MULTI-MODALITY CONDITIONED GENERATIVE METHODS

Supervisor: Pietro Lio' (pl219@cam.ac.uk)
Co-supervisor: Helena Andres Terre (ha376@cam.ac.uk)

Generative methods such as Variational Autoencoders and GANs can be trained to learn integrative embeddings from different data modalities. However, little is known about the influence of each modality over the integrated latent space.

In this project you will learn how to build generative models capable to integrate different data types. Moreover, you will explore and interpret the latent space distributions conditioned to the different modalities from partial observations.

Different techniques can be used to establish the mapping of modalities over the latent space, such as perturbation/variational approaches, saliency maps or information/entropy estimation.

By learning such interpretable embeddings, one can generate unseen modalities and explore potential correlations that may not be discovered otherwise due to restricted or a lack of data.

In principle this project will be based on integration of images with significantly different data types such as gene expression or categorical data. We are flexible in terms of the datasets used and application, although originally the implementation would be over clinical data.

LEARNING GENERAL GENERATIVE FUNCTIONS VIA NEAREST NEIGHBOURS

Supervisor: Pietro Lio' (pl219@cam.ac.uk)
Co-supervisor: Helena Andres Terre (ha376@cam.ac.uk)

Generative models such as VAEs learn a set of generative functions that lay in a latent space defined by a set of multivariate Gaussians.

Due to the unsupervised nature of these methods, each individual sample or observation is considered independent, and therefore it is assigned its own particular generative function. Most samples with similar features or characteristics will have similar generative distributions, being close and in local areas within the latent space. However, VAEs are not specifically trained to capture these similarities among samples. Instead their original design aims to minimise the distance to multivariate independent priors, which optimises the disentanglement between the latent components.

In physics, we refer to a degenerate eigenvalue or energy when they have more than one linearly independent eigenvector or multiple energy functions associated.

This project will introduce a perturbation approach to solve the degenerate nature of VAEs embeddings. From the original set of independent generative functions, we will implement a k-Nearest Neighbours constraint to minimize the perturbation term and learn the non-degenerate distributions corresponding to each cluster or family of samples.

This will provide a dataset-independent set of functions that better generalise to unobserved samples - and can be potentially used for transfer learning.

Applications of Algorithmic Information Theory to Learning Algorithms

Supervisor: Pietro Lio'
Co-supervisors: Paris Flood <pdlf3@cam.ac.uk>,Ramon Viñas Torné <rv340@cam.ac.uk>

In 2010, the well-known AI pioneer Marvin Minsky gave the following quote about Algorithmic Information Theory (AIT):

'It seems to me that the most important discovery since Gödel was the discovery [of Algorithmic Information Theory] ... This is a beautiful theory ... Everybody should learn all about that and spend the rest of their lives working on it.'

Minsky's enthusiasm is shared by many other contemporaries including Jürgen Schmidhuber, Marcus Hutter, and Gregory Chaitin (to name a few). But what is AIT and why is it considered such a fundamental and potentially useful tool for AI? At its essence, AIT studies the complexity of mathematical objects by asking, 'what is the length of the shortest computer program that generates the object?' This is called the Kolmogorov complexity of an object. A mathematically rich structure such as pi has a low Kolmogorov complexity (there are simple formulas for pi) while a mathematically random structure has a high Kolmogorov complexity. However, AIT is far more than just Kolmogorov complexity; a particularly interesting topic is the Chaitin Omega number - a noncomputable number (as is Kolmogorov complexity) which if known to k bits can be used to decide the truth of any provable/finitely refutable theorem that can be written in less than k bits.

But what does AIT have to do with AI? Well, in the words of Marcus Hutter ' being able to compress well is closely related to acting intelligently'. AIT defines compression in a more fundamental sense than Shannon's Information Theory by investigating a computable description of an object. In this project we seek to advance this application by exploring classic topics in machine learning such as generalization and initialization through the lens of AIT.

If interested, the following references might be helpful:

1. The most comprehensive textbook written about this subject is 'An introduction to Kolmogorov complexity and its applications' by Li and Vitányi. However, this textbook can be intimidating to newcomers, therefore I would recommend one starts out by reading the following:

2. Chapter 14 (Kolmogorov Complexity) of 'Elements of Information Theory' by Cover and Thomas.

3. Hutter had compiled a list of introductory online material on his website at http://www.hutter1.net/index.htm

4. An enjoyable application paper is: Schmidhuber, Jürgen. "Discovering neural nets with low Kolmogorov complexity and high generalization capability." Neural Networks 10.5 (1997): 857-873.

