Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

Application of Capsule Net for automated DNA sequencing using tunnelling spectroscopy

Interim Report

Project Dates: 01/07/2019 - 31/12/2019

University of Birmingham, School of Chemistry

Professor Tim Albrecht & Dr Anton Vladyka
University of Birmingham

Report Date: 01/09/2019

AI3SD-Project-Series:Report-1_Albrecht_Interim

Application of Capsule Net for automated DNA sequencing using tunnelling spectroscopy
AI3SD-Project-Series:Report-1_Albrecht_Interim
Report Date: 01/09/2019
DOI: 10.5258/SOTON/P0036

Network: Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery
This Network+ is EPSRC Funded under Grant No: EP/S000356/1

Principal Investigator: *Professor Jeremy Frey*
Co-Investigator: *Professor Mahesan Niranjan*
Network+ Coordinator: *Dr Samantha Kanza*

Contents

1 Project Details

Title	Application of Capsule Net for automated DNA sequencing using tunnelling spectroscopy
Funding reference	AI3SD-FundingCall1_016
Lead Institution	University of Birmingham
Project Dates	01/07/2019 - 31/12/2019
Website	N/A
Keywords	Machine Learning; single-molecule detection; Capsule Nets; Deep Learning

2 Project Team

2.1 Principal Investigator

Name and Title: Tim Albrecht Professor of Physical Chemistry
Association: University of Birmingham, School of Chemistry
Work Email: t.albrecht@bham.ac.uk
Website Link: albrechtlab.com

2.2 Co-Investigators

Name and Title: Dr Eduardo Alonso, Reader and Head of Department
Association: City, University of London, School of Mathematics, Computer Science & Engineering, Department of Computer Science
Work Email: e.alonso@city.ac.uk
Website Link: https://www.city.ac.uk/people/academics/eduardo-alonso

2.3 Other Researchers & Collaborators

Dr Anton Vladyka is the postdoctoral research associate currently working on the project. He has replaced Dr Joseph Hamill, who was originally named on the application, as Joseph continued his work on another project taking place in the group.

Mr. Christropher Weaver has completed his Master project in the Albrecht during the 2018/19 academic year with a project on the detection of DNA nucleotides using Scanning Tunnelling Microscopy in solution. He was able to generate the first experimental proof-of-concept data, which formed the basis for the current proposal. As Chris has a strong background in AI, having completed the Chemistry with a Year in Computer Sciences MSci degree in Birmingham, he was also able to repeat some of the work we had published in our 2017 Nanotechnology paper on Deep Learning in Single-Molecule Science. He will join the group again in October, as PhD student, with the aim to systematically investigate various experimental parameters to optimise the detection of DNA nucleotides via tunnelling. He will then also benefit from the foundation in AI techniques that has been set up in the current project. As a consequence, however, we decided to put greater focus on the AI aspects of the project at this early stage and then intensify our effort around the generation of experimental data towards the end of the project. Chris will then also be able to continue the development of AI-based approaches, given his background.

3 Publicity Summary

In this project, we will take quantum tunnelling-based biosensing and sequencing to a level that would make the most promising contender for label-free 'next-next' generation sequencing of biopolymers. This will be achieved by combining state-of-the-art surface chemistry, nanoscience and chemical sensing with Capsule Nets (CN) as a novel Deep Learning methodology, to maximise the extraction of information from the tunnelling data.

4 Executive Summary

The aim of this project is to explore the limits of quantum-based sensing for individual DNA nucleotides A, T, C and G in combination with Capsule Nets, which is a realistic goal for a 4-month feasibility study. If its performance is comparable to or even exceeds current commercial sequencing techniques at this early stage, then this makes a strong case for further, significant investment towards the implementation of a complete sequencing platform on the medium- and long-term. In terms of future funding applications and our longer term plans, our primary target would be EPSRC, but given the rapid dynamics of the sequencing market potentially also venture capital funding, e.g. from Illumina Ventures, where TA already has existing links.

5 Aims and Objectives

The goal of the project is to improve automatic recognition of the nucleotides in the spectroscopy-based approach. We propose to use Capsule Net as an alternative to already studied CNN. Existing CNN have some translational invariant, but still require some pre-prosessing of input data. In contrary, the architecture of capsule net allows to store the location of the 'event' within analysed data to improve recognition. In addition, CN can detect more parameters of event signal such as duration, magnitude etc.

