

Public Abstract

First Name:Ming

Middle Name:

Last Name:Du

Adviser's First Name:Yi

Adviser's Last Name:Shang

Co-Adviser's First Name:

Co-Adviser's Last Name:

Graduation Term:SP 2017

Department:Computer Science

Degree:MS

Title:EMPIRICAL STUDY OF DEEP NEURAL NETWORK ARCHITECTURES FOR PROTEIN SECONDARY STRUCTURE PREDICTION

Protein secondary structure prediction is a sub-problem of protein structure prediction. Instead of fully recovering the whole three dimensional structure from amino acid sequence, protein secondary structure prediction only aimed at predicting the local structures such as alpha helices, beta strands and turns for each small segment of a protein. Predicted protein secondary structure can be used for improving fold recognition, ab initial protein prediction, protein motifs prediction and sequence alignment.

Protein secondary structure prediction has been extensively studied with machine learning approaches. And in recent years, multiple deep neural network methods have pushed the state-of-art performance of 8-categories accuracy to around 69%. Deep neural networks are good at capturing the global information in the whole protein, which are widely believed to be crucial for the prediction. And due to the development of high level neural network libraries, implementing and training neural networks are becoming more and more convenient and efficient.

This project focuses on empirical performance comparison of various deep neural network architectures and the effects of hyper-parameters for protein secondary structure prediction. Multiple deep neural network architectures representing the state-of-the-art for secondary structure prediction are implemented using Tensor-Flow, the leading deep learning platform. In addition, a software environment for performing efficient empirical studies are implemented, which includes network input and parameter control, and training, validation, and test performance monitoring. An extensive amount of experiments have been conducted using popular datasets and benchmarks and generated some useful results. For example, the experimental results show that recurrent layers are useful in improving prediction accuracy, achieving up to 5% improvement on 8-category accuracy. This work also shows the trade off between running speed and building speed of the model, and the trade off between running speed and accuracy. As a result, a relatively small size recurrent network have been build and achieved 69.5% 8-category accuracy on dataset CB513.