MOLECULES IDENTIFICATION FROM ^{1}H (ONE H) NMR SPECTRUM USING DEEP LEARNING TECHNIQUES

M. H. L. Chamali

A dissertation submitted in partial fulfillment of the requirement for a Master of Science degree in Computer Science

Department of Computing
Informatics Institute of Technology, Sri Lanka
In collaboration with
Robert Gordon University, UK

2021

Abstract

Integrating the chemistry concepts about NMR and the deep learning technology advance concepts, in here predict the SMILES structure of molecules for the given ^{1}H NMR spectrum. The NMR is the principal method to identify the molecular structure and identify the content and purity of a sample. ^{1}H NMR which is also called proton NMR is more popular for identifying organic molecule structures among chemistry experts and analyzing it manually, so it is time-consuming. However, it is really challenging to implement a system to predict molecular structure for a given NMR because it is required to identify the ^{1}H NMR image correctly and required to generate a correct sequence of SMILES with a meaningful molecule. That means this research is a combination of computer vision and natural language processing (NLP) capabilities.

Manually downloaded ^{1}H NMR public data set, and open-source SMILES data set was used to carry out this research. The resized NMR image was converted into its feature vector using the InceptionV3 model in CNN referring to the transfer learning techniques. Before training the model, SMILES was tokenized into the characters and combined with the ^{1}H NMR image feature vector. Then SMILES was converted into a vector. The LSTM technique in deep learning was used to identify the hidden pattern of SMILES characters. The Seq2seq approach is utilized to merge this CNN and RNN model and return a single model following the encoder-decoder concepts to predict molecule structure based on the given ^{1}H NMR spectrum. The hyper-parameter techniques were carried out to maximize the predictivity of this deep learning model. This implemented model was evaluated using ROUGH techniques and with the benchmark study. The implemented system predicts molecular structure in SMILES notation for provided ^{1}H NMR spectrum less than 2.5 seconds and the model accuracy around 90%.

Key Words: ^{1}H NMR, SMILES, Seq2Seq, encoder-decoder, CNN, RNN, LSTM, Computational Chemistry, Deep Learning