Master's Thesis Pitch
Open AI Day at NTNU Norwegian Open AI Lab
March 26, 2020

How can we give artificial intelligence systems used for medical diagnosis the ability to say... "I have never seen this before. Don't trust me on this one." ?

Problem Description

Medical diagnosis is difficult because not every case is a typical case. Sometimes patients have signs and symptoms that are unusual. In extreme cases, due to genetic mutation, the patient is the only person in the world with the disease and its associated symptoms.

When a doctor comes across an unusual case she may say "I've never seen this before." and proceed carefully. An AI system that diagnoses patients should also have the ability to say... "I have never seen this before. Don't trust me on this one." Assuming the AI system is based on supervised learning, how can we add capability to the AI system so that it tells users to what extent its training dataset is representative of the current case? Maybe this requires a secondary AI algorithm built with unsupervised learning?

We can attempt to tackle this problem by looking at a system which performs automatic analysis of lung sounds. Such a system could perform better in the real-world if it detected input data which will perform poorly due to either the input data not being representative of the dataset or the data is of poor quality. The excluded data can then be assessed through traditional means such as review by a doctor.

Company

dedeX is a med tech startup. dedeX is looking to solve the problem: how do we get the data needed to build AI algorithms that can empower medical history taking and physical examination?

dedeX understands both real-world healthcare and AI and can give the student feedback from this unique perspective.

Data

Two separate datasets of annotated lung sounds with different bias and variance should be used. One is the 'development' dataset and the other is a 'real-world' dataset. The development dataset will be used to create a 'primary' supervised machine learning algorithm and a 'secondary' unsupervised machine learning algorithm.

The primary algorithm will be used to classify lung sounds in both datasets. This algorithm can be based on existing work by other researchers.

The secondary algorithm will identify data from the real-world dataset that will perform poorly when used in the primary algorithm. In simplified terms it is an unsupervised learning algorithm which looks for the data that is the least similar to data in the development dataset. When this data is excluded from the real-world dataset then the primary algorithm should achieve higher accuracy on the real-world dataset.

We have identified four sources of annotated lung sound data. The student can pick two of these for this project or alternatively find other data sources if more convenient.

Int. Conf. on Biomedical Health Informatics - ICBHI 2017
http://www.auditory.org/mhonarc/2018/msg00007.html
https://www.kaggle.com/vbookshelf/respiratory-sound-database

Department of Community Medicine at UiT – The Arctic University of Norway
http://bdps.cs.uit.no
"In the Tromsø Lung Sounds project we are building a database with more than 36.000 lung sound recordings. The recordings are done as part of Tromsøundersøkelsen 7, which is an Epidemiological study that was started in 1974. The database will be used to provide educational and analysis services for lung sounds. Our contributions are methods for automated classification and similarity search for the sounds. This project is done in collaboration with Hasse Melbye at the Department of Community Medicine, University of Tromsø. The results from this project are further developed by our Medsensio AS startup."

Thinklabs
https://www.thinklabs.com/lung-sounds

Interactive Systems for Healthcare
https://is4health.com

Task

1. Investigate the possibility of applying existing state-of-the-art deep learning methods to solve the above tasks. As part of this, the student is expected to perform a state-of-the-art literature review and implement the most relevant method(s) that can solve the problem.
2. Make the chosen method time effective, feasible and available for researchers or developers of AI technologies for diagnosis.

Thesis Information

Timeframe: 6 or 12 months
Supervisor:
dedeX contact: Jon Bekker - jon@dedex.ai