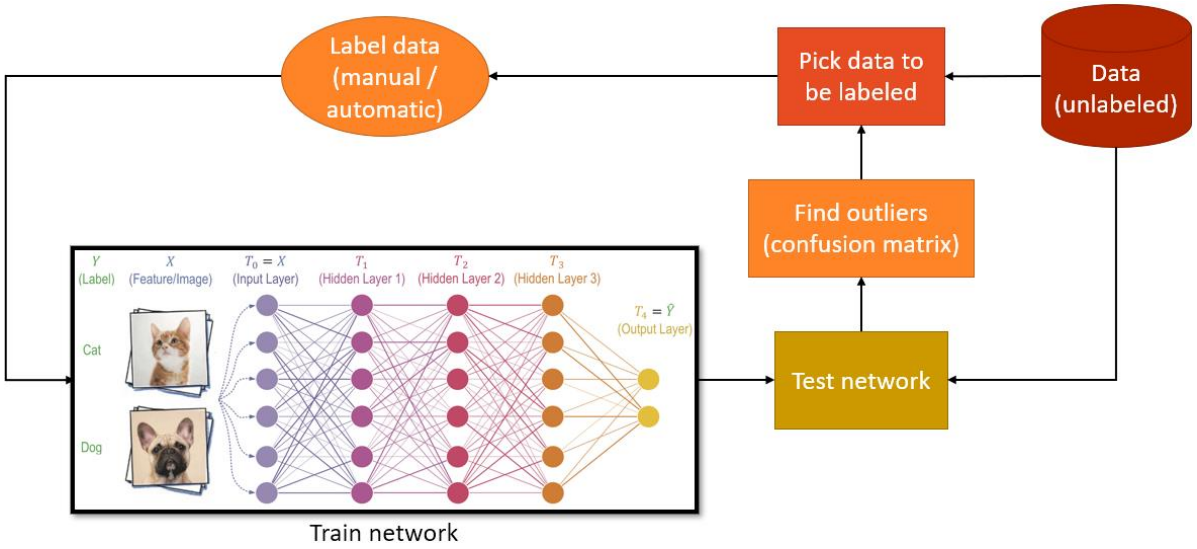
Graduation project:
Deep learning infrastructure for efficiently and iteratively labeling and training deep neural networks

Project description

Machine learning has witnessed a tremendous amount of attention over the last few years. The developments in this field are mainly driven by breakthroughs in artificial neural networks, often termed deep learning, a set of techniques and algorithms that enable computers to discover complicated patterns in large data sets. Feeding the breakthroughs is a.o. the increased access to data ("big data"). These models nowadays form the state-of-the-art approach to a wide variety of problems in computer vision, language modeling and robotics. However, access to labeled big data is not always available for many complicated problems that could potentially be solved using a deep neural network. Since labeling data can be time-consuming, automating this process as much as possible is beneficial. This can possibly be done by iteratively labeling part of the data and (re-)training the neural network using the extended labeled data set. Choosing the correct part of the dataset to be labeled is of great important here. This means retraining the network on data points that previously led to network outputs with high uncertainty.



The goal of this project is to design and implement an infrastructure for i) easily labeling data points in a (large) data set, ii) training neural networks on the labeled part of this dataset, iii) testing the network on (part of) the unlabeled data set, iv) automatically selecting data points that yield uncertain neural network outputs, and v) labeling these outliers for retraining. See also the scheme above.

The result will be evaluated on a dataset of 10,000 dermatoscopic images of common pigmented skin lesions. Part of the evaluation is a comparison of network quality between a network trained on the full dataset versus a network trained on a small dataset that contains mostly outliers (the aforementioned iteratively trained network infrastructure).

Project activities

- Literature research
- Design and implementation of the iterative labeling, testing and training infrastructure
- Programming in Python
- Application of the infrastructure on a dataset of 10,000 images
- Evaluation of performance
- Documentation

The exact scope and goals of the project will be determined in consultation.

Prerequisites

- Energetic master student in electrical engineering, biomedical engineering, computer science or a related field
- Smart student who is able to apply technical skills (design machine learning algorithms, conceptual design of complex infrastructure)
- Programming experience in Python
- Good documenting skills to ensure that the project output is properly captured
- Team player

Details

Duration: 9-12 months
Location: Nobleo Technology, Heggeranklaan 1, 5643BP, Eindhoven
Start date: 1 September 2020
Contact information: Rik Wetzels (0649937195)