Causal discovery of biological interactions from omics data *

Supervisor: Pietro Lio' (pl219@cam.ac.uk)
Assistant supervisor: Ramon Viñas Torné <rv340@cam.ac.uk>, Nikola Simidjievski <ns779@cam.ac.uk>,"P.D.L. Flood" <pdlf3@cam.ac.uk>

A strong correlation between the expression of two genes X and Y can be explained by a large number of reasons, including the presence of another gene Z that simultaneously upregulates both of them. The goal of this project is to infer causal relationships between pairs of biological entities (e.g. genes) from one or multiple omics modalities (e.g. transcriptomics or epigenomics). Depending on your interests, the project may focus on link prediction (e.g. some of the relationships are assumed to be known), bivariate causal discovery, or multivariate causal discovery (the full biological network is unknown).

During this project you will learn: bioinformatics, causal inference, and machine learning. The ideal candidate for this project will have a strong background in mathematics and statistics as well as good Python programming skills.

Suggested readings (more material can be provided):

1. Elements of Causal Inference (free pdf).
https://mitpress.mit.edu/books/elements-causal-inferenc
2. Causal Inference book (free pdf).
https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/
3. Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks.
https://jmlr.org/papers/v17/14-518.html
4. Comparative Analysis of Gene Regulatory Networks: From Network Reconstruction to Evolution.
https://www.annualreviews.org/doi/abs/10.1146/annurev-cellbio-100913-012908
5. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1810318/
6. Kolmogorov Regularization for Link Prediction. https://arxiv.org/abs/2006.04258
7.Structural Agnostic Modeling: Adversarial Learning of Causal Graphs.
https://arxiv.org/abs/1803.04929

Improving Neural network models for dynamic graphs

Supervisor: Pietro Lio' (pl219@cam.ac.uk)
cosupervisor: Felix Opolka

Description:

Continuous-time dynamic graphs are the most general form of dynamic graphs. Events such as the insertion (or deletion) of edges and nodes or a change in node or edge features can appear at any point in time (rather than at discrete time steps like in discrete-time dynamic graphs).

The current state-of-the-art architecture for machine learning on dynamic graphs is the Temporal Graph Network (TGN) [2]. The broad goal of this project is to further improve the model and adapt it to be suitable for new tasks. Potential directions are:

* Time Prediction: one of the core machine learning tasks on dynamic graphs is link prediction: Given a future timestamp and a pair of nodes, the network outputs whether a link exists between the two nodes at the given timestamp. While useful in some scenarios, many practical applications demand that the network predicts when the next interaction between two nodes is taking place. A tailor-made architecture will most likely use temporal point processes (generalisation of a Poisson point process, see also DyRep paper linked below) or neural ODEs.

* Scalability: Improve the scalability of the model to make it applicable to larger graphs. The current model definition requires using some slow procedural patterns (e.g. for-loops). By relaxing model assumptions in the right way and/or by designing a more efficient data loading technique, the scalability of the model can be improved.

* Improving the memory: the memory is one of the key theoretical components of the TGN model. However, it is still not completely clear why the memory makes the model perform so much better. Understanding of this would be an important advance. Moreover, variations of the memory, such as using key-value memories and using a global memory for the entire graph (together with the local node memories already used by TGN).

* Other directions that might come apparent during the project.

Reference starting points:

[1] DyRep: Learning Representations over Dynamic Graphs

[2] [2006.10637] Temporal Graph Networks for Deep Learning on Dynamic Graphs

[3] [1905.11485] Representation Learning for Dynamic Graphs: A Survey

Other resources:

The traditional benchmark data sets for link and node-level prediction on dynamic graphs, wikipedia and reddit, can be used for evaluation.

