12:00PM · Stage
Cristian Bodnar
Topological Representation Learning

12:00PM–12:15PM · Stage
Edgaras Liberis
Deep learning in resource-constrained environments

12:15PM–12:30PM · Stage
Tiago Pimentel Martins Da Silva
An Informative Exploration of the Lexicon

12:30PM–1:30PM · Stage
Lunch

1:30PM–2:30PM · Sessions
- Machine learning and Artificial Intelligence theme meeting
- Programming Languages, Semantics and Verification theme meeting
- Systems and Networking theme meeting

2:30PM–2:45PM · Stage
Yuxiao (Sean) Ye
Argument Mining with Real-world text

2:45PM–3:00PM · Stage
Dan Andrei Iliescu
Representing Grouped Data

3:00PM–3:15PM · Stage
Josef Valvoda
Robust Legal Reasoning

3:15PM–3:30PM · Stage
Paris Flood
Practical Minimal Description Lengths of Neural Networks

3:30PM–3:45PM · Stage
Mahwish Arif
Cinnamon: A Domain-Specific Language for Binary Profiling and Monitoring

3:45PM–4:00PM · Stage

Mahwish Arif
Let's talk girls - About the Women@CL initiative

4:00PM–4:15PM · Stage
Tea break

4:15PM–5:15PM · Networking
Happy hour chat
Chat roulette: grab a glass of something wet and join our research students for a random chat with someone interesting

5:15PM–5:25PM · Stage
End day 3
Thank you and good night

Dhruv Makwana *supervised by Dr N. Krishnaswami*

Translation Validation for Optimised Binaries

Translation validation is a way of ensuring compilation correctness without relying on the correctness of the compiler. That is, instead of relying on the compiler outputting the correct program for all possible input programs , it instead check whether it's output the correct program for a particular program. In this talk, I'll be discussing some work I did over the past year on this problem.

Chi (Ian) Tang *supervised by Prof. C. Mascolo*

Improving Human Activity Recognition through Self-training with Unlabelled Data

Machine learning and deep learning have shown great promise in mobile sensing applications, including Human Activity Recognition. However, the performance of such models in real-world settings largely depends on the availability of large datasets that captures diverse behaviours, which is costly and difficult to obtain. In the talk, I present SelfHAR, a semi-supervised model that effectively learns to leverage unlabelled mobile sensing datasets to complement small, labelled datasets. The approach combines teacher-student self-training, which distils the knowledge of unlabelled and labelled datasets while allowing for data augmentation, and multi-task self-supervision, which learns robust signal-level representations by predicting distorted versions of the input. SelfHAR was evaluated on various HAR datasets and showed state-of-the-art performance over supervised and previous semi-supervised approaches. Furthermore, SelfHAR is data-efficient, reaching similar performance using up to 10 times less labelled data compared to supervised approaches.

Smita Vijaya Kumar *supervised by Dr E. Kalyvianaki and Dr A. V. S. Madhavapeddy*

Packing Compute Resources using Decentralized Schedulers for Effective Cloud Scale Scheduling

Our research proposes Murmuration, a fully decentralized datacenter scheduler that aims to increase CPU utilization and minimize job completion times. In Murmuration, scheduling decisions are taken independently by each scheduler in a loosely coordinated manner based on information exchanged among the schedulers. Scheduling at each node is performed using a loosely consistent global view of the data center resources which is made available at every scheduler. Increased CPU utilization is achieved by packing tasks on unused gaps in task queues on worker nodes. Job completion times are minimized by placing tasks on the workers with the smallest wait times.

Vadim Safronov *supervised by Dr R. M. Mortier*

Towards distributed and protocol-independent IoT interoperation in smart spaces

In the next couple of decades every smart building will contain thousands of heterogeneous IoT devices which sense, measure and actuate things and work on different protocols. Current centralised management platforms (aka BMSs) are not ready for that large amount of new heterogeneous IoT traffic and, therefore, will be a source of overloads and unacceptable delays if that integration happens. Most current smart space application protocols assume the presence of the IP layer for most devices, thus making non-IP IoT protocols, such as LPWANs and LoWPANs, incompatible or only partly compatible with that services.

To prevent that near-future interoperation problems, I am exploring a decentralised way for smart space automation, where, in order to enforce management workflows, IoT devices interoperate directly, without the regular need of going through the central platform. The focus of my PhD work is the design, development and evaluation of the distributed and protocol-independent IoT interoperation model for smart spaces. The model is inspired by the Plutarch protocol-neutral internetworking architecture [1] which is tested in practice and supplemented with new design choices and deployment recommendations for smart-space use cases in order to perform the automation more efficiently compared to the legacy centralised operation options.

