Article

NIR news

Perspectives on deep learning for near-infrared spectral data modelling

NIR news
2022, Vol. 33(7–8) 9–12

DOI: 10.1177/09603360221142821
journals.sagepub.com/home/nir


Dário Passos[1] and Puneet Mishra[2]

Abstract

Deep learning for near-infrared spectral data is a recent topic of interest for near-infrared practitioners. In recent years, applications of deep learning are flourishing from analyses of point spectrometer data to hyperspectral image analysis. However, there are also some cases where simple partial least-squares based models are sufficient. This paper provides a concise view of the state of the art of deep learning for near-infrared data modelling, particularly discussing when deep learning is useful. Discussion is also provided on what is already achieved and what ideas would be interesting to pursue regarding deep learning modelling of near-infrared data.

Keywords

Artificial intelligence, neural networks, non-linear

Introduction

Near-infrared (NIR) spectroscopy is a technique that allows to infer information about diverse physical systems ranging from distant galaxies to molecular interactions in biological micro-organisms. NIR has been applied to many tasks such as high-end pharmaceutical manufacturing to food quality control including field analysis of agricultural produce. Decades have passed since NIR spectrometers were first developed and utilised. However, in recent years, the scientific community and commercial spectrometer manufacturers have worked towards improving both the instruments as well as the multivariate calibration modelling approaches for NIR data. One of the major technical developments is the miniaturisation of NIR spectrometers and consequently their ease of use. Nowadays, NIR spectrometers are available as pocket devices which are fully standalone or can be operated with smartphones. Additionally, due to recent developments in cloud computing approaches and high-speed internet connection, heavy model computations and deployments for doing predictive modelling can also be performed anywhere rapidly from lab to remote locations. Stability is still an issue in some of these handheld devices and the quality of the signal they acquire is still not comparable to benchtop instruments which are more stable but much more expensive. Consequently, spectral data obtained by these portable instruments are more prone to noise (e.g., due to instrumental thermals) and other sources of variability (e.g., operator handling), requiring more attention on how the data are processed during the analysis. With proper exploration of pre-processing methods, outlier detection, wavelengths selection and optimization of models based on partial least-squares, calibration models can be obtained. Notwithstanding some specific cloud analysis services, this analysis pipeline is still time consuming and requires expert knowledge to correctly implement.

Recently, with the advancement in artificial intelligence (AI), deep learning (DL) techniques have gained wide traction as a tool for all purposes. From popular augmented reality apps available in smartphones to the most complex fields of sciences, DL is having a significant impact and it keeps gaining momentum. Some of the most attractive features of DL algorithms are their extraordinary ability to extract hidden patterns from data and to efficiently encode information into complex latent variables (akin to a space of concepts that is exceedingly difficult to hardcode). Translation between different languages or image descriptions are two clear examples of these capacities. In the latter, DL algorithms can identify different objects in an image and describe what is on the scene based on complex concepts such as the contextual information of objects, its position, characteristics, etc. Given their popularity, the chemometric community working with NIR

[1]CEOT and Physics Department, Universidade do Algarve, Faro, Portugal
[2]Wageningen Food and Biobased Research, Wageningen, The Netherlands

Corresponding author:
Puneet Mishra, Wageningen Food and Biobased Research, Bornse Weilanden 9, P.O. Box 17, 6700AA, Wageningen, The Netherlands.
Email: puneet.mishra@wur.nl

spectra has also started to experiment with these types of algorithms in their research.[1]

Deep learning for NIR data modelling

The huge progress experienced by DL, as a core subject of computer sciences, is a result of gradual research process over many years that involved entire scientific communities dedicated to developing algorithms aimed at solving very specific problems, e.g., computer vision (CV). In parallel, other challenges such as protein folding or autonomous driving are being addressed by large, well-funded and highly focused teams, for example, in Deepmind and Tesla. Advances in other scientific areas such as chemometrics are lagging. In NIR chemometrics, DL adoption is still in an "experimental phase" where existing DL architectures are being applied to different tasks; however, these have not yet reached a high level of maturity and acceptance. Common and simpler DL architectures such as Convolutional Neural Networks (Figure 1) and Auto-Encoders are by far the most used for NIR chemometric modelling. As with all well-established areas (including chemometrics) the adoption of new algorithms and analysis methodologies in the current workflows is sometimes slow. In chemometrics, specifically, there are several reasons for this: most of the linear models used in chemometrics work remarkably well for most of the problems they have to solve, DL implementations require a steep learning curve which may not compensate the performance gains in terms of modelling metrics, chemometricians have seen similar hypes in the past with the previous wave of machine learning algorithms (Random forests, Support Vector Machines, Classical Artificial Neural Networks, etc.), dataset size is still viewed as a potential problem for the application of this type of algorithms, and finally the lack of proven DL architectures specifically aimed at spectral analysis. These are legitimate concerns that can be explained by the fact that the goal for most chemometricians is to solve a specific task that is related to chemical/physical characterization of their samples. Notwithstanding a small percentage of chemometricians dedicated to algorithm development, most of the research in this area is done by using algorithms as tools that allow finding the answer to a specific problem (e.g., in analytical chemistry, food quality, industrial monitoring, etc.). Unlike certain multivariate analysis methods that evolved or were created for chemometric tasks, DL models developed for other areas are being repurposed and applied to chemometrics applications. One can see this as the first wave of DL works associated with chemometrics where current existing solutions/architectures are explored. Despite a certain lack of specialization, some very encouraging results have emerged in the literature.

