Component identification for Raman spectra with deep learning network

Xin Wang¹,², Qi-hang Pan¹, Xian-guang Fan¹,² Ying-jie Xu ¹,²

¹Department of Instrumental and Electrical Engineering, School of Aerospace Engineering, Xiamen University, Xiamen, Fujian 361005
²Fujian Key Laboratory of Universities and Colleges for Transducer Technology, Xiamen, Fujian 361005

Email: xuyingjie@xmu.edu.cn

Abstract. Raman spectroscopy is widely used in the research of the molecular structure of substances because of the advantages of no invasion, no damage and no interference from water. Meanwhile, component identification for mixtures is still challenging in Raman spectra. In this paper, a graphics-based sample-generating method and a model based on deep-learning for component identification was proposed. Convolution neural network (CNN) model is an important part of deep learning network and CNN models was utilized to assess the possibility of the presence of components in samples. As is shown in the comparative studies, the model was sensitive to the relative position of the characteristic peaks and could learn spectra features in mixtures. The deep-learning based component identification method showed more robustness than conventional linear fitting methods. Therefore, the method provided a valid approach to component identification for mixtures and has the potential in spectra component analysis.

Keywords: Raman spectra, component identification, data augmentation, deep-learning; convolutional neural network

1. Introduction

Raman spectrum was firstly discovered in 1928. Its features of non-invasive and high-efficiency [1] allows widely used as finger technique in spectra analysis. Whereas, the research on molecular information, especially the component identification faced a challenge that lack of accuracy and stability [2]. Thus, various methods had been proposed to solve this problem.

In the early years, people differentiated the Raman spectra of components with naked eyes. With the development of the mathematical algorithm, identification methods based on linear fitting [3] were proposed, the principle of these methods is to compare the fitting curve of the spectra with the standard spectra in database by the similarity of peak or curve [4]. Currently, most common methods for Raman spectra are based on linear fitting due to its convenience. However, mathematics methods are more adept at the recognition of the pure spectra, while in the actual situations, the spectra contain multiple components, noise signal and peak aliasing. Moreover, the limitation of mathematics methods contained lack of flexibility between different instruments [5] Novel methods on mixtures spectra identification were in urgent need.

In recent years, with the progress of computer technology, new algorithms that aim at learning the features and characteristics have been applied to various fields. The concept of deep learning was proposed by Hinton [6] in 2006 accompanied by the proposal of DNN (Deep Neural Network) model,
which has the ability of intelligent identification and adaptability. In order to achieve higher efficiency, CNN (Convolutional Neural Network) [7] was proposed with convolution kernel sharing property that could learn the features of samples better.

The Underlying logic of deep learning is similar to component identification, methods based on CNN or other deep-learning models had been widely applied to identification of various spectra. Current approaches to identify the component of various subjects had made full use of deep-learning. High accuracy and reproducibility had been showed in the experiments of Xiaqiong Fan [8] who used the 1D CNN model trained by spectra dataset. Jiaji Zhu [9] realized the rapid detection of pesticide in tea by CNN. Since the differences in the Raman spectra of different molecule can be subtle, the use of deep learning algorithms helps distinguish these differences[10]. Thus, the feature-extraction for the Raman spectra of the molecules could be vital to the accuracy of the component identification.

The first stage of the proposed method is establishing the algorithm for the features of the spectra. The identification results were greatly influenced by the quality and accuracy of the training samples. In order to collect all the characteristic peaks, a method based on the Lorenz-peak fitting and second difference was proposed.

The second stage is generating the graphics training samples and training the CNN prediction models. Traditional CNN models for spectra identification usually use 1D samples for training, however, CNN models were mainly applied in image classification[11]. In this study, graphics samples were generated from the characteristic peaks of each component and mixture.

Figure 1 Flowchart of the data augmentation and CNN training

Figure 1 showed the steps of the overall method. The current results showed that the graphics samples-based CNN model maintain the features of the spectra and get relatively high accuracy with few raw spectra data.

2. Methods

2.1. Data augmentation for Raman spectra

The identification of Raman spectra mainly depends on the location of the particular characteristic peaks. While spectra data from different instruments had different scales that are not suitable for deep-learning. The pre-processing method was needed to captured the characteristic peaks of the analytes with an algorithm based on Lorentz peak fitting and differential peak-finding.

The augmentation dataset of spectra was created randomly within the limitation. In this study, 5 percent variation in location and intensity of the characteristic peaks was set for the positive dataset.
Larger variation and exchange of the peaks was applied in the negative dataset. 2% RMSE (Root mean square error) boundary was used to differentiate the positive and negative. Moreover, the negative dataset added some irrelevant signal to enhance the robustness of the model. With the augmentation of the positive and negative dataset, CNN could learn the features of the spectra.