[1] Jon Crowcroft, Steven Hand, Richard Mortier, Timothy Roscoe, and Andrew Warfield. 2003. Plutarch: an argument for network pluralism. SIGCOMM Comput. Commun. Rev. 33, 4 (October 2003), 258–266. DOI:https://doi.org/10.1145/972426.944763

Jack Hughes *supervised by Dr A. J. Hutchings*

Using machine learning and big data approaches to understand the role of gaming within cybercrime pathways on underground discussion platforms

Cybercrime forums provide a place for members of varying skill levels to exchange knowledge and tools, and have been well studied in the literature: from social network analysis of forum communities, to measurements of marketplaces on forums, and the use of natural language processing (NLP). Most research use a static snapshot of data, missing the time-series nature of these forums. Longitudinal analysis has not been well studied, which can give us an insight into the evolution of forums and their users. In addition, while static snapshots reducing complexity to support off-the-shelf analysis tools, they assume forum activities do not significantly change over time, or limit analysis to one time window. Typically, research has focused on specific topics found by manual inspection of forums, from small datasets collected over limited timeframes. This work uses on the CrimeBB dataset from the Cambridge Cybercrime Centre, with scrapes of multiple forums spanning up to 10 years, supporting longitudinal approaches. Observing and modelling forums at scale, using large datasets, allows the comparison of behaviours within cybercrime discussion platforms, to understand how groups operate and change over time. This leads to the research question of whether gaming has an impact on groups becoming interested in cybercrime.

Felix Opolka *supervised by Prof. P. Liò*

Bayesian Link Prediction with Deep Graph Convolutional Gaussian Processes

Link prediction aims to reveal missing edges in a graph. We introduce a deep graph convolutional Gaussian process model for this task, which addresses recent challenges in graph machine learning with oversmoothing and overfitting. Using simplified graph convolutions, we transform a Gaussian process to leverage the topological information of the graph domain. To scale the Gaussian process model to larger graphs, we introduce a variational inducing point method that places pseudo-inputs on a graph-structured domain. Multiple Gaussian processes are then assembled into a hierarchy whose structure allows skipping convolutions and thus counteracting oversmoothing. The proposed model represents the first Gaussian process for link prediction that makes use of both node features and topological information. We evaluate our model on multiple graph data sets with up to thousands of nodes and report consistent improvements over state-of-the-art graph neural network approaches.

Ryan Kortvelesy *supervised by Dr A. S. Prorok*

A Modular Graph Neural Network Framework

Recent work in the multi-agent domain has shown the promise of Graph Neural Networks (GNNs) to learn complex coordination strategies. However, most current approaches use minor variants of a Graph Convolutional Network (GCN), which applies a convolution to the communication graph formed by the multi-agent system. In this paper, we investigate whether the performance and generalization of GCNs can be improved upon. We introduce ModGNN, a decentralized framework which serves as a generalization of GCNs, providing more flexibility. To test our hypothesis, we evaluate an implementation of ModGNN against several baselines in the multi-agent flocking problem. We perform an ablation analysis to show that the most important component of our framework is one that does not exist in a GCN. By varying the number of agents, we also demonstrate that an application-agnostic implementation of ModGNN possesses an improved ability to generalize to new environments.

Cristian Bodnar *supervised by Prof. P. Liò*

Topological Representation Learning

How can topology helps us learn better representations for problems where the data has an underlying structure? I will be discussing how graph pooling can be understood as computing homotopy-equivalent spaces and how simplicial complexes can improve the expressive power of graph neural networks.

Edgar Liberis *supervised by Dr N. D. Lane*

Deep learning in resource-constrained environments

We can bring computational intelligence to personal IoT-type devices by employing deep learning. Such devices are typically powered by microcontroller units (MCUs) which are extremely resource-scarce, with orders of magnitude fewer computational resources (RAM, storage, number of cores, etc.) than is typically required for deep learning. We'll look into the difficulties of designing neural networks for such a platform which requires an intricate balance between keeping high predictive performance (accuracy) while achieving low memory and storage usage and inference latency.

Tiago Pimentel Martins Da Silva *supervised by Prof. S. H. Teufel*

An Informative Exploration of the Lexicon

During my PhD I've been exploring the lexicon through the lens of information theory. In this talk, I'll give an overview on results detailing the distribution of information in words (are initial or final positions more informative?), cross-linguistic compensations (if a language has more information per character, are their words shorter?), and on it's general "efficiency" (how far is it from an information-theoretic optimum?).