Analyzing / Solving the Graph Bottleneck

Supervisor: Pietro Lio' (pl219@cam.ac.uk)
cosupervisor: Felix Opolka

Description:

Graph Neural Networks have been shown to be extremely effective at modeling relational data. GNNs operate by sending messages between adjacent neighbors in the graph. However, a recent paper [1] has shown how standard message-passing GNNs incur a bottleneck of information due to the oversquashing of an exponentially growing amount of information into a fixed size vector. This work clearly shows an important problem of current GNNs, and opens the road to further investigation of the issue, as well as the development of methods to overcome the issue, and establishing the next generation of state-of-the-art GNNs.

The project will involve further analysis of the graph bottleneck problem, as well as possibly proposing new solutions to overcome it.

For example, an interesting analysis would be to investigate to what extent graph sampling (aggregating from only a subset of the neighbors) helps alleviate the graph bottleneck.

Reference starting points:

[1] On the Bottleneck of Graph Neural Networks and its Practical Implications

Dynamic Graphs (Physics/Biology) Application

Supervisor: Pietro Lio' (pl219@cam.ac.uk)
cosupervisor: Felix Opolka

Description:

Many graphs in the real world are dynamic, i.e. they change over time. New nodes and edges may appear or disappear, and their features can change over time. New state-of-the-art models for dynamic graphs have recently been proposed [1]. This project will explore interesting applications of models for dynamic graphs, possibly in the field of physics or biology. The goal of the project will be to identify one interesting and important (phyiscal/biological) problem which can be modelled as a dynamic graph, craft a new data set and adapt the TGN model [1] to obtain state-of-the-art results on the selected task.

Designing a suitable benchmark data set is a crucial step in advancing the field of machine learning for dynamic graphs and can be highly influential if it becomes adapted as a standard in the community (similar to ImageNet for images or Cora for static graphs). Current dynamic data sets have limitations that prevent the field from evaluating more advanced architectures, hence a new, well-crafted data set will be in high demand.

Reference starting points:

[1] Temporal Graph Networks for Deep Learning on Dynamic Graphs

[2] Representation Learning for Dynamic Graphs: A Survey

Deep Convolutional Gaussian processes for graph-level predictions

Supervisor: Pietro Lio' (pl219@cam.ac.uk)
cosupervisor: Felix Opolka

Description:

Gaussian processes are powerful non-parametric, Bayesian machine learning models. Unlike neural networks, they are effective in a low-data regime and provide reliable confidence estimates.

When applying Gaussian processes to graph-structured data, adapting them to the new domain is crucial to increasing their inductive bias. The work towards achieving this is still in its infancy. While there has been some work on Gaussian processes for graphs, even on graph-level predictions, there are still many ideas, some of them borrowed from graph neural networks, that can be applied to Gaussian processes. Examples include applying multi-hop convolutions, using deep Gaussian processes, counteracting oversmoothing, and much more. The goal is to consider these techniques and devise a (deep) Gaussian process method for graph-level prediction tasks. A real-world application of this research is for example molecule property prediction.

The amount of math required for understanding Gaussian processes might initially seem intimidating but it is actually an elegant, self-contained framework that is useful to understand for machine learning in general. Introductory material on Gaussian processes can be provided.

Reference starting points:

[1905.05739] Graph Convolutional Gaussian Processes

[1809.04379] Bayesian Semi-supervised Learning with Graph Gaussian Processes

https://grlplus.github.io/papers/50.pdf; earlier, more detailed version: [2002.04337] Graph Convolutional Gaussian Processes For Link Prediction

Deep Gaussian Processes

[1705.08933] Doubly Stochastic Variational Inference for Deep Gaussian Processes

Data imputation for graph-structured data

Supervisor: Pietro Lio' (pl219@cam.ac.uk)
cosupervisor: Felix Opolka

Description:

While in research we usually deal with clean and complete data sets, in real world scenarios data is often messy and incomplete. This necessitates data imputation, i.e. filling in missing values, as a key step in data preprocessing. Many techniques exist for data imputation on unstructured, vector-valued data where independence between data samples can reasonably be assumed. However, in graph machine learning, samples are not statistically independent, hence demanding novel techniques for filling in missing values.

The goal of this project is to develop a data imputation method for graph-structured data, for example for the various node-classification benchmark data sets. As there is less recent related work on this topic, it will require a more thorough literature review and familiarising oneself with some statistical techniques.

