RE-EVALUATING THE SCATTERING TRANSFORM

Francisco Rodríguez-Algarra, Bob L. Sturm
Centre for Digital Music, Queen Mary University of London, U.K.
{f.rodriguezalgarra, b.sturm}@qmul.ac.uk

ABSTRACT

We reproduce published experiments with the scattering transform [1, 2] and consider the contents of the *GTZAN* benchmark music dataset in the results.

1. INTRODUCTION

Andén and Mallat [1] propose the scattering transform as a way to incorporate invariances into audio features, as opposed to learning them from data. In fact, Mallat [2] points out the resemblance of scattering transform features to the first layers of trained convolutional deep neural networks. The scattering transform uses a cascade of wavelet transforms to achieve invariances to global time-shifts and local deformations like time-warping. Moreover, the scattering transform allows for other kinds of invariances, e.g., frequency invariance when applied along the log-frequency scale. Table 2 in [1] shows that scattering transform-based SVM systems reproduce a large proportion of ground truth in the benchmark music dataset *GTZAN* [3,4]. The reasons behind these low error rates are not clear. Furthermore, the experiments do not consider the contents of *GTZAN*, in particular its known faults [3].

2. RE-EVALUATION METHODOLOGY

The experiments in [1] are reproducible, [1] and so we adapt their code in the following ways. [2] Due to memory constraints, we decrease by a factor of 4 the number of scattering features in the pre-computation of the Gaussian kernel of the SVM. This reduces computational cost without sacrificing much performance. While [1] uses 10-fold stratified cross-validation, we use two different hold-out train-test partitions:

(i) 75/25% randomised stratification
(ii) 640/290 fault-filtered selection.

We create the fault-filtered selection by hand.

3. RESULTS

Table 1 shows the normalised accuracies (mean recalls) of our re-evaluations along with those reported in [1] for six different feature sets (a - f). We see that the differences between the results in [1] and ours in condition (i) are small, and most of them within reason considering the standard deviations reported in [1]. Only our results with feature sets (e) and (f) differ more than two standard deviations from the accuracies reported in [1]. In condition (i), we see an increase of accuracy when we include second-order time-scattering features, (b) to (c). Adding depth to the features, however, does not increase the performance further, contrary to what is reported in [1]. Wee see accuracy decrease when we include third-order features, feature set (c) to (f).

We observe a considerable decrease in performance of the systems between conditions (i) and (ii). Feature set (e) achieves the highest normalised accuracy in condition (ii), which is almost 20 percentage points lower than the highest in condition (i). Similar to the results in [1], adding depth to the features in condition (ii) increases accuracy, except in feature set (f), which reproduces slightly less *GTZAN* ground-truth than (e).

When looking at the changes in performance within each of the *GTZAN* categories between conditions (i) and (ii), we see similarity in the variations across scattering-based feature sets. We observe different trajectories for each individual category, however, similar to what Fig. 2 shows for feature set (e). Figure 3 shows the figures of merit obtained with feature set (e) in condition (ii). In all feature sets, most of the categories suffer a decrease in the achieved accuracy, ranging from an average of 1.34 percentage points in the case of "Pop" up to 49.17 in the case of "Rock". The accuracies increase in two categories, however: "Metal" increases in average 2.2 percentage points, and "Classical" moves to perfect for every feature set (an average increase of 8.8 percentage points). This variability suggests that the interactions between the feature sets and the partitioning conditions are non-trivial.

Regarding the particular excerpt labelling errors that all scattering-based systems make, in condition (i) we find the following excerpts are consistently mislabelled: "Richard Strauss - Konzert Fur Waldhorn Mit Orchester, Op. 11, Allegro" (`cl48`), [3] "Leonard Bernstein - Candide Overture" (`cl52`), and "Janet Jackson - If" (`po69`), are classified as "Jazz"; "Queen - Tie Your Mother Down" (`me58`) is classified by all variants in the "Rock" category. Two excerpts are consistently mislabelled in both conditions: "Al-

[1] `http://www.di.ens.fr/data/software/`

[2] `https://code.soundsoftware.ac.uk/projects/scatter_reeval/`

© Francisco Rodríguez-Algarra, Bob L. Sturm. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Francisco Rodríguez-Algarra, Bob L. Sturm. "Re-Evaluating the Scattering Transform", Extended abstracts for the Late-Breaking Demo Session of the 16th International Society for Music Information Retrieval Conference, 2015.

[3] We refer to an excerpt using the notation "`XXNN`", where "`XX`" refers to the first two letters of the *GTZAN* category (e.g., `cl` for "Classical"), and "`NN`" is the ID of the audio filename within the category

Set	Features	Normalised Accuracy		
		Reported in [1]	Randomised Stratification (i)	Fault Filtering (ii)
a	Δ-MFCC (T=740 ms)	82.0 ± 4.2	78.00	53.29
b	Time Scat., l=1	80.9 ± 4.5	79.20	54.96
c	Time Scat., l=2	89.3 ± 3.1	88.00	66.46
d	Time & Freq. Scat., l=2	90.7 ± 2.4	87.20	68.49
e	Time & Freq. Scat., l=2, Adapt Q_1	91.4 ± 2.2	85.60	68.61
f	Time Scat., l=3	89.4 ± 2.5	83.60	68.32

Table 1. Normalised accuracies (mean recall) in *GTZAN* dataset with six feature sets used in [1], and our conditions.

