

Use of conditional generative variational autoencoder networks to improve representativity of data in optical spectroscopy

Alexander Efitorov^{a,*}, Sergey Burikov^a, Kirill Laptinskiy^a, Tatiana Dolenko^{ab},
Sergey Dolenko^a

^aSkobeltsyn Institute of Nuclear Physics, Lomonosov Moscow State University, Leninskie gory 1(2), Moscow, Russia

^bDepartment of Physics Lomonosov Moscow State University, Leninskie gory 1(2), Moscow, Russia
E-mail: a.efitorov@sinp.msu.ru, dolenko@srd.sinp.msu.ru

In this study, the solution of the inverse problem of spectroscopy of water-ethanol liquid solutions by neural network models is considered. The process of training a neural network requires a large number of patterns, which cannot be obtained by laboratory measurements. In this paper, we demonstrate the possibility of generating an additional array of patterns using a conditional variational autoencoder. The generated patterns have a form similar to real spectra, and they are used to train the neural network for classification, along with the original patterns. As a result of applying this approach, it was possible to improve the quality of solving the inverse problem on real patterns that were not used in the training process.

*The 5th International Workshop on Deep Learning in Computational Physics (DLCP2021)
28-29 June, 2021
Moscow, Russia*

*Speaker

1. Introduction

In recent years, artificial neural networks have shown significant success in various applied problems of real world [1]. Although active research in the field of computer science is still ongoing [2], the application of neural networks has gone beyond academic tasks formulated on large datasets collected from the global web (images, video, texts, voice samples) [3], and appears in production fields that use data obtained in laboratory experiments which require expensive equipment, qualified staff for operating and experts to make the annotation of the patterns. To minimize the efforts required to collect representative datasets, sparse grids of measuring conditions\parameters are generated to cover a wide range of possible realizations. In case of classical deep learning tasks, there is a well-known set of recommendations [4], such as adding augmentations to original sample (noise adding, affine and brightness transformations etc.), the use of a semi-supervised learning, training on additional samples from open sources etc. Unfortunately, there is no way to use such recommendations for many specific applications, for which there are no relevant data from another sources, and the type of permissible augmentations is not known. An example of such specific problem (the inverse problem of Raman spectroscopy of water-ethanol solutions) is considered in this article. The specificity of the object of research is determined by a set of factors: the properties of measuring equipment (the characteristics of laser, diffraction grating, the CCD matrix of the detector etc. [5]) and the subject of the study - water-ethanol solutions with impurities poisonous in a certain concentration range. In laboratory conditions, more than 5 thousand spectra were measured with various concentrations of the investigated components. This number of patterns already makes it possible to train the baseline ML-model, but further improvement of its results requires additional tricks. It is impossible to implement the augmentation strategy for spectral data, since the Raman spectrum of the low-frequency band has narrow strictly localized peaks of natural oscillations, whose intensity is very sensitive to the concentration of presented components, and their intensity has a nonlinear dependence on simultaneous presence of components dissolved in the solution. It should be noted that although there is still no theoretical model that could accurately describe the Raman spectra of water solutions, an empirical calibration model for 1-2 component solutions was created 4 decades ago [6], and the possibility of using neural networks [7] and other ML models was also demonstrated [8] for more complex Raman spectroscopy inverse problems.

The idea of generating additional spectra using neural networks was proposed in [9]; however, the rapid development of the topic of generative adversarial neural networks (GAN) [10] and variational autoencoders (VAE) [11] made it possible to obtain significant benefits from the exploitation of this approach in the application of spectroscopy problems [12] by using the conditional versions [13] of these models (cGAN and cVAE respectively). The classical-ML way of processing spectral data is the construction of the so-called “chemometrical” linear models (based on Principal Component Analysis [14], Partial Least Squares [15]), which allow one to build stable models even on small amounts of data by considering only significant eigenvalues, but have a limited accuracy due to nonlinear effects appearing in complex subjects.

The object of our research are water-ethanol solutions, additionally containing impurities (fusel oil, methanol, ethyl acetate), the concentrations of which (max concentrations 11.2%,

24%, 18%) vary in intervals similar to those present in alcoholic beverages, as well as the concentration of ethanol (which is much higher and varies in the range 38-57%). The observed pattern solutions may not contain additional impurities at all (water-ethanol only) or any combination of them (from 1 to 3 impurities at the same time).