Meet your future in the mirror

Supervisor: Pietro Lio' (pl219@cam.ac.uk)
Assistant supervisor: Pietro Barbiero <pb737@cam.ac.uk>,Ramon Viñas Torné <rv340@cam.ac.uk>

Description Modern medicine needs to shift from a wait and react, curative discipline to a preventative, interdisciplinary science aiming at providing personalised, systemic and precise treatment plans to patients. This project aims at developing a framework based on interpretable machine learning approaches to run probabilistic simulations where the entire organism is considered as a whole. Depending on your interests, the project may focus on: (i) integrating constraint-based systems and reasoning methods with graph neural networks; (ii) proposing and developing interactive visual inference methods helping physicians in formulating and testing hypotheses by considering the whole organism; (iii) investigating the advantages of generative methods on graphs to infer multiscale relationships among biological entities.

References Barbiero, Pietro, Ramon V. Torné, and Pietro Lió. "Graph representation forecasting of patient's medical conditions: towards a digital twin." arXiv preprint arXiv:2009.08299 (2020). Barbiero, Pietro, and Pietro Lió. "The Computational Patient has Diabetes and a COVID." arXiv preprint arXiv:2006.06435 (2020). Bodnar, Cristian, Cătălina Cangea, and Pietro Liò. "Deep Graph Mapper: Seeing Graphs through the Neural Lens." arXiv preprint arXiv:2002.03864 (2020). Marra, Giuseppe, et al.
"Integrating learning and reasoning with deep logic models." Joint European Conference on Machine Learning and Knowledge Discovery in Databases.
Springer, Cham, 2019. Li, Yujia, et al. "Learning deep generative models of graphs." arXiv preprint arXiv:1803.03324 (2018).

Title: Applications of Geometric Jensen-Shannon divergence

Supervisor: Pietro Lio' (pl219@cam.ac.uk)
Assistant supervisor: Nikola Simidjievski and Jacob Deasy

Recently, a geometric mixture of statistical divergences has been shown to outperform traditional divergences when regularising the latent space of Variational Auto Encoders (VAEs) [1,2]. The regularisation mechanism introduced, based on the skew-geometric Jensen-Shannon divergence, leads to an intuitive interpolation between forward and reverse KL in the space of both distributions and divergences. This project will focus on applications of this work, investigating GJS-VAEs for properties such as robustness to noise, lossy compression, and generalization to other types of data [3].

During this project, you will learn: deep learning (PyTorch framework), variational/Bayesian inference, and make use of multiple data types - strong preparation for a Phd or industry role in machine learning. The ideal candidate for this project will have a strong background in mathematics and statistics as well as good Python programming skills.

Familiarity with bayesian methods is a plus.

[1] https://arxiv.org/abs/1312.6114

[2] https://arxiv.org/abs/2006.10599

[3] https://www.frontiersin.org/articles/10.3389/fgene.2019.01205/full

Title: Cell-type deconvolution of bulk RNA-seq

Supervisor: Pietro Lio' (pl219@cam.ac.uk<mailto:pl219@cam.ac.uk>) Assistant
supervisor: Ramon Viñas Torné <rv340@cam.ac.uk<mailto:rv340@cam.ac.uk>>

Knowledge of cell type composition in disease relevant tissues is an important step towards the identification of cellular targets of disease.
The goal of this project is to develop a deep learning method to infer the proportions of each cell-type in bulk RNA-seq samples. Depending on your interests, the project may focus on data from certain cell-types, tissue-types, diseases, or organisms.

During this project you will learn: bioinformatics, generative models, topic modelling, and deep learning. The ideal candidate for this project will have a strong background in mathematics and statistics as well as good Python programming skills.

Suggested readings (more material can be provided):

1. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference.
https://mitpress.mit.edu/books/elements-causal-inference
2. DeconRNASeq: A Statistical Framework for Deconvolution of Heterogeneous Tissue Samples.
https://pubmed.ncbi.nlm.nih.gov/23428642/
3. Non-negative Matrix Factorization.
https://en.wikipedia.org/wiki/Non-negative_matrix_factorization
4. Latent Dirichlet Allocation. https://www.jmlr.org/papers/v3/blei03a
5. Topic Modeling in Embedding Spaces. https://arxiv.org/abs/1907.04907

Project: New Methods for Learning Distributed Representations of Graphs

Supervisor: Pietro Lio
Cosupervisor: Paul Scherer

Description:
Representation learning of graphs using neural networks has turned into a large and exciting hub of research driven by successive proposals of graph representation learning methods and datasets to apply them onto. A significant part of the activity has focused on Graph Neural Networks (GNN). Such neural networks are characterised by graph convolutional operators that serve as useful inductive biases for learning representations of nodes and other graph substructures.

