Ansøgning om Stibofondens IT-rejsestipendie til ph.d.-studerende

Projekt plan for eksternt forskningsophold ved University of Cambridge

Beskrivelse af studieopholdets relevans og udbytte for ph.d.-afhandlingen

Det forventede udbytte er et 4 måneders eksternt forskningsophold på University of Cambridge fra September 2019 til start Januar 2020, som faciliterer et samarbejde på tværs af forskningsgrupper (CogSys på DTU Compute og MLG på Cambridge) og udgivelse af 2 artikler med Assistant Professor, José Miguel Hernandez-Lobato. Idéen der ligger til grund for opholdet er, at bruge deep learning til data drevet modellering af Raman spektre og den bagvedliggende proces givet den molekylære struktur til automatisk design, detektering og screening af medicin i micro-containers. Dette er i tråd med mit PhD projekt i "Active Deep Learning for Nano-Sensor Systems", som omhandler machine-learning assisteret forskning til nano-sensor og udvikling af medicin. Projektet er et led i forskningsmålet for hele grundforskningscenteret, IDUN Center of Excellence (http://www.idun.dtu.dk/research), og forventes, at bringe stor værdi til både forskere i nano-teknologi og Dansk farmaceutisk industri. Dette opnåes ved, at bruge de dybe prædiktive modeller til effektiv automatisk screening og bestemmelse af egenskaber for ny medicin. Dette kan anvendes til en strømlining af designprocessen af både medicin og de micro-containers der skal indeholde det, således at vi kan levere medicinen mere målspecifikt, effektivt og med større sikkerhed for patienten.
Da værts-vejleder og forskningsgruppen i Cambridge har en stærk international profil indenfor de metoder jeg anvender i mit projekt, såsom usikkerhedsestimering, Bayesianske neural netværk, aktiv læring og automatisk design af eksperimenter, vil dette ophold være den perfekte mulighed for, at udvikle min forskning, viden og idéer i machine-learning assisteret eksperimentering i et anerkendt og aktivt forskningsmiljø. Da University of Cambridge kræver et bench fee på 500 GBP/mdr (i alt 17.300kr), bliver denne udgift dækket af mit projekt. Med højere leveomkostningerne i Cambridge og udgifter til udstyr og udgivelser, anmoder jeg derfor om støtte af jer til, at dække merudgifterne på 34.000 kr. forbundet med opholdet. Jeg ansøger hermed om et rejsestipendie på 35.000 kr. Tak for jeres tid.

4 June 2019
MFV

De bedste hilsner,

Maximillian Fornitz Vording
PhD Student
Section for Cognitive Systems

DTU Compute
Department of Applied Mathematics
Computer Science
www.compute.dtu.dk

Building 321, Room 106
Richard Petersens Pl,
2800 Kgs. Lyngby, Denmark

Dir. +45 61603013
Tel.. +45 45 25 30 31 and
maxvo@dtu.dk

Detailed plan for external stay

For my external research stay I have ideas for combining deep generative models (DGM), uncertainty estimation, multi-task learning and active learning (AL) for molecular exploration and detection through Raman spectra as a molecular fingerprint. This is mainly inspired by the work of my host-supervisor, Assistant Professor, José Miguel Hernández-Lobato, on ChemVAE, GVAE, BNN-LV, PBP, the recent work on Deep Learning Spectroscopy by my supervisor, Associate Prof. Mikkel N. Schmidt and my own work w. Prof. Ole Winther developing a Gaussian-mixture-VAE for gene expressions and exploration of cell types in latent space. These ideas lead to Miguel, my supervisors and me agreeing on and planning a 4 months research project in the Machine Learning Group, University of Cambridge, which now needs additional external funding from you.

The project for this 4 months external stay is a part of my PhD project in Active Deep Learning for Nano-Sensor Systems, which is described in detail below, and will therefore be focused on developing deep learning models for Raman spectroscopy, which is the most commonly used technique for detection of molecules. Therefore I will here motivate Raman spectroscopy and explain why a data-driven approach will benefit the research and industry in pharmaceutical engineering.

