

Deep Learning and Chemometrics: Quantitative and Qualitative Spectroscopy Interpretation of Aqueous Solutions

Colin Murphy

Drexel University: CS-615 Deep Learning

Abstract

Spectroscopy can be simply defined as the measurement of interactions between light and material to better understand the chemistry of the material or system. Analytical chemists have relied on traditional spectroscopy methods for over a century to determine unknown analytes and concentrations. However, the interpretation of spectroscopy data requires substantial domain knowledge and processing. Further complications arise from systematic error of measurements, of which may make it difficult to compare spectra from two different instruments. A generalizable and robust model for analysis of spectral data is needed to achieve rapid and accurate processing for chemometric application. In this study, a Convolutional Neural Network (CNN) machine learning algorithm is trained for NIR spectroscopy classification and concentration prediction of 5 different analytes. The model has achieved high classification performance without additional hyperparameter optimization and outperforms traditional chemometric methods of NIR processing. This simple CNN architecture, along with the absence of data preprocessing of data, is intended to generalize the raw spectrum.

Background

Within the domain area of Chemometrics, Process Analytical Technology (PAT) methodologies have become useful methods of monitoring the trajectory of processes and gaining novel insight through modeling. With the FDA's promotion of "Quality by Design", it has adopted PAT methods as, "...a system for designing, analyzing, and controlling, manufacturing through timely measurements (i.e., during processing) of critical quality and performance attributes of raw and in-process materials and processes, with the goal of ensuring final product quality" (Hinze 2006) thus, PAT has been proposed in industry quality plans.

Among the most widely used methods for process monitoring is spectroscopy - the measurement of interactions between light and material. There are various types of spectroscopy measurements, they are generally applied due to the low cost, well established theoretical background, and their ability to provide both quantitative and qualitative information of an analyte.

Many areas of AI application exist within spectroscopy, however, the focus of this study will consider spectroscopy in relation to PAT. The application of deep learning techniques will focus on Near-Infrared Spectroscopy (NIR).

Spectroscopy has been widely used for centuries and the interpretation of spectra data can be difficult. NIR spectra are broad, up to 100–150 nm wide (Walsh, Guthrie, and Burney 2000), with the spectrum being broad bands of overlapping absorption from molecular overtone and combination vibrations.

Previous applications of spectral analysis have required extensive domain expertise - applying feature engineering to extract information such as peak width, peak ratios, and peak gaps. These feature engineering methods still suffer from the realities of non-ideal conditions for measurement - numerous environment factors - which lead to varying levels of noise and spectral

shift of spectra taken at different times (the effect being even greater for two different instruments). This exacerbates the difference between validated reference spectrum (used to identify analyte) and the test spectrum. Thus, a more generalized model is needed to identify minor differences in spectral data for classification and regression.

Related Work

Traditional NIR interpretation methods have relied on varying levels of preprocessing of spectral data, including background subtraction, Principle Component Analysis (PCA), moving average (data smoothing), derivatives, and a method known as multiplicative scatter correction (MSC) (Magwaza et al. 2012).

More complex and robust machine learning models have been applied to spectroscopy as certain algorithms have become popular. The widespread use of Convolutional Neural Networks (CNN) and their application for signal processing has made them ideal for Chemometrics. Liu et al. (2017) applied a CNN architecture for classification of Raman spectroscopy from the RRUFF mineral dataset. Spectra preprocessing was also performed for baseline correction and data augmentation (synthetic chemical shift data).

Recently, Chatzidakis and Botton (2019) applied the same approach as Liu et al. to classify electron energy loss spectroscopy - using a relatively simple CNN architecture and digitized spectroscopy images to train a model resilient to common spectroscopy calibration variations.


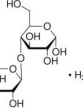
Similar to these aforementioned studies, the work done by Bjerrum, Glahder, and Skov (2017) had applied CNN architectures to NIR data, making qualitative comparisons of the features extracted by individual kernels, and is the primary source of inspiration for this current study.

Methodology

Data Source

The International Diffuse Reflectance Conference (IDRC), organized by the Council for Near-Infrared Spectroscopy (CNIRS), hosts a "shootout" competition for the analysis of NIR spectroscopy using chemometrics and data science methods. The most recent competition, hosted in 2018, focused on *The Application of Aquaphotomics in Data Evaluation* - analysis of the water absorption band in the range of 1300-1600nm (0.5nm resolution). The primary objective of the competition was to predict solute concentration and/or classify the precise solute present in a given spectrum.