2. Methods

2.1 Raman spectroscopy

Each pattern was recorded with laser excitation at 532nm, power 300 mWatt in two ranges - in the low-frequency region (190-2230 cm^{-1}) and in the area of the valence band of water (2170-3840 cm^{-1}), resolution 2 cm^{-1} . In each range, there were 1024 spectral channels.

2.2 Neural network classifier

To solve the inverse problem of spectroscopy, a fully connected NN with 2 hidden layers (128 and 64 neurons, respectively) was used. This NN model had 8 outputs, each output corresponding to a unique combination of components present in the explored water-ethanol solutions. To minimize possible overfitting of the NN, the Dropout [16] regularization approach was used, which provided random disconnection (with a probability of 50%) of individual neurons of the hidden layers during training; also, for the same reason, early stopping mechanism was implemented: the NN training process was stopped if during a defined number of epochs there was no improvement of the average loss function calculated for the pairs (predicted class vector - ground vector for the corresponding pattern) over the validation dataset.

2.3 Conditional Variational Autoencoder

A vanilla variational autoencoder consists of a pair of encoder-decoder NN models that map the original input patterns into a lower dimension latent space, and implement a generation process based on sampling from this lower dimension space. Unlike classical autoencoders, which consist of a single NN model and are trained on estimation of the quality of reconstructed samples passed through the bottleneck of a neural network (a layer of the autoencoder NN with a small number of neurons), VAE attempts to create a lower-dimension latent space with good statistical properties that enable the generative process. For VAE architecture, the encoder creates an embedding not of a single point, but of the corresponding distribution (determining the parameters of the normal distribution) to the latent space. The decoder is trained to generate patterns from a vector sampled from the corresponding distribution in the latent space. Because we are dealing not only with reconstructed patterns, but with distributions, an additional regularization term to the loss function is added to estimate the Kulback-Leibler divergence [17] between the approximated and the target distributions. In order to add information about the target class, in the conditional VAE, the concatenation of the one-hot encoded output vector with a vector sampled from the latent space is introduced during the training process. Thus, after completing the training of the model, it's possible to use the decoder model standalone to generate spectra corresponding to target classes. The examples of such generated spectra are presented in Fig 1. We also experimented with conditional generative adversarial networks; however, was hard to achieve stable convergence of the training process, and the generated

spectra were significantly corrupted by noise, which led to the loss of information in some important (according to physics) spectral regions. This leads to decrease in the quality of the trained model.

