Università degli studi di Milano-Bicocca

Advanced Machine Learning

Final Project

Deep Learning-Enabled Decoding of Raman Spectroscopy

Authors:
Nina Singlan - 867646- n.singlan@campus.unimib.it
Romain Michelucci - 867644- r.michelucci@campus.unimib.it

January 22, 2021

Abstract

Raman spectroscopy is commonly used in chemistry to provide a structural fingerprint by which molecules can be identified. Machine Learning methods are used to decode these spectra. The purpose of this study is to identify patients infected with the Amyotrophic Lateral Sclerosis (ALS) by analysing their Raman spectrum with Machine Learning methods. This report present a Transfer Learning (TL) approach combined with data augmentation techniques. The main problem for an efficient classification is the small number of data, indeed, medical data are subject to "medical secrecy". Regarding the low number of data, the use of a pre-trained model, trained on a big amount of bacterial Raman spectrum, seems to be a good solution to obtain a better classification.

1 Introduction

Raman spectroscopy ally with methods of Machine Learning promise a new rapid and non-invasive technique to diagnose patients. This capacity leeds to a certain number of search in this domain showing that Convolutional Neural Networks (CNN) in the context of Raman spectroscopy is outperforming other machine learning methods with baseline corrected spectra. However, CNNs are "data-hungry", they require a massive amount of data in order to extract the right features. Usually, a CNN ca be well-trained on a few thousands of sample.

Unfortunatelly, the dataset used is composed of less than 600 spectra, which is a not sufficient to properly train a model. In addition to this, a problem lies in the repartition of the number of samples per patients and per groups. Transfer learning aims at improving the performance on a target domain by transferring the knowledge contained in different but related source domains. Due to the fact that the dependence on a large number of target domain data can be reduced, TL has become a popular and promising area in machine learning.

In this work, the effort is focused on trying different transfer learning experiments. The first consists in fine-tuning which consists of unfreezing a pre-trained models and re-training it on the new data. This can potentially achieve meaningful improvements by incrementally adapting the pre-trained features to the new data. The second experiment is the most common incarnation of transfer learning : after taking the layers from a pre-trained model, the layers are freezed and some new trainable layers are added on top of the frozen ones. These new layers will learn to turn the old features into predictions on a new dataset.

Finally some data augmentation methods will be applied. It is a well-known technique for improving robustness and training of neural networks. The idea is to expand the number of training samples by adding some noise and small variations resulting in a more robust training. For spectral data, random offsets, multiplication and Gaussian noises are commonly used.

2 Datasets

The dataset provided is composed of 591 spectra, beloging to 30 patients, divided into two classes : patients affected by ALS and healthy ones (CTRL). Only 393 spectra are provided for the ALS patients and 198 for the CTRL ones. Thus, it is umbalanced as it is shown in table 1. Furthermore, all the spectra belonging to a same patient are correlated between each other, so it is necessary to treat them together.

Thanks to Dario Bertaziolli who provide the data, they are already filtered and pre-processed. In the processing of the data, the only add of this project is to remove the negative values from the spectra. In fact, these points are meaningless, having negative value would literally mean that energy is produced by scattering.

The dataset on which the model used for TL has been trained consists of 30 bacterial and yeast isolates. The reference training dataset consists of 2000 spectra each for the 30 reference isolates.

3 The Methodological Approach

This is the central and most important section of the report. Its objective must be to show, with linearity and clarity, the steps that have led to the definition of a decision model. The description of the working hypotheses, confirmed or denied, can be found in this section together with the description of the subsequent refining processes of the models. Comparisons between different models (e.g. heuristics vs. optimal models) in terms of quality of solutions, their explainability and execution times are welcome.

Do not attempt to describe all the code in the system, and do not include

Table 1: This table represents the repartition of patients and spectra belonging to the ALS and CTRL groups

Repartition of spectra and patients in dataset				
Groups	Patient ID	Samples range	Samples count	Number of samples per group
ALS	ALS01	1-60	60	393
	ALS02	61-78	18	
	ALS05	79-114	36	
	ALS07	115-150	36	
	ALS08	151 -194	44	
	ALS09	195-210	16	
	ALS10	211-225	15	
	ALS11	226-242	16*	
	ALS12	243-256	14	
	ALS13	257-281	25	
	ALS14	282-300	19	
	ALS15	301-314	14	
	ALS16	315-324	10	
	ALS17	325-334	10	
	ALS18	335-344	10	
	ALS20	345-354	10	
	ALS22	355-364	10	
	ALS23	365-374	10	
	ALS24	375-384	10	
	ALS25	385-394	10	
CTRL	CTRL01	1-33	33	198
	CTRL02	34-76	43	
	CTRL03	77-91	15	
	CTRL04	92-138	47	
	CTRL05	139-149	11	
	CTRL06	150-158	9	
	CTRL07	159-168	10	
	CTRL08	169-178	10	
	CTRL09	179-188	10	
	CTRL10	189-198	10	

* The 227th sample is missing. It has probably been removed while processing because it was ”too bad”.

large pieces of code in this section, use pseudo-code where necessary. Complete source code should be provided separately (in Appendixes, as separated material or as a link to an on-line repo). Instead pick out and describe just the pieces of code which, for example:

- are especially critical to the operation of the system;
- you feel might be of particular interest to the reader for some reason;
- illustrate a non-standard or innovative way of implementing an algorithm, data structure, etc..

You should also mention any unforeseen problems you encountered when implementing the system and how and to what extent you overcame them. Common problems are: difficulties involving existing software.

4 Results and Evaluation

The Results section is dedicated to presenting the actual results (i.e. measured and calculated quantities), not to discussing their meaning or interpretation. The results should be summarized using appropriate Tables and Figures (graphs or schematics). Every Figure and Table should have a legend that describes concisely what is contained or shown. Figure legends go below the figure, table legends above the table. Throughout the report, but especially in this section, pay attention to reporting numbers with an appropriate number of significant figures.

5 Discussion

The discussion section aims at interpreting the results in light of the project's objectives. The most important goal of this section is to interpret the results so that the reader is informed of the insight or answers that the results provide. This section should also present an evaluation of the particular approach taken by the group. For example: Based on the results, how could the experimental procedure be improved? What additional, future work may be warranted? What recommendations can be drawn?

6 Conclusions

Conclusions should summarize the central points made in the Discussion section, reinforcing for the reader the value and implications of the work. If the results were not definitive, specific future work that may be needed can be (briefly) described. The conclusions should never contain "surprises". Therefore, any conclusions should be based on observations and data already discussed. It is considered extremely bad form to introduce new data in the conclusions.

References

The references section should contain complete citations following standard form. The references should be numbered and listed in the order they were cited in the body of the report. In the text of the report, a particular reference can be cited by using a numerical number in brackets as [**?**] that corresponds to its number in the reference list. LaTeXprovides several styles to format the references