Alongside ongoing research into GCNNs and its variants, another approach has focused on extending graph kernel methods with neural language embedding methods that exploit the distributive hypothesis to learn representations of graphs. This is a useful alternative inductive bias to model the vector space embeddings of graphs over the distribution of the discrete substructure patterns contextualising them. Much like how the semantic meaning of words is similar to words that have similar context words around them, comparability can also be defined for graphs with the appropriate specification of what constitutes context and the entities (nodes, subgraphs, substructure patterns) that are involved.
Such vector representations of graphs are inductively biased to be close when they contain similar substructure patterns, and distant when they do not. This perspective enables the construction of a powerful class of unsupervised representation learning methods.

I believe that distributed representations of graphs are a woefully understudied area of graph representation learning in comparison to GNN methods which are at the height of popularity as of this moment. The silver lining is that there are plenty of unexplored avenues for methods learning distributed representations of graphs and many datasets at hand as well. The candidate will be expected to study the theoretical foundations of both distributed representations of graphs (with support) and contribute towards the definition of new methods. Project directions may be defined across many avenues which can be discussed and refined.

Suited to:
Students who have less experience with graph representation learning, are starting to explore it, or wish to explore it from a non-GNN approach. Experience with NLP, especially word and document embeddings will be helpful. Familiarity with linear algebra will also be helpful.
The project will carry both a research and coding element with support from myself and is particularly suited to students aiming for a publication at a workshop or conference.

References (In no particular order):
Kristian Kersting, Nils M. Kriege, Christopher Morris, Petra Mutzel, and Marion Neumann. 2016. Benchmark Data Sets for Graph Kernels. Datasets available at http://graphkernels.cs.tu-dortmund.de.
Zellig S. Harris. 1954. Distributional Structure. WORD 10, 2-3 (1954), 146–162.
Sergey Ivanov and Evgeny Burnaev. 2018. Anonymous Walk Embeddings. In
Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80) Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32 (Beijing, China) (ICML'14) Annamalai Narayanan, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu, and Shantanu Jaiswal. 2017. graph2vec: Learning Distributed Representations of Graphs.
Paul Scherer and Pietro Lio. Learning Distributed Representations of Graphs with Geo2DR. ICML 2020 Graph Representation Learning and Beyond.

Project: Learning Multiple Representations of Nodes

Supervisor: Pietro Lio
Cosupervisor: Paul Scherer

Description:
As anyone reading this is likely aware, graph representation learning has to a large extent taken over machine learning research. The canonical task in GRL is node representation learning wherein we try to learn a vector representation of a node which contains information about its local/global neighbourhood. As an example of a citation network where nodes represent research papers and edges citations between papers. Each node is also associated with a feature vector describing the content of the papers. The task is to classify the papers into different scientific categories such as “physics” or “cs” and so on.
This is where graph representation learning methods made their mark as they are able to not only use the paper’s feature vector as part of the learning process but also the papers it is related to (and their feature vectors).

Anyhow, fast forward a couple years and there's a plethora of node representation learning techniques. A bulk of these methods are centered around learning a single representation for a node. That is to say within a node classification scenario we seek to identity a single classification or role that characterises the node within the networks.
In the context of clustering a social network this would equate to partitioning social network into non-overlapping sets. However, in my opinion a much more realistic view is that people would belong to multiple communities to reflect that we have multiple personae.

This project aims to co-develop methods for learning node representations which jointly learn multiple representations of a node with respect to the various communities it is involved in.

Suited to:
To those interested in unsupervised learning, specifically clustering, graph neural networks, other graph representation learning algorithms.
There is significant freedom to this project to propose your own ideas and approaches; just as much this means there will be more independence expected. It is not a well trodden path so there will be less material around; at the same time there is a higher chance of making a research contribution, and the co-supervisor (myself) will also actively work on the project.

References (in no particular order):
Allesandro Epasto, Bryan Perozzi; “Is a Single Embedding Enough?
Learning Node Representations that Capture Multiple Social Contexts”