Yuxiao (Sean) Ye *supervised by Prof. S. H. Teufel*

Argument Mining with Real-world text

Yuxiao's research addresses the question of understanding and analysing the kind of multi-person argumentation that appears on the platform Quora. He has argued extensively that this platform provides an environment that will advance research in Argument Mining (AM) community, particular because humans on these platforms argue informally and because the massively parallel availability of somewhat similar arguments allows for multi-person summarisation.To this end, he plans to build an argument graph representing a large number of opinions and arguments.

Yuxiao has presented at the EACL conference his research on applying end-to-end neural dependency parsing to traditional annotations in AM, with a performance that is surpassing current SoA. His current work extends the simple, one-person, claim and premise-based model currently used, to the multi-person situation. The new annotation model must ensure that topically related sub-arguments are grouped together, while ensuring that within-argument support and attack links are still meaningful. He is currently working out the theoretical properties of this new representation, along with a plan for human annotation and gathering of ideas for the neural architecture such a model could use.

Dan Andrei Iliescu *supervised by Dr D. J. Wischik*

Representing Grouped Data

Deep representations have revolutionised numerous scientific fields through their power to distill information and separate the factors of variation in data. One widely applicable subproblem is learning representations of datasets partitioned into groupings, such that the latent factors common within groups are disentangled from the factors which uniquely characterise each data instance. This framework connects seemingly disparate research topics in Machine Learning, including Fair Representations, Causal Inference, Translation, Domain Adaptation and Sequence Modelling. In my first year, I investigated how to achieve such a separation between Group and Instance factors in the latent

space of a Variational Autoencoder, drawing upon insights from Probabilistic Modelling and Information Theory. I also evaluated my model on a set of diverse tasks which lend themselves to this Group-Instance disentanglement, ranging from synthesising images of objects from novel viewpoints to normalising student exam scores across schools. My current research focuses on extending my model to accommodate sequence data, in order to address more challenging real-world tasks in the medical field, climate forecasting and process analytics.

Josef Valvoda *supervised by Prof. S. H. Teufel*

Robust Legal Reasoning

Legal AI has generated a lot of interest in the past few years. But what are the tasks that should be automated and how can they be automated remains an area of research. In this talk I will discuss legal precedent, the doctrine that drives legal reasoning, and my research on what about the precedent becomes law.

Paris Flood *supervised by Prof. P. Liò*

Practical Minimal Description Lengths of Neural Networks

The relationship between compression and learning algorithms has long been a rich area of research, inviting analysis from a variety of methodologies including the minimum description length (MDL) principle. Recent work has shown that the MDL technique known as prequential coding can achieve extremely tight compression bounds when applied to deep neural networks solving supervised learning tasks. In this paper we introduce the prequential coding optimization problem for finding practical, minimal length versions of these codes. Although the core problem proves to be non-convex, we also show that an equivalent reformulation belongs to a problem class known as ‘difference of convex’ programming, for which there exists a well-developed literature on optimization strategies. Accordingly, we present disciplined, general procedures for minimizing the aforementioned prequential code lengths under a variety of practically relevant constraints.

Mahwish Arif *supervised by Dr T. M. Jones*

Cinnamon: A Domain-Specific Language for Binary Profiling and Monitoring

Binary instrumentation and rewriting frameworks provide a powerful way of implementing custom analysis and transformation techniques for applications ranging from performance profiling to security monitoring. However, using these frameworks to write even simple analyses and transformations is non-trivial. Developers often need to write framework-specific boilerplate code and work with low-level and complex programming details. This not only results in hundreds (or thousands) of lines of code, but also leaves significant room for error.

To address this, we introduce Cinnamon, a domain-specific language designed to write programs for binary profiling and monitoring. Cinnamon's abstractions allow the programmer to focus on implementing their technique in a platform-independent way, without worrying about complex lower-level details. Programmers can use these abstractions to perform analysis and instrumentation at different locations and granularity levels in the binary. The flexibility of Cinnamon also enables its programs to be mapped to static, dynamic or hybrid analysis and instrumentation approaches. As a proof of concept, we target Cinnamon to three different binary frameworks by implementing a custom Cinnamon to C/C++ compiler and integrating the generated code within these frameworks. We further demonstrate the ability of Cinnamon to express a range of profiling and monitoring tools through different use-cases.

Mahwish Arif *Women@CL*

Let's talk girls - About the Women@CL initiative