In the situation of low concentration, the dataset of measured spectra was limited, the superposed spectral data with different proportion was applied to simulate the dataset. In consideration of the relation expression of the Raman intensity and the concentration:

$$\lg R = \lg C + \lg A$$

which $R$ representing relative concentration, $C$ representing concentration, $A$ representing constant. Neglecting the self-reversal phenomenon, we could suppose the relationship between concentration and peak intensity follow logarithmic relation, a low change in relative intensity means a high change in relative concentration. In this study, the simulative mixture sample generated from the superposition of spectra varying in $[0.5,1.0]$ proportion.

2.2. Dataset partition and graphic samples generation for identification

Since the CNN models show advantages in the performance of image recognition, the graphic samples utilized the performance of the deep learning model. Graphics based sample could get higher dimension feature map compared with 1D spectra serial. Figure 2 showed how Raman spectra transfer into a graphic sample, the dark vertical columns in the samples stand for the location and the intensity of the characteristic peaks of the simulative spectra data. For the convenience of the model training, the pre-processing of the compound normalized the spectra into the same size to avoid the influence of different measured instruments and errors.

![Figure 2](image)

**Figure 2** Illustration of how graphics sample generate from Raman spectra.

With the data augmentation, the number of each mixture spectra sample were up to 2000 for positive and negative dataset, among 4000 samples, 70% would be the training set and the rest of 30% would be the validation set. The mixture dataset contained the positive partition made up with mixed spectra of each component and the negative partition composed of spectra of other molecules and spectra with large disturbance.

Figure 3 showed the training logic of a trinary-component identification model. For each component, features of possible mixture Raman spectra were concerned. In the model for $A$, the positive dataset contained the Raman spectra in the presence of $A$ while the negative dataset contained the spectra that $A$ is absent. A new sample would be tested by each model to predict its component.

2.3. Convolutional neural network for component identification

CNN is a representation algorithm of deep learning, it consists of neurons with learnable weights and bias. CNN has 3D volumes of neurons, which makes the feedforward function more efficient. CNN features shared weights and sparse connectivity, providing high efficiency, strong generalization and sensitive response to local input pattern. As the prediction results of the model could be differentiated by the Softmax function:
The results would be classified into $n$ classes. In this study, $n$ would be 2, representing the positive and the negative.

$$\text{softmax}_k = \frac{e^{W_k^T x}}{\sum_{j=1}^n e^{W_j^T x}}$$  \hspace{1cm} (2)

The logic of component identification model training in ternary mixture.

CNN model has several important parts for component identification. The input layer introduces the samples for training. The following layer is convolutional layer. The convolution is performed by sliding a kernel over the entire image with a certain step called ‘stride’ to produce a new representation of the image that representing the features. Then, the output feature map is passed to an activation function to implement nonlinearity transformation within the network layer. Pooling is a layer that could reduce dimensionality of the outcome of a convolutional layer to drop down the amount of the parameters and computation. Full connection layer plays a role of classifier in convolutional neural networks. It is usually placed at the end of the network.

To avoid over-fitting and enhance the generalization ability of the model, Dropout function and Adam optimizer was applied. The CNN model illustration is shown in Figure 4.

2.4. Measurement of prediction accuracy

To evaluate the prediction accuracy of the method proposed and other method, accuracy (ACC), true
positive rate (TPR, sensitivity) and false positive rate (FPR, specificity) were used in this study. The calculation of these data are as follow:

\[
\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)
\]

\[
\text{TPR} = \frac{TP}{TP + FN} \quad (4)
\]

\[
\text{FPR} = \frac{FP}{TN + FP} \quad (5)
\]

where TP, FP, TN, and FN represent true positive, false positive, true negative and false negative. After training, randomly-generated samples of corresponding components would be used as test dataset.

3. Results and discussion

3.1. Accuracy of pre-processed algorithm

To demonstrate the effect of the pre-processed algorithm, Figure 5 showed the fitting result of Lorenz-fitting in carbon tetrachloride sample. As the result showed, Lorenz peaks fitting could get smooth curve of the spectra in the single-peak and multi-peak. Thus, second-order differentiate algorithm could get the characteristics peaks rid of undulating curve.

