Spectral Identification via Convolutional Neural Network models

M.A.Z. Chwodhury and M.A. Oehlschlaeger

Department of Mechanical, Aerospace and Nuclear Engineering,

Rensselaer Polytechnic Institute, Troy, New York

Abstract:

Keywords:

1.Introduction

Convolutional neural networks (CNN) are widely used to extract features from data. These features are used to represent the data. CNNs have the added advantage over fully connected neural networks since the output from each of the applied filter on the data can viewed and interpreted to better understand the limitation of the training data and how the model is making its decision when performing a classification task.

Spectral identification is important for …….

Spectral identification is difficult because of overlapping of spectral shapes across different molecules and blending of spectral features owing to broadening of the spectra.

CNNs can be 1 dimensional. Commoly used CNNs are two dimensional since they are mostly used on image data. In literature, numerous work have been documented for using 2D CNNs to learn useful representations such as edges, contours, contrat, saturation etc. The difficulty in using a 2D CNN on spectral data is that spectral data is high-dimensional along a single axis. The rotational spectra in THz frequencies and the rovibrational spectra in IR range contains many sample points depending on their resolution. This can range from as low as 200 to a maximum of 360001 for the data presented in this paper. If all these data points are fed to a ML model as features , a supervised classification problem requires fitting a large number of variables. A 2D CNN is unable to process such data since it cannot move the filter kernel along multiple axes of such a high dimensional data point (spectrum). Therefore, converting spectral data to images is necessary before using a 2D CNN. This conversion may cause overall reduction in the model’s performance if the spectra are not resampled correctly. [Cite Lecun and other imageNet model papers]

A 1D CNN, however, can handle high dimensional data. They are widely used to learn time-series data for classification and regression purposes. 1D CNN moves a sliding window filter kernel along a single axis of the spectrum to learn features from the data. We can evaluate the output from the filter kernel to observe if the model basing its classification decision on spectral peaks, shapes and locations for each of the molecules.

In prior work, we have demonstrated that conventional ML models, such as random forest, fully connected neural networks and support vector machines can achieve high classification accuracy by fitting available spectral data points. While the performance of these models based on specifically designed features in a frequency range is good, they may fit specific spectral frequency locations instead of generalizing the spectral peaks, shapes and widths.

In this paper, we report the results from two different approaches to classifying THz and IR spectra based on CNN models. In the first approach, we convert spectra to images and then apply a 2D CNN model on the images for identification. The model’s output are inverted to better understand the underlying behavior. In the second approach, spectra is fed to a 1D CNN. Afterwards, we investigate the feature maps from both 1D and 2D CNNs. We interpret the feature maps from an spectroscopy point of view to understand if the models are learning the data by identifying relevant peaks and shapes of the spectra. Finally, we evaluate the performance of the CNN. The effect of pressure and size of the training data is explored. From receiver-operating-characteristics curves (ROC) we analyze the model’s performance across different molecule classes.

2.Method

2.1. Training, testing and experimental data

The data consist of simulated spectra and experimental spectra measured in our laboratory [] and found in literature. The simulated spectra is calculated based on line positions found in HITRAN and JPL using the HAPI tool. There are two datasets used in this paper. One in the THz range (7.33-11 cm-1) and the other is in the mid-IR range (400-4000 cm-1) . The THz dataset consists of 12 molecules and ranges from 0.1 to 16.5 Torr pressure values. The IR dataset comprises 34 molecules with pressures below 1 atm. Each spectrum only contain a single molecule. For details of these two dataset, readers are referred to [xx] and [xx]. The suuplementary data also consists tables and plots to better understand the data.

2.2 Experimental method

The experimental data in the THz range was measured in our laboratory. The details of which can be found in [xx] and [xx]. The IR experimental data is from NIST [xx]. The hydrogen chloride IR spectra is from PNNL database.

2.3. 1. 1D CNN

Given a spectrum x consisting of xn features. We apply filter f(?), max pooling (argmax()), dropout and dense connections to map to the label integers y while minimizing categorical cross entropy loss. The data may require normalization.

Equations here

2.3.2 2D CNN

Given a spectrum x consisting of xn features, we convert it to a binary image, d belonging to image set D. We then apply, 2D convolutions, max pooling and dense layers to map to y while minimizing categorical cross entropy loss.

Equations here

Fig 1. a) 1D CNN flowchart b) 2D CNN and image conversion flowchart

3. Results

Fig 2. Accuracy vs epoch. a) 1D CNN, b) 2D CNN on THz data.

Fig 3. Accuracy vs epoch. a) 1D CNN, b) 2D CNN on IR data.

Table 1. Accuracy, precision, recall and F1 scores for 1D CNN on THz and IR data

Table 2. Accuracy, precision, recall and F1 scores for 2D CNN on THz and IR data

Fig 4. ROC curves for 1D CNN on a) THz data and b) IR data

Fig 5. ROC curves for 2D CNN on a) THz data and b) IR data

Fig 6. Feature maps/embeddings/representations from 1D CNN on THz data

Fig 7. Feature maps/embeddings/representations from 1D CNN on IR data

Fig 8. Feature maps/embeddings/representations from 2D CNN on THz data

Fig 9. Feature maps/embeddings/representations from 2D CNN on IR data

The supplementary data contains more feature maps corresponding to individual molecules in THz and IR range.

4. Discussions

Interpret the ROC curves with further explanation.

Interpret the feature maps from spectroscopy point of view.

I need to make the plots before writing what more is going to be in the discussion.

Compare performance of CNN against SVM and other methods.

5. Conclusion

We demonstrate the use of CNNs for automated identification of spectral data in THz and IR range. Parameters affecting the performance of CNN are discussed. The feature maps of the CNN provide valuable insight into the spectral data and elucidate how and what CNN models learn from spectral data.