What is already achieved?

In different areas that use NIR spectra, DL algorithms have shown improved model-predictive performance when compared to classical chemometric methodologies. This is more evident in the availability of large datasets containing thousands of samples. In a small sample set, the main performance improvement can be noted when the response variables are complex, such as multiple continuous responses or classifications. Recently, in the chemometric literature, wide applications of DL such as calibration transfer, model update, multiblock modelling for data fusion and spectral image processing have been demonstrated. Furthermore, different articles have also demonstrated how the chemometric knowledge and DL can be used complementarily to both understand the system and improve the predictive power. These early results further

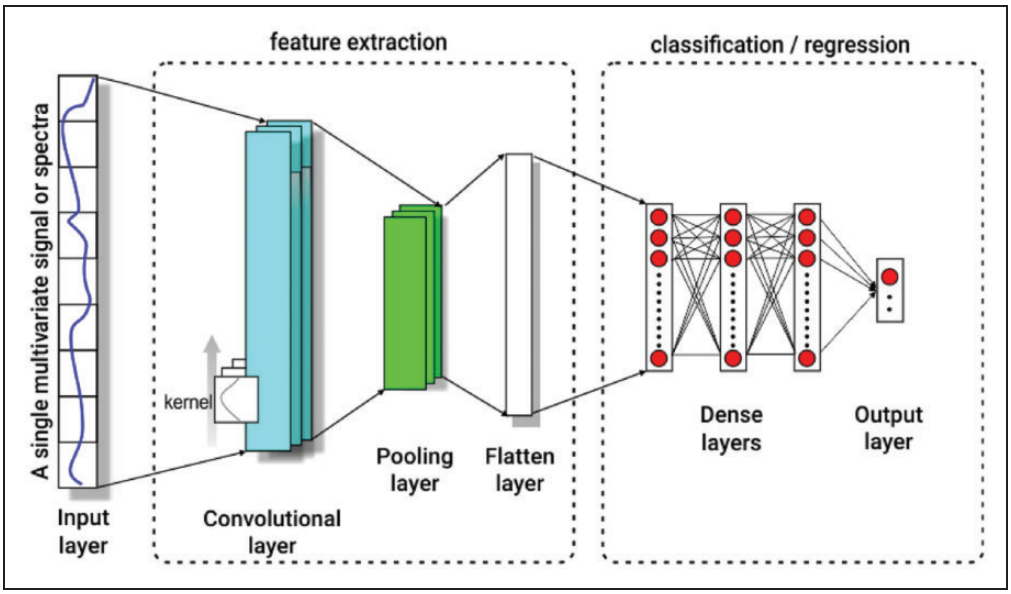

Figure 1. Representation of a typical Convolutional Neural Network architecture used for spectral analysis.

sparked interest and concerns around the topic. Mishra et al.[1] and Zhang et al.[2] provided recent reviews about the current state of DL applied to NIR data modelling. These reviews cover most of the recent results including the most used DL architectures, data-augmentation proposals, interpretability of the models, implementations of model automatic optimization and some of the open questions. Passos and Mishra[3] also provided open access codes to do DL on NIR data for regression and classification tasks combining automatic optimisation of hyperparameters (Figure 2). These are definitely a very nice starting point for any reader that wants to learn more about the subject and start practising DL for NIR data modelling.

What still needs to be explored?

Currently, there is lack of focused research towards the development of DL aimed specifically at solving NIR spectroscopic problems. On a more fundamental level, recently, DL models aimed at dealing with the root causes of nonlinearities in NIR spectra are being explored. Such models aim to automatically find the most suitable pre-processing that minimizes several types of these interferences. There is also work initiated around the hybridization between PLS and DL models, where PLS is assumed to account for the linear part of the signal and the DL model with the nonlinear residual. Beyond these initial steps, a fountain of ideas can be found in the recent literature on DL applied to image analysis. Interesting concepts and properties observed in some DL architectures applied to CV problems can serve as drivers for a focused second wave of development. Some DL architectures present what is called "spatial invariance" for object identification, where the object itself can be in any position and orientation in an image, meaning that the DL model is able to create a concept (i.e., an abstraction) for that type of object and identify it independently of its position. This is like spectra taken by different spectrometers and subjected to shifts in wavelength. Is there a DL architecture capable of dealing with the problem of spectral shifting and present the same type of invariance properties mentioned? Most language models that are based on a specific DL architecture, Transformers, can translate between different languages by encoding the sentences of a certain language into a latent space of features related to the meaning of that sentence, and then find the closest latent features generated by other languages and perform the translation. Could something analogous be used to "translate" spectra between different spectrometers and simplify the process of calibration transfer? In CV, there are also DL architectures capable of performing "style transfer," where high-frequency features (e.g., textures) and low-frequency features (e.g., large shapes) can be separately extracted from an image and applied to another. This allows, for example, to convert a photo into a painting on the style of a famous painter. Could an analogous concept be developed for NIR spectra and used in "calibration transfer" tasks as well or in spectra separation? These questions are just a few examples of possible research suggestions, that might (or might not) be feasible. A different research direction could be related with the inclusion of expert knowledge (chemical or physical) into DL models to help modulate their behaviour by imposing constraints. This has been successfully done in Physics with the development of physics-informed neural networks that are nowadays revolutionizing the field of computational fluid dynamics.