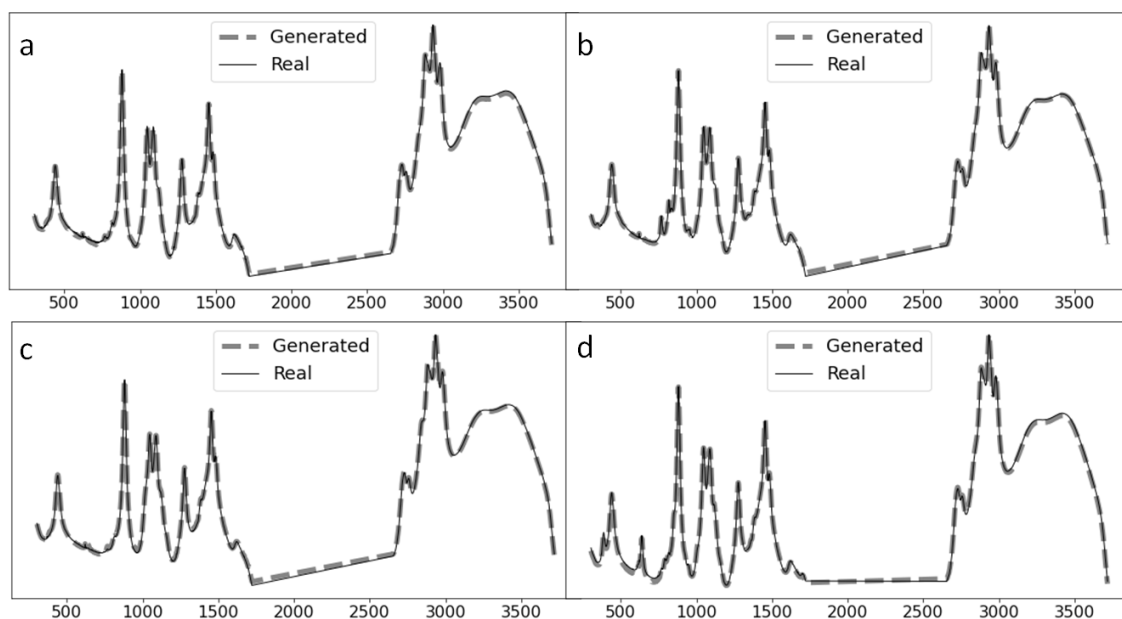


Figure 1: A comparison of the spectra (observed in laboratory and generated by cVAE, \log_{10} processed) of water-ethanol solutions with different substances: a - no impurities, b - fusel oil, c - methanol, d - ethyl acetate.

3. Data processing

As a part of the routine of collection and initial processing of the raw data recorded by laboratory equipment, the spectral baseline correction and normalizing to the integral of the valence band of water were performed. Also, uninformative spectral regions were excluded, since no intensity changes corresponding to dissolved components could be observed there. The \log_{10} and rescaling to the range (0, 1) were applied as the next steps, to reduce the extreme variation of regions of natural spectral oscillations. Since different classes have a different number of possible component concentrations (for solutions of 2 components - 560 spectra were measured, for solutions of 3 components - more than 3000), an equal number of patterns from each class was randomly selected to balance the classes for the training set (used to adjust the weights of NN model) and the test set (used only for independent estimation of the quality of the NN model, after the training was ended). This procedure was performed 5 times to eliminate the dependence on the empirical distribution of the selected patterns. The final training and test sets consisted of 3560 and 440 unique spectral patterns, respectively.

POS(DLCP2021)013

4. Computational experiments

According to the described pipeline, the cVAE and classifier NNs were trained. cVAE was fitted on a training dataset consisting of real spectra only. After the cVAE training was finished, 1100 examples were generated for each class by its decoder NN model. Then three different NN classifiers were trained: the first classifier was trained on real data (the same dataset that was used to train the cVAE model), the second classifier - only on simulated data generated by the decoder model, the third classifier - on a mixed dataset (generated + real spectra). The accuracy of the classifiers after training was estimated on the test set (patterns that were used neither for training cVAE nor for training the classifier NNs). The early stopping criteria - the loss function (categorical cross entropy) computed on the validation set (20% of patterns randomly selected from the training set) was not improved during 200 consecutive epochs of training process; the maximum possible total number of epochs was 2000, but it was never achieved. The network accuracy values for various cases are presented in Table 1.

It should be noted that similar performance of classifier NNs observed in various experiments (for various splits) demonstrates that the effect obtained reflects a stable tendency, despite the relatively low values of accuracy. However, the proposed approach significantly improves the classification accuracy on an independent test dataset of real patterns. As expected, the second classifier model, trained only on the generated patterns, demonstrates the lowest performance (classification accuracy less than 30%), although during the training process the accuracy on the training set and on the validation set (consisting of generated patterns only) was greater than 90%. However, despite such a strong difference in the results computed on artificial and real datasets, the model trained on a mixed dataset demonstrated fast convergence (after about 400 training epochs) and higher accuracy on the test dataset. The best results for the first classifier model (trained on real samples only) required more than 1000 training epochs and still had lower accuracy. We can say that to some extent adding generated patterns provides some kind of augmentation for the initial experimental data array.

	Dataset		
	Real	Generated	Mixed
Split 1	0,33	0,27	0,37
Split 2	0,36	0,28	0,4
Split 3	0,35	0,29	0,41
Split 4	0,37	0,3	0,43
Split 5	0,38	0,29	0,42
Average	0,358	0,286	0,406
Standard deviation	0,019	0,011	0,023

Table 1. The accuracy of solving the 8-class classification problem of Raman spectroscopy using an artificial neural network. Types of training datasets: Real - spectra recorded in the laboratory experiment, Generated - spectra generated by the cVAE, mixed - dataset including real and generated patterns. The training set is a class-balanced, random subsample of laboratory experiment patterns. Similar patterns within the same split were used to train cVAE and classifier models. All values are computed for real spectra of the test set, which were not used for model training.

