Supporting information for Raman Spectrum Matching with Contrastive Representation Learning

Bo Li, Mikkel N. Schmidt, Tommy S. Alstrøm

April 7, 2022

1 Dataset statistics

Fig. 1 shows the distribution of spectra per class in the Mineral dataset. The majority of the minerals (50%) have less than five spectra.

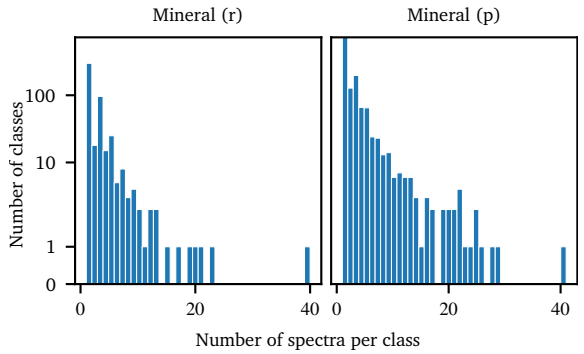

Fig. 1 The distribution of spectra per class in the Mineral dataset.

2 Influence of the augmentation methods

We also evaluate the matching accuracy on the Mineral dataset using different augmentation methods. Since it is common to generate the synthetic ones by taking the interpolation with randomly simulated coefficients if the number of spectra is larger than one[1–3], we here only experiment with applying different augmentation methods on the classes that only have one spectrum in the training data. We choose to use different augmentation methods as described below:

- None: no augmentation is applied
- Repeat: we duplicate the spectrum for each class to a certain number
- G-noise: we apply randomly generated Gaussian noise
- D-noise: we simulate noise based on the variation of the first derivative of the intensities as described in Section 2 and add those to the original spectrum

We experiment with different augmentation methods using the same data splits and we repeat this process five times. The results are shown in Table 1. The model is relatively more stable when we apply augmentations as described in section 2. Besides, we also achieved a slightly higher accuracy.

Table 1 Matching accuracy with 95% confidence interval for the Mineral dataset using different augmentations. Generating synthetic spectra with *D-noise* leads to a more stable and slightly higher spectrum matching accuracy.

	None	Repeat	G-noise	D-noise
Mineral (r)	93.58±0.68	93.94±0.54	93.64±0.32	**94.05±0.25**
Mineral (p)	91.50±0.25	91.58±0.38	91.42±0.55	**92.34±0.26**

3 Influence of the positive-negative ratio

One of the problems in training a Siamese network is how to balance the number of positive pairs and negative pairs per batch α such that the model is able to retrieve the similar ones and also discriminate the dissimilar ones. Therefore, we here examine the influence of the ratio α on the model performance for all the dataset on a fixed data splits. The results are shown in Table. 2.

Table 2 Influence of the ratio between the number of positive pairs and negative pairs per batch α over five different data splits (±95% confidence interval). The choice of the α has more visible influence on the Bacteria dataset compared to other datasets

	0.05	0.1	0.5	1	2
Mineral (r)	93.84±0.49	93.74±0.64	94.20±0.67	**94.66±0.55**	94.05±0.47
Mineral (p)	91.73±0.16	92.09±0.28	92.22±0.37	**92.24±0.32**	92.19±0.27
Organic (r)	94.44±0.76	96.11±1.19	**97.50±0.91**	97.22±0.77	96.66±0.97
Organic (p)	**97.22±0.77**	**97.22±0.77**	**97.22±0.77**	97.22±1.33	96.66±0.97
Bacteria	82.44±0.39	82.28±0.45	**83.06±0.19**	82.14±0.48	82.12±0.40

4 Influence of the distance calculation

To demonstrate the need of using both the element-wise product and the absolute difference between the feature maps as input to the neural network that computes the spectral similarity, we conduct a study where we drop one of the distance metrics. The results over multiple data splits are shown in the Table 4. Absolute difference alone performs better than the element-wise product on the Mineral (r) and Organic datasets, but worse than the element-wise product on the Bacteria dataset. Using both metrics together gives better and more stable performance on all the datasets.

5 Error analysis on S. lugdunensis

S. Lugdunensis is an example of a class in the Bacteria dataset that is difficult to classify correctly using spectral matching, because the spectrum in the test dataset is different than the ones in the reference and reference-finetune dataset (see Fig. 2.)

Table 3 Number of spectra that can be matched per second in the inference time

	Mineral (r)	Mineral (p)	Organic (r)	Organic (p)	Bacteria
Number of reference spectra	1190	3554	144	144	5400
Number of matched spectra per second	100	37	921	901	33

Table 4 Matching accuracy with 95% confidence interval over five data splits on all the datasets.

	$D_{\text{prod.}}$	$D_{\text{diff.}}$	$D_{\text{prod.}}$ and $D_{\text{diff.}}$
Mineral (r)	94.20±0.39	94.25±0.81	**94.51±0.15**
Mineral (p)	**92.25±0.49**	90.23±0.36	91.85±0.34
Organic (r)	92.78±3.20	96.94±2.35	**96.39±0.59**
Organic (p)	94.21±1.18	96.52±1.06	**96.76±1.38**
Bacteria	82.22±0.72	81.13± 0.34	**82.63±0.32**

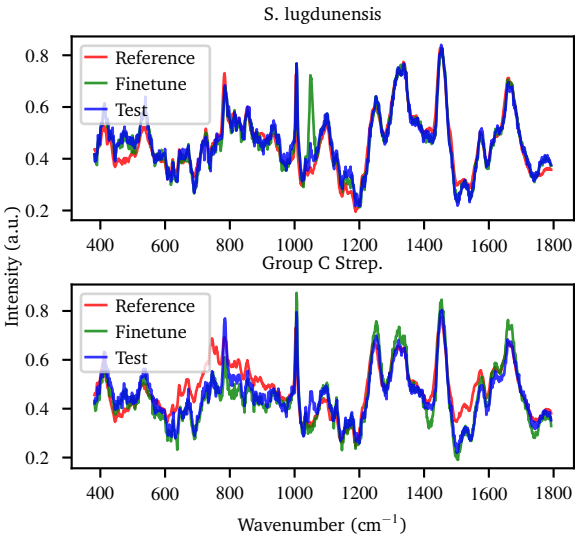

Fig. 2 Averaged spectrum for Bacteria *S. lugdunensis* and *Group C Strep.* in the reference, reference-finetune, and the test dataset.

6 Computation cost

We train and evaluate our model using an Intel(R) Xeon(R) CPU E5-2620 v4@2.10GHz with TITAN Xp 12GB machine. Depending on the size of the reference dataset and the use of the early stopping strategy, the training time per experiment is between 4 min (organic dataset) and 5 hours (bacteria dataset). During inference, the number of spectra that can be matched per second ranges from 33 to 920 depending on the size of the reference dataset and the input resolution of the spectra (see Table 3).

Notes and references

1 J. Liu, M. Osadchy, L. Ashton, M. Foster, C. J. Solomon and S. J. Gibson, *Analyst*, 2017, **142**, 4067–4074.

2 J. Liu, S. J. Gibson, J. Mills and M. Osadchy, *Chemometrics and Intelligent Laboratory Systems*, 2019, **184**, 175–181.

3 R. Zhang, H. Xie, S. Cai, Y. Hu, G.-k. Liu, W. Hong and Z.-q. Tian, *Journal of Raman Spectroscopy*, 2020, **51**, 176–186.