6 Methodology

6.1 Scientific Methodology

Experimental data for the network training are acquired from the tunnelling spectroscopy of the nucleotides solution in STM, for which a limited amount of data exists, or from realistic simulated data (i.e., informed by initial experimental data, in terms of their statistical properties). In this context, simulated data can be advantageous, because they allow for specific signal parameters, such as amplitudes or noise levels, to be modulated and their effect on prediction performance investigated. Experimental or simulated input data represent current-time (I(t)) traces, which are then split into 500 data point sequences and subjected to AI-based analysis (1-dimensional input arrays [1x500]).

6.2 AI Methodology

To realize automatic recognition of the nucleotides from the I(t) signal, a Capsule Network was built based on the 'Dynamic Routing Between Capsules' paper (G. Hinton et al., 2017) and implemented in Python using the Pytorch framework. As an input of the network, the slices of the original trace with 500 data point were used. The structure of CN is mimicking the architecture of CNN for nucleotide detection (Albrecht, T. et al. Nanotechnology 28, 2017). The encoder of the network includes two capsule layers which inter-capsule routing. The goal of the routing is to make hierarchical connections between features extracted at previous layers of the network.

Overall, the structure of the network is the following:

1. 1D-convolutional layer (CL), 128 filters with 1x9 kernel and a stride of 2, followed by ReLU
2. 1D-convolutional layer (CL), 128 filters with 1x9 kernel and a stride of 2, followed by ReLU
3. Capsule layer with 8 capsules, where each capsule includes 32 convolutional filters with 1x9 kernel and a stride of 2
4. Capsule layer with 6 capsules (=number of classes of the network) with 16 output parameters in each
5. Decoder: a linear sequence of 3 FC layers each followed by ReLU

Overall, the network has 2477300 trainable parameters (1.8M in the encoder and 0.65M in the decoder stages).

Test evaluation of the network is performed on the simulated data (Albrecht, T. et al. Nanotechnology 28, 2017). 1000 sequences were generated with 10000 points in each, which were then split into 500 points slices. Overall, full dataset consisted of 20000 slices with 500 point in each for 6 classes (no event, A, G, C, T, M).

7 Interim Results

- With simulated current-time data, a CNN and a Capsule Net were shown to yield similar results.
- Because of the possibility to extract spatial information of the features in data, Capsule Nets do not require pre-processing of the data. As one consequence, Capsule Net is more steady for edge effects (when the only event in the trace occurring in the very beginning or very end of the trace). Secondly, Capsule Net is a promising technique for the classification of mixed data, where individual trace contains events from different nucleotides, and analysing the capsule states every event can be related to its class (to be studied additionally).
- Because of larger number of computing operations, Capsule Net is more computationally expensive for training, but being trained, the prediction is almost as fast as using CNN with similar architecture.
- In related work, we could show that Transfer Learning can enhance the performance of Autoencoders in unsupervised classification tasks for single-molecule charge transport data.

8 Outputs

The results from this project have so far not been presented in talks or published in research papers. However, two research papers originating from this project are currently in preparation:

- Anton Vladyka, Tim Albrecht et al., "A Comparison of Capsule Net and Deep Convolution Neural Networks during the Single-Molecule Detection of DNA Nucleotides"
- Anton Vladyka, Tim Albrecht, "Unsupervised Classification of Single-Molecule Data with Autoencoders and Transfer Learning", manuscript in preparation*

* This work is a very interesting spin-off of the present project, focused around the use of transfer learning and autoencoders for unsupervised classification of single-molecule data.