5. Conclusions and future work

In this paper, a successful application of the approach of generating Raman spectra using a conditional variational autoencoder for the subsequent training of a neural network model for classifying multicomponent water-ethanol solutions has been demonstrated. Despite the complexity of the object of study (nonlinear interactions of the components, a strong spread in the concentrations of different components, which differently affect the shape of the observed spectrum), we succeeded to achieve a significant improvement in the performance of the classifier NN model within the proposed approach. Further development of this work is the transition from one-hot encoded unique classes to the concentrations of the components simultaneously present in the studied samples. To obtain a significant difference between spectra of different substances and their different concentrations, it is supposed to use conditional GAN networks. However, to guarantee similarity with real spectra, the weights of the decoder of a pre-trained conditional VAE could be used as the initialization of the weights of the generator network.

6. Acknowledgements

This study has been performed at the expense of the grant of the Russian Science Foundation, project no, 19-11-00333

References

- [1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature* 521, pp. 436–444 (2015)
- [2] M. Abdar et al., A review of uncertainty quantification in deep learning: Techniques, applications and challenges, *Information Fusion*, volume 76, pp. 243-297 (2021)
- [3] Russakovsky, O., Deng, J., Su, H. et al. ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis* 115, 211–252 (2015)
- [4] I. Goodfellow, *Deep Learning (Adaptive Computation and Machine Learning series)*, MIT Press, (2016)
- [5] M. Pelletier, Quantitative Analysis Using Raman Spectrometry, *Appl. Spectrosc.* 57, pp. 20A-42A (2003)
- [6] K. Dean, G. Wilkinson, Precision Raman investigation of the ν_1 mode of vibration of SO_2 -4, WO_2 -4 and MoO_2 -4 in aqueous solutions of different concentrations, *Journal of Raman Spectroscopy*, volume 14(2), pp. 130-134 (1983)
- [7] I. Gerdova, S. Dolenko, T. Dolenko, I. Persiantsev, V. Fadeev, I. Churina, New opportunity solutions to inverse problems in laser spectroscopy involving artificial neural networks, *Izvestiya Akademii Nauk SSSR Seriya Fizicheskaya*, volume 66 (8), pp. 1116-1124 (2002)
- [8] S. Burikov, S. Dolenko, T. Dolenko, I. Persiantsev, Application of artificial neural networks to solve problems of identification and determination of concentration of salts in multi-component water solutions by Raman spectra. *Optical Memory and Neural Networks*, volume 19, pp. 140-148, (2010)
- [9] S. Dolenko, S. Burikov, T. Dolenko, I. Persiantsev, Adaptive methods for solving inverse problems in laser Raman spectroscopy of multi-component solutions. *Pattern Recognition and Image Analysis*, volume 22, (2012)

- [10] I. Goodfellow et al. Generative Adversarial Nets. Proceedings of the International Conference on Neural Information Processing, pp. 2672–2680 (2014)
- [11] Kingma, Diederik P. and Max Welling. “Auto-Encoding Variational Bayes.” CoRR abs/1312.6114 (2014)
- [12] K. Davaslioglu and Y. E. Sagduyu, Generative Adversarial Learning for Spectrum Sensing, IEEE International Conference on Communications, pp. 1-6, (2018)
- [13] Yan X., Yang J., Sohn K., Lee H. (2016) Attribute2Image: Conditional Image Generation from Visual Attributes. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV Lecture Notes in Computer Science, volume 9908. Springer (2016)
- [14] I.T. Jolliffe. Principal Component Analysis. Springer, New York, (1986)
- [15] M. Hubert and K. Branden, Robust methods for Partial Least Squares Regression. Journal of Chemometrics, volume 17, pp. 537-549. (2003)
- [16] N. Srivastava et al., Dropout: A Simple Way to Prevent Neural Networks from Overfitting, The Journal of Machine Learning Research, volume 15(56): pp. 1929–1958 (2014)
- [17] D. MacKay, Information theory, inference, and learning algorithms, Cambridge University Press, Cambridge, UK, (2003)