The first research question is: Can we determine and detect the fingerprint of the building blocks of nature itself in a smarter more data efficient way? The hypothesis is: We can use deep neural networks for learning the underlying process behind how light is scattered by linking the Raman spectra and various representations of molecular structure. Learning how to generate the fingerprint of molecules, will make it possible to detect the molecules, their concentrations and properties as medicine contained in micro-containers by non-invasive methods, e.g. lasers.

Raman scattering stems from driven molecular vibration coming from symmetric and asymmetric stretching of bonds between atoms in molecules, when the incident laser hits the molecule. The intensity is below 0.01% of the direct Raleigh scattered light, so it needs filtering and the signal-to-noise-ratio is low, which also argues for using Surface-Enhanced Raman Spectroscopy (SERS) or Coherent anti-Stokes Raman scattering spectroscopy (CARS). SERS leads to a very complex behaviour and makes it hard to simulate with DFT. The frequency of the Raman scattered light (Raman bands) will depend on the strength of the atomic bonds and atomic masses and can be modelled in the time domain with an ordinary differential equation like Hooke's law and Newton's 2nd law of motion. Through quantum mechanical equations the differential functional theory (DFT) can be used to simulate Raman spectra, but for complicated molecular compounds and environment like SERS, the simulation rarely fits with real world measurements

and will often be re-adjusted manually to fit this. This suggests using a data-driven approach by generating spectra through deep graph neural networks, which are commonly used for linking molecular properties to the molecular structure.
Here is the representations of molecules and what we need for predicting molecular properties and Raman spectra. We need both molecular structure, atomic mass and band strength to determine vibrational modes and differential equations.

A list of these commonly used representations of molecular structure:
- SMILES (Text representation of molecules)
- Coulomb matrix (Energy interaction based on distances between all atoms)
- Bag of bonds
- Histograms
- Radial distribution functions
- Chemical environment
- ACSF

We can find these resources in the following datasets:
- https://www.researchgate.net/post/Free_Database_with_Raman_spectra
- https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/
- https://serc.carleton.edu/research_education/crystallography/xldatabases.html
- http://oqmd.org/
- http://quantum-machine.org/datasets/

At the IDUN Center of Excellence, researchers can as well provide huge amounts of data for our project through their ongoing experiments with Raman spectroscopy on micro-container drug-delivery. With computational resources at both University of Cambridge and DTU Compute, we can scale up to big data and thereby the representational strength of our deep learning models tremendously.

The plan for the project during the external stay is:
For this project the goal is for me to be the main author on two publications in high-impact conferences or journals and will be conducted in collaboration with the following researchers:

Host-supervisor: *José Miguel Hernández-Lobato - Assistant Professor (CAM)*
Principal supervisor: *Mikkel N. Schmidt - Associate Professor (DTU)*
Co-supervisor: *Tommy Sonne Alstrøm - Senior Researcher (DTU)*

We will have a weekly 30 min. Meeting with Host-supervisor and additional 30 min. Meeting with supervisors at DTU Compute, so we synchronise ideas and work.
The work will follow the following study plan roughly dependent on results along the way.

Study plan for external stay:

1. **September:**
    1. Test deep generative models, Grammar variational autoencoder (Work by host J. M. Hernández-Lobato) and Graph neural networks for spectra (work by supervisor, M.N. Schmidt), for various representations of molecules, to choose best candidate.
    2. Based on this screening, collect datasets with Raman spectra and chosen molecular structure and additional molecular properties.
2. **October:**
    1. Create a combined model linking the latent representation of molecules with a generative neural network for Raman spectra.
    2. Test different likelihood distributions and peak shape functions (Pseudo-Voigt or Lorentz).
    3. Write paper on results and submit to either journal IEEE Transactions on Pattern Analysis and Machine Intelligence or given it's state wait to January and submit to ICML.
3. **November**:
    1. Use the deep generative model for active learning and estimating uncertainties on Raman spectra and predicted properties.
    2. Design and test an automated experimental design process out of this with actual humans-in-the-loop interacting with the model.
4. **December**:
    1. Write paper based on using the model for active learning and experimental design and efficient screening of molecules.
5. **January:**
    1. Submit one or two papers to Thirty-seventh International Conference on Machine Learning (ICML) in the end of January, 2020.