The dataset included spectra samples and corresponding experimental conditions, with a traditional calibration curve of samples for each solute in the range of 1-100mM: (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100). Additional experimental data included environmental conditions (e.g. date spectrum was taken, room temperature, and relative humidity).

Analyte	Solutes Included in Dataset		Structure
	Formula	Molar Mass (g/mol)	
Acetic acid (Ace)	CH ₃ COOH	60.052	
Lactose monohydrate (Lac)	C ₁₂ H ₂₂ O ₁₁ ·H ₂ O	360.312	
Sodium-chloride	NaCl	58.44	Na-Cl
Potassium-chloride	KCl	74.548	K-Cl

Two additional spectra datasets were provided, of which included two different preprocessing methods for pure water subtraction, this study will focus on raw spectral data to create a robust generalized model.

The complete dataset and documentation is publicly available and can be found on the CNIRS website: https://www.cnirs.org/content.aspx?page_id=22&club_id=409746&module_id=276203. Dataset summary and documentation provided by CNIRS can be found in supplementary information attachment.

Dataset Preparation

- The datasets provided on the site were already split into subsets: 512 "Calibration" spectra (78.6%, training set) and 172 "Testing" spectra (21.4%, described as a validation set in the documentation, but used for testing in this study). Each provided as a txt file containing both spectral data and experimental data. For classification training, two datasets were concatenated for training and later split using Stratified Cross Validation.
- The TxT files provided were cleaned (datatype conversions, string editing, etc.) and converted to CSV using the Pandas library — the spectra and experimental information were stored in separate dataframes. The original TxT files and CSV files can be found in supplementary information.
- Standardized (zero-mean and unit-variance) the spectral data of each dataset respectively.
- Created one-hot encoded vectors for the solute class of each sample, such that the following vectors can be used for classification training:

Pure Water	[1,0,0,0,0]
Ace	[0,1,0,0,0]
Lac	[0,0,1,0,0]
NaCl	[0,0,0,1,0]
KCl	[0,0,0,0,1]

Experiments and Results

Data Augmentation

Data augmentation is a common method for improving CNN training of images and can be understood as simulating variations of an image that may be easily understood by a human, but can confuse a machine learning algorithm — an image rotated by 90° may not be as easily classified if trained on the original image.

Data augmentation is especially useful for spectroscopy applications, where several translations of the spectrum may occur between measurements (e.g. frequency shifts, peak broadening, and intensity changes). Augmentation of the spectrum was made by randomly offsetting the data by $\pm 0.10 \cdot \sigma$, amplifying by $1 \pm 0.10 \cdot \sigma$, and adjusting the slope randomly between 0.95-1.05. This augmentation was repeated 10 times for each sample and an example output can be seen in Figure 1.

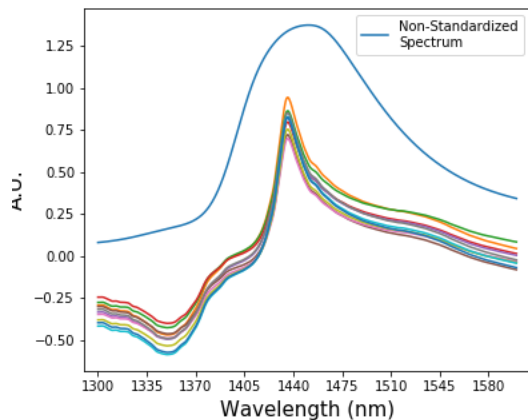


Figure 1: Example of spectra data augmentation

Network Architecture

A similar network architecture as was used by (Bjerrum, Glahder, and Skov 2017) and (Liu et al. 2017) was adapted for this dataset and application.

The spectrum was first passed through a Gaussian noise filter to aid in regularization of the model (Holmstrom and Koistinen 1992), then through two 1D convolutional layer consisting of 8 and 16 filters, respectively — each with a kernel size of 32 and a rectified linear(ReLU) activation. The output of the convolution was flattened and passed through a dropout layer before being passed to a fully connected neural network layer with a ReLU activation and finally the output layer with a softmax activation for classification. A visualization of the architecture can be seen in Figure 2. This same CNN architecture was applied for regression modeling of the analyte concentration — the target being a continuous value (concentration in g/100mL) with a linear activation for the output.