Dataset size is considered by some as one of the possible bottlenecks for the development of DL architectures in chemometrics. It is a known fact that the large size of the datasets used in CV and natural language processing (NLP) were crucial for the development of larger and more complex architectures that constitute the current state of the art. The creation of NIR spectral datasets (labels included) of the proportions of image datasets is not on the foreseen horizon

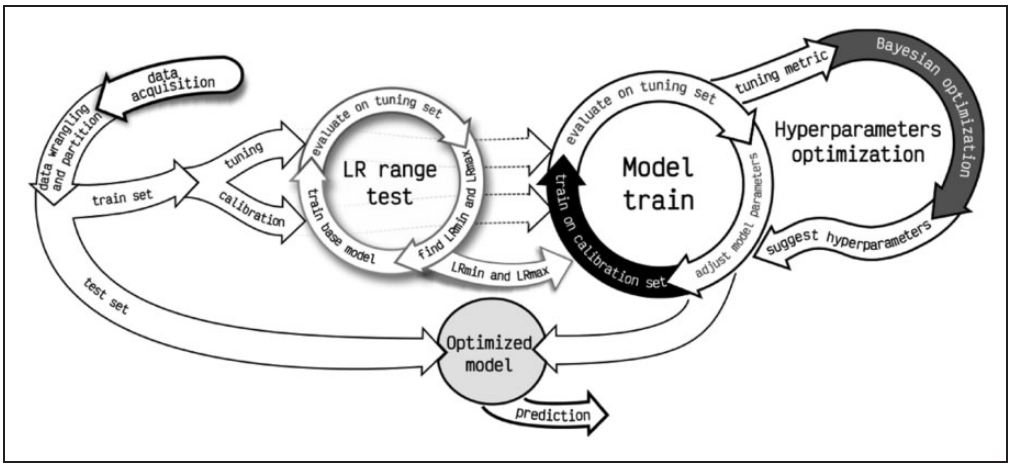

Figure 2. Schematic representation of an automatic optimization pipeline for deep learning models. It includes train/test split of the data, a learning range (LR) test and model hyperparameters optimization loop.

because the cost of labelling the spectra (the wet chemistry analysis of the samples) is orders of magnitude more expensive and time consuming than labelling an image. Nonetheless, there are some alternatives that could help with this process of building large NIR spectra datasets. Besides the development of custom spectral data augmentation techniques, data could be produced by simulations. Monte Carlo simulations of light scattering in diverse types of parametrised materials or biological tissues can be used to produce large synthetic datasets for DL models training. Also, quantum systems simulations are also advancing rapidly, and it is possible that in the coming years, this type of simulations can produce NIR template spectra for many molecules and compounds.[4] This information can then be assimilated by DL models and used as constraints to real spectra in the training of calibration models.

Conclusions

At this point in time, DL applications to NIR data modelling are still new, sparsely distributed among applications to different areas and a bit unfocused regarding algorithmic development. Some DL architectures, especially from CV, seem to present interesting properties regarding image-related tasks that could find parallels in NIR spectral analysis and solve, or aid to solve, some of the known problems behind chemometric modelling. It would be very interesting to see what kind of developments around this topic could be achieved by multidisciplinary teams that combine computer science researchers with chemists and physicists. One of the important steps behind the boom seen in CV- and NLP-related DL models was the development of standardized image and text corpus datasets that different groups around the world could use to benchmark their models. This seems like a logic step to be implemented by the chemometrics community as well, one that would allow a direct comparison on how different novel chemometric-DL algorithms perform on different tasks.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Dário Passos acknowledges FCT – Fundação para a Ciência e a Tecnologia, Portugal, for funding CEOT project UIDB/00631/2020 CEOT BASE and UIDP/00631/2020 CEOT PROGRAMÁTICO.

References

1. Mishra P, Passos D, Marini F, et al. Deep learning for near-infrared spectral data modelling: hypes and benefits. *TrAC Trends Anal Chem* 2022; 157: 116804.
2. Zhang X, Yang J, Lin T, et al. Food and agro-product quality evaluation based on spectroscopy and deep learning: a review. *Trends Food Sci Technol* 2021; 112: 431–441.
3. Passos D and Mishra P. A tutorial on automatic hyperparameter tuning of deep spectral modelling for regression and classification tasks. *Chemometr Intell Lab Syst* 2022; 223: 104520.
4. Beć KB and Huck CW. Breakthrough potential in near-infrared spectroscopy: spectra simulation. A review of recent developments. *Front Chem* 2019; 7: 48.