9 Progress Summary

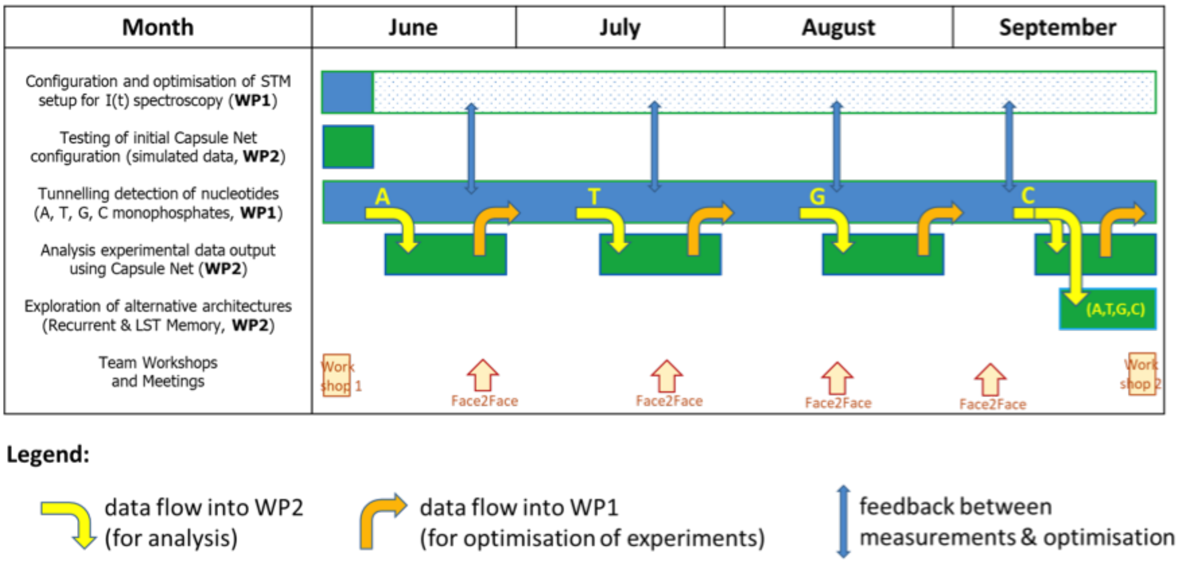
Month
June
July
August
September
Configuration and optimisation of STM setup for I(t) spectroscopy (WP1)
Testing of initial Capsule Net configuration (simulated data, WP2)
Tunnelling detection of nucleotides (A, T, G, C monophosphates, WP1)
Analysis experimental data output using Capsule Net (WP2)
Exploration of alternative architectures (Recurrent & LST Memory, WP2)
Team Workshops and Meetings
A
T
G
C
(A,T,G,C)
Work shop 1
Face2Face
Face2Face
Face2Face
Face2Face
Work shop 2
Legend:
data flow into WP2 (for analysis)
data flow into WP1 (for optimisation of experiments)
feedback between measurements & optimisation

Figure 1: Original Gantt Chart, as included in the proposal:

- The actual project start date was 01/07/2019 and the new end date is 31/12/2019 (6 months, as a non-cost extension), That is, the project has been active for approximately two months by now. Formally, Dr Vladyka's involvement in this project only starts in September, but his current project allows for sufficient flexibility to incorporate its objectives and generate some first results (see Section 8).
- The initial workshop took place on 04th June 2019, with participation from the project team (Albrecht, Vladyka, Weaver, Alonso) as well as additional specialists from City (Alex Ter-Sarkisov (Deep Learning in Computer Vision), Carlos Reyes-Aldasoro (image analysis/biomedical applications)). Apart from personal introductions and presentations around people's previous track record, means of communication were established and first steps initiated.
- As mentioned in a previous section, due to the involvement of additional project participants (Weaver) who nominally focus on the experimental aspects of the work programme, we decided to prioritise the AI-related elements at the early stages of the project, essentially to expand the portfolio of available data analysis tools once systematic experimental data become available.
- Nevertheless, we have been able to generate meaningful simulated data based on known statistical properties of preliminary experimental data. These simulated data were then used as the basis for a comparative study of CNNs vs. Capsule Nets. This was initially planned for month 1 (setup phase) with repeated testing throughout the project. While a quantitative comparison of the network performance is non-trivial, overall we observed that for 'unperturbed' events both approaches yielded similar results (vida supra). However, the Capsule Net outperformed the CNN, when events occurred at the edges of the current-time traces, i.e. they were not fully resolved or data were missing.

10 Next Steps

The project will broadly follow the original timeline, while focus will gradually shift from establishing the AI-based toolbox towards testing the respective methods with experimental data, to be generated during the second half of the current project and beyond. New AI methods mentioned in the proposal, such as RNNs and LST memory networks, will be investigated over the coming weeks.

The immediate next steps are to finalise our efforts in regards to the first two outputs, namely the manuscripts listed in Section 8.

11 References

See Text

12 Data & Software Links

Capsule Net used in this project was implemented in Python using pyTorch framework for deep learning, based on publicly available implementation https://github.com/gram-ai/capsule-networks. Generator of simulated nucleotide tunnelling was also implemented in Python.