Inverting the rules of the game

Supervisor: Bianca Dumitrascu (bmd39@cam.ac.uk)

Description:

Morphogenesis is one of the most striking examples of emergent behavior. Cells divide, migrate, reorganize to give rise to complex patterns that make up tissues and organs. Understanding the underlying mechanisms responsible for these processes is a daunting task. Recently, this problem has seen renewed interest through methods focused on computation and inference [1,2,3,4]. In this project we will look at simple models for morphogenesis inspired by cellular automata models [1] and amend them to resemble the more richer structures encountered in nature [2]. Potential directions:

- Evaluate the robustness of current methods for evolving patterns using cellular automata inspired models [1,2,3] in the context of noisy/probabilistics formulations of both planar and 3d cellular automata.

- Evaluate the ability of recurrent neural network structures to invert Conway's Game of Life [1,2] and contrast it with the recent results for small convolutional networks [3].

- Using tools from graphical neural networks [5], consider a planar graph extension of Conway's Game of Life and evaluate the ability of the resulting recurrent neural networks to evolve such a structure.

- Learn local rules that are able to evolve realistic patterns encountered in spatial single cell transcriptomics data [6] and in simulated data [7].

This project is suitable for a student with medium programming experience and a strong statistical background, who enjoys coding, simulation and visualization, and wants to learn more about dynamical systems and spatial transcriptomics.

References:

[1] Mordvintsev, Alexander, et al. "Growing neural cellular automata." *Distill* 5.2 (2020): e23.
[2] Clamons, Samuel, Lulu Qian, and Erik Winfree. "Programming and simulating chemical reaction networks on a surface." *Journal of the Royal Society Interface* 17.166 (2020): 20190790.
[3] Springer, Jacob M., and Garrett T. Kenyon. "It's Hard for Neural Networks To Learn the Game of Life." *arXiv preprint arXiv:2009.01398* (2020).
[4] Cichos, Frank, et al. "Machine learning for active matter." Nature Machine Intelligence 2.2 (2020): 94-103.
[5] Sanchez-Gonzalez, Alvaro, et al. "Learning to simulate complex physics with graph networks." *arXiv preprint arXiv:2002.09405* (2020).
[6] Goltsev, Yury, et al. "Deep profiling of mouse splenic architecture with CODEX multiplexed imaging." *Cell* 174.4 (2018): 968-981
[7] https://www.kaggle.com/c/conways-reverse-game-of-life-2020/overview

Active Learning in Graph Neural Networks for Interventional Studies Design

Supervisor: Bianca Dumitrascu (bmd39@cam.ac.uk)

Description:

Active Learning is a powerful statistical framework that allows predictive inference when data collection is difficult and data labeling is scarce [1, 2]. In this setting, models are trained on a small amount of data, followed by the active selection of a set of new samples (based on an acquisition function) which are deemed most informative. The selected samples are then used to expand the existing training data and retrain the model. This process is then repeated over multiple iterations. Latent structure in the input data can often lead the procedure to state-of-the-art performance with significantly less training data than required by a non-active counterpart. The goal of this project is to use an active learning framework
in the context of graph neural network training. Potential directions are:

- [Empirical] Active Learning Link Prediction: Consider a large weighted graph in which all the nodes are known and where edge weights are collected in batches over a finite number of batches. This problem has two important applications.

  First, it is related to the following pharmacogenomic problem: given an existing library of repurposed, pre-approved drugs and a limited budget for performing drug pair interventions, find favorable drug pairs over a limited number of iterations.

  Second, it can be used to answer a question related to combinatorial experiment prioritization (e.g Perturb-seq, CRISPR): given singular genetic perturbations[3], can we predict combinatorial effects at gene expression level. No familiarity with biology is assumed, yet an excitement for learning about it is a must!

- [Theoretic] Active Learning Community Detection: Consider well characterised, probabilistic models such as stochastic block models. Recently [4] studied misclassification error for community detection via weighted message passing. How do these results change in the light of an active learning setting?

- Other directions that can come up during research

References:
[1] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. arXiv preprint arXiv:1711.04043, 2017.
[2]Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489, 2017.
[3]Replogle, Joseph M., et al. "Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing." Nature biotechnology (2020): 1-8.
[4]Cai, T. Tony, Tengyuan Liang, and Alexander Rakhlin. "Weighted Message Passing and Minimum Energy Flow for Heterogeneous Stochastic Block Models with Side Information." *Journal of Machine Learning Research* 21.11 (2020): 1-34.

Linguistics of batch correction: Lessons from language modelling for single cell analysis

Supervisor: Bianca Dumitrascu (bmd39@cam.ac.uk)

Description:

Learning representation of words and articles is a task central to the field of Natural Language Processing and has been the heart of many machine learning algorithms and applications. Common areas empowered by efficient word embeddings include translation tasks, words association tasks, topic modelling, and recommendation. Similarly, finding embeddings of genes and cells in order to identify cell type, disease status, or developmental trajectories has become the bread and butter of the computational biologist following single cell sequencing advancements. While governed by different grammars, the world of words and that of genetics are intrinsically related. As articles can be characterized as distributions over words, cell identity can be thought of as a distribution over counts of gene products. As related articles share topics, which in turn share specific words and their associations, cells of similar type share functional relationships, pathways, all largely determined by their genetic markup.

There are a few reasons for why many questions to modern computational biology questions can find related formulations in the computational linguistic literature. First, the easy access to text data in large amounts benefits both training and testing, thus creating a space accessible to computer scientists and statisticians to explore their ideas without the need for deep domain knowledge. Second, the ability to easily interpret the results as a post processing step has allowed the quick development of methodology that can be easily evaluated. This project explores possible synergism between computational tools for linguistics and biology. Consider multiple gene expression datasets collected by researchers in different laboratories, studying the diversity of cells in the same brain region. How can these datasets be integrated into a single dataset while correcting for study specific noise? This question, also known as batch correction or data integration, is central to biological research. Potential directions:

- Compare and contrast existing computational biological tools for batch correction with a novel technique from computer science developed in the context of word translation without parallel data [3] which is based on adversarial training. The student will compare and benchmark the methods on real, as well as synthetic datasets inspired by the statistical literature of Factor Models.
- In Euclidean space, the standard Procrustes problem is finding the optimal transformation (rotation, translation, reflection, scaling) that aligns two point clouds in a way that preserves pairwise distances within the clouds. Recently, a number of methods have been developed for representing data points in hyperbolic space, with the motivation that this representation is most appropriate for data exhibiting hierarchical structure. Can such a structure be exploited in the context of the alignment problem?
- Other directions that can come up during research

References:

[1] Luecken, Malte D., et al. "Benchmarking atlas-level data integration in single-cell genomics." BioRxiv (2020).

[2] Alvarez-Melis, David, and Tommi S. Jaakkola. "Gromov-wasserstein alignment of word embedding spaces." arXiv preprint arXiv:1809.00013 (2018).

[3] Conneau, Alexis, et al. "Word translation without parallel data." *arXiv preprint arXiv:1710.04087* (2017).