![Figure 5. Visualization of Lorenz fitting in carbon tetrachloride spectrum](image)

3.2. Accuracy of CNN models in multi-component dataset

Herein, Table 1 showed the validation results of five mixture training models. In each model, test dataset contained 250 positive samples and 250 negative samples. The samples in the test dataset were generated from Raman spectra of corresponding components with the same data augmentation as the training steps above.

| Components   | Prediction | TP   | TN   | ACC(%) |
|--------------|------------|------|------|--------|
| Methanol     | √          | 249  | 246  | 99     |
| Ethanol      | √          |      |      |        |
| Propylene glycol | √      | 248  | 245  | 98.6   |
| Isopropanol  | √          |      |      |        |
| Acetone | √ | 250 | 247 | 99.4 |
|---------|---|-----|-----|------|
| Methanol| √ |     |     |      |
| Toluene | √ | 246 | 243 | 97.8 |
| Ethanol | √ |     |     |      |
| Sulfa   | √ | 239 | 232 | 94.2 |
| Naphthylacetic acid | √ |     |     |      |
| Sodium bicarbonate | √ |     |     |      |

As shown in the Table 1, the prediction accuracy of CNN models in binary-mixtures all exceeded 98% and the accuracy of trinary-mixture reached 94.2%. Obviously, the CNN model showed high prediction accuracy and sensitivity in the samples of varying measurement and concentration.

3.3. Comparison of CNN model with typical classifiers

To verify the performance of CNN model, Table 2 showed the prediction results of CNN model, RF (Random Forest), k-NN (k-Nearest Neighbor) and Naive Bayes. The samples generated based on mixture of acetone and methanol with data augmentation. CNN model showed the best prediction result both in TPR and FPR.

| Method       | TP | TN | FP | FN | TPR(%) | FPR(%) | ACC(%) |
|--------------|----|----|----|----|--------|--------|--------|
| CNN          | 100| 100| 0  | 0  | 100    | 0      | 100    |
| RF           | 98 | 95 | 5  | 2  | 98     | 5      | 96.5   |
| k-NN         | 93 | 89 | 11 | 7  | 93     | 11     | 91     |
| Naive Bayes  | 90 | 86 | 14 | 10 | 90     | 14     | 87     |

4. Conclusion

In this study, a novel method for Raman spectra component identification was proposed. Firstly, a Lorenz-peak fitting based feature detection algorithm was introduced. Secondly, with the augmentation of the Raman spectra dataset, the prediction model could get enough training samples to formulate and learn the various cases in spectra measurement. Thirdly, the prediction accuracy of CNN model was significantly higher than several typical machine-learning classifiers in binary and trinary mixture identification. Moreover, the results demonstrated that the CNN model based on graphics samples bring the advantage of figure feature recognition into Raman spectra classification. In summary, the method proposed provided an effective way in Raman spectra component identification with relatively high accuracy in low data volume.

5. References

[1] Raman C V. A new type of secondary radiation[J]. Nature, 1928, 120(3048): 501-502.
[2] Griffen Julia and Owen Andrew and Matousek Pavel. Comprehensive quantification of tablets with multiple active pharmaceutical ingredients using transmission Raman spectroscopy--a proof of concept study.[J]. Journal of pharmaceutical and biomedical analysis, 2015, 115 : 277-82.
[3] Surface enhanced Raman spectroscopy signals of mixed pesticides and their identification[J].Chinese Chemical Letters,2011,22(12):1477-1480.
[4] Anonymous. The SI Report: Our Progressive Review of DNN[J]. M2 Presswire, 2008.
[5] Tivive Fok Hing Chi and Bouzerdoum Abdesselam. Efficient training algorithms for a class of shunting inhibitory convolutional neural networks.[J]. IEEE transactions on neural networks, 2005, 16(3) : 541-56.

[6] Anonymous. The SI Report: Our Progressive Review of DNN[J]. M2 Presswire, 2008.

[7] M. Ranzato et al. Automatic recognition of biological particles in microscopic images[J]. Pattern Recognition Letters, 2006, 28(1) : 31-39.

[8] Fan X , Ming W , Zeng H , et al. Deep learning-based component identification for the Raman spectra of mixtures[J]. Analyst, 2019.

[9] Jiaji Zhu et al. Rapid on-site identification of pesticide residues in tea by one-dimensional convolutional neural network coupled with surface-enhanced Raman scattering[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2021, 246

[10] Determination of Active Components in a Natural Herb with Near Infrared Spectroscopy Based on Artificial Neural Networks[J]. Chemical Research in Chinese Universities, 2005(01):36-43.

[11] Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning. MIT Press, Cambridge.

Acknowledgments
The authors would like to acknowledge financial support from Natural Science Foundation of China (No.21874113, NO.21974118).