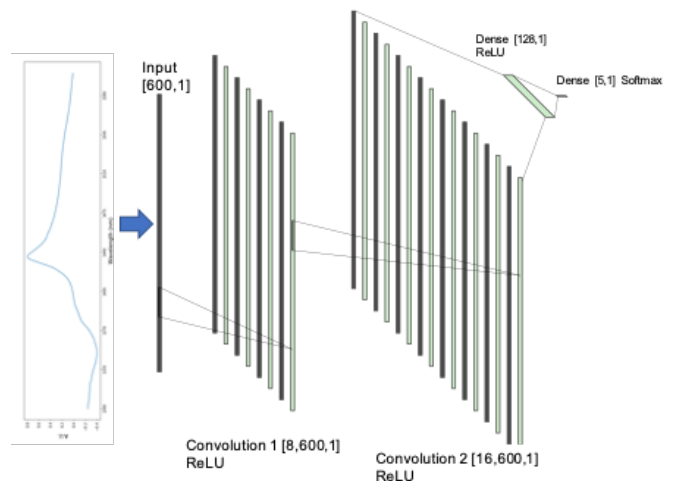


Figure 2: Network Architecture

Multi-Classification Results

The CNN was trained for classification of the 4 different solutes and pure water (5 classes total) with a cross entropy loss function and 5-fold Stratified Cross Validation.

The accuracy impact of data augmentation was evaluated by training the network using the dataset with and then without data augmentation. The results of both implementations are shown in Table 1.

Table 1: Multi-Classification Results

Training Set	Epochs	Batch Size	Accuracy
Data Aug	10	32	82.17 \pm 0.79%
No Data Aug	10	16	71.07 \pm 3.51%

Visualization of the trained model’s feature maps can provide crucial understanding of how a CNN is learning the unique features of an image that allow it to be properly classified.

In Figure 3, the feature maps of the 8 filters that constitute the first convolution layer are reshaped to be visualized as a spectrum. These plots indicate the extraction of the spectrum frequencies corresponding to combinations and overtones of vibrational frequencies of the molecules being classified.

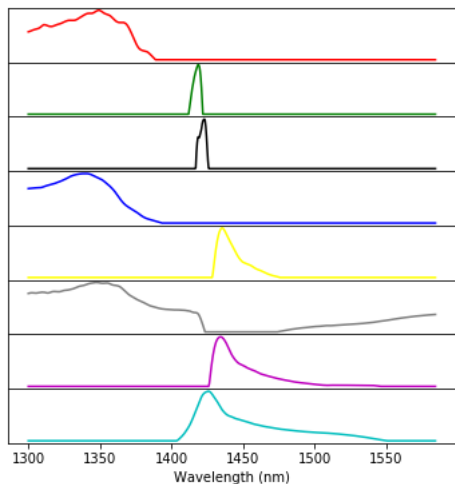


Figure 3: Feature Map Visualization of First Convolution Layer

Regression Results

The CNN was trained for concentration regression for all analytes (concentrations ranging between 0-3.60312 g/100mL) with a Mean Squared Error loss function and 5-fold Stratified Cross Validation.

The Root Mean Squared Error and Huber Error evaluation of the augmented training dataset and the test dataset are summarized in Table 2. The use of the Huber loss function as an error function is intended to better account for outliers, as it is less sensitive than MSE (Faraway 2014). The piecewise Huber loss is defined as follow:

$$L_{\delta}(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for } |y - f(x)| \leq \delta \\ \delta |y - f(x)| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases}$$

The use of Huber loss for training causes outliers to have a linear impact on the gradient, thus a much greater impact on the gradient. In the case that the sample is not an outlier, the function becomes a quadratic (this tolerance being the parameter, δ), at which point it essentially becomes the MSE (Friedman 2001). Thus, it may potentially reach the minimum faster than MSE when handling outliers.

Table 2: Concentration Prediction by Regression

Loss Function	Dataset	R ²	RMSE	Huber ($\delta = 1.0$)
MSE	Training	0.9726	0.1616	0.0124
	Test	0.9643	0.1888	0.0167
Huber Loss ($\delta = 1.0$)	Training	0.9865	0.1174	0.0067
	Test	0.9798	0.1422	0.0097

Conclusions

This application of deep learning for spectroscopy has been applied in a somewhat naïve manner, without implementing significant chemistry domain knowledge to prepare the data — this being the overall goal of the study. The model for classification of spectra is best

if trained for a generalized use, with minimal preprocessing. For qualitative spectra interpretation, a robust model could be used for sample analysis, regardless of the experimental and environmental conditions of the data acquisition (e.g. different instruments, baselines, or resolution).

Ideally, the model’s relative accuracy and robustness could be compared to those reported for the IDRC Shootout competition winners; however, the preprocessing methods and modeling algorithms are not reported, thus future work may be done to compare the CNN model reported here with other NIR datasets. The performance benchmark reported for the competition, RMSE, is also difficult to directly compare, as the documentation does not indicate dataset scale.

Nevertheless, the CNN model applied here exhibits promising classification capabilities for aqueous-based analytes. Having only trained the model using the raw spectrum data (i.e. without subtraction of the pure water spectrum), only applying standardization to the data, and without substantial optimization of hyperparameters, an accuracy of $71.07 \pm 3.51\%$ was obtained. The addition of minor data augmentation improved model accuracy by over 10% and decreased overall error. For concentration prediction, the model exhibited outstanding performance across all metrics and datasets, with the Huber loss function expressing a modest improvement. These results indicate that the data augmentation can replicate the systematic errors that occur in spectroscopy methods.

For additional model implementation details and results please refer to attached source code.

Future Work/Extensions

Although not used for training or data augmentation, the experimental datasets included environmental variables, such as temperature and relative humidity. Such environmental variables are abundant in PAT-integrated monitoring systems, and therefore, they can be investigated for potential model improvement — knowing other state changes of the process could potentially allow the model to discern the root cause of a change in spectrum. Many monitoring systems also utilize an ensemble of PAT instruments to gain insight, with some instruments gathering different spectroscopic information (e.g. UV-vis, Raman, fluorescence) that can be modeled using the methods proposed in this study to build an ensemble modeling structure or multi-headed CNNs.

Considerable space exists for optimization of the network hyperparameters. The impact of different convolution layer parameters (e.g. kernel size and stride) or pooling layers were not investigated in this study. The optimization of which might be best understood by interpreting the features of the spectrum learned by the current kernels.

References

- [Bjerrum, Glahder, and Skov 2017] Bjerrum, E. J.; Glahder, M.; and Skov, T. 2017. Data augmentation of spectral data for convolutional neural network (cnn) based deep chemometrics. *arXiv preprint arXiv:1710.01927*.
- [Chatzidakis and Botton 2019] Chatzidakis, M., and Botton, G. 2019. Towards calibration-invariant spectroscopy using deep learning. *Scientific reports* 9(1):1–10.
- [Faraway 2014] Faraway, J. J. 2014. *Linear models with R*. CRC press.
- [Friedman 2001] Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* 1189–1232.
- [Hinz 2006] Hinz, D. C. 2006. Process analytical technologies in the pharmaceutical industry: the fda’s pat initiative. *Analytical and bioanalytical chemistry* 384(5):1036–1042.
- [Holmstrom and Koistinen 1992] Holmstrom, L., and Koistinen, P. 1992. Using additive noise in back-propagation training. *IEEE Transactions on Neural Networks* 3(1):24–38.
- [Liu et al. 2017] Liu, J.; Osadchy, M.; Ashton, L.; Foster, M.; Solomon, C. J.; and Gibson, S. J. 2017. Deep convolutional neural networks for raman spectrum recognition: a unified solution. *Analyst* 142(21):4067–4074.
- [Magwaza et al. 2012] Magwaza, L. S.; Opara, U. L.; Nieuwoudt, H.; Cronje, P. J.; Saeys, W.; and Nicolai, B. 2012. Nir spectroscopy applications for internal and external quality analysis of citrus fruit—a review. *Food and Bioprocess Technology* 5(2):425–444.
- [Walsh, Guthrie, and Burney 2000] Walsh, K. B.; Guthrie, J. A.; and Burney, J. W. 2000. Application of commercially available, low-cost, miniaturised nir spectrometers to the assessment of the sugar content of intact fruit. *Functional Plant Biology* 27(12):1175–1186.