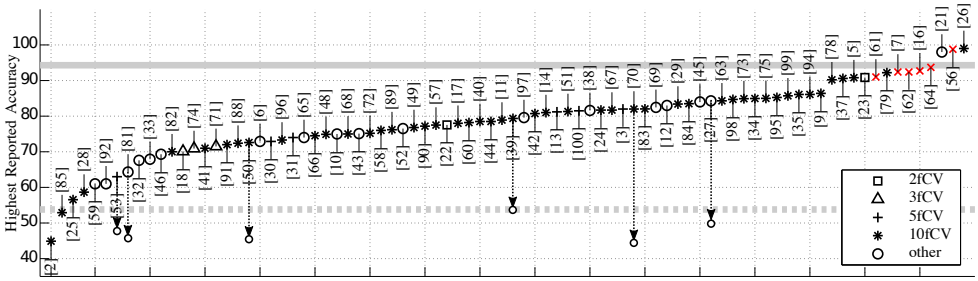

Figure 1. Highest published accuracies of music classification systems tested in *GTZAN* (numbers refer to publications in [3]). Results marked "x" are due to methodological errors in the evaluation implementation. A downward pointing arrow shows the difference between published accuracy and our re-evaluation in condition (ii).

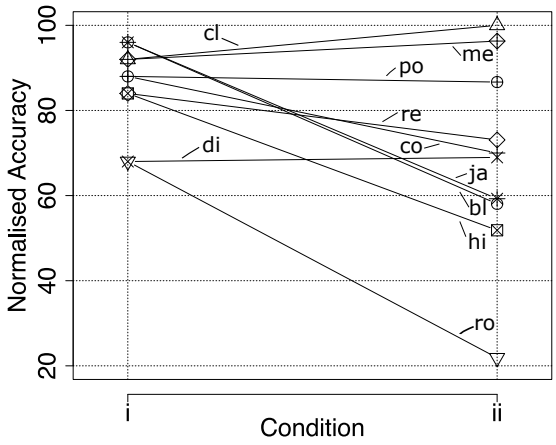

Figure 2. Interaction between the partitioning conditions (i) and (ii) and *GTZAN* categories for feature set (e). The trend is similar across scattering-based feature sets.

bert Collins - Iceman" (`bl98`) is classified as "Rock", and "Jimmy Cliff - Many Rivers to Cross" (`re83`) is classified as "Jazz". In condition (ii), we find 29 excerpt consistently mislabelled, including four by Clifton Chenier (always from "Blues" to "Rock"), and four by A Tribe Called Quest (always from "Hip-Hop" to "Pop").

As Table 1 and Figs. 1-3 show, the faults in *GTZAN* [3] considerably affect the amount of ground truth reproduced by music classification systems, and in ways unique to each system and each *GTZAN* category. Our current work is investigating why systems using these features are reproducing the most *GTZAN* ground truth of all systems we have re-evaluated in condition (ii) (see Fig. 1). In particular, we are performing a deeper review of system performance and how music content relates to it.

ACKNOWLEDGMENTS

Thanks to Joakim Andén and Stéphane Mallat for the reproducible research package accompanying their paper, and additionally to Vincent Lostanlen for advice.

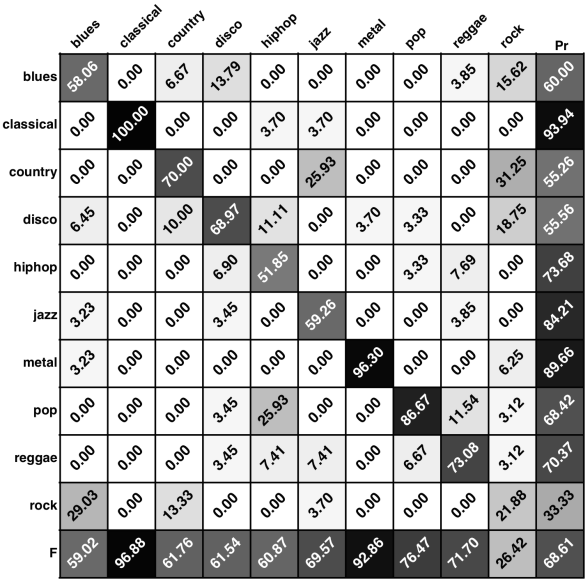

Figure 3. Figure of merit (x100) in condition (ii) using second order time and frequency scattering features, and adaptive wavelet octave bandwidth Q_1 (e). Column is ground truth, row is prediction. Far-right column is precision, diagonal is recall, bottom row is F-score, lower right-hand corner is normalised accuracy. Off-diagonals are confusions.

4. REFERENCES

[1] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Trans. Signal Process.*, vol. 62, pp. 4114–4128, Aug 2014.

[2] S. Mallat, "Group Invariant Scattering," *Communications on Pure and Applied Mathematics.*, vol. LXV, no. 11, pp. 1331–1398, 2012.

[3] B. L. Sturm, "The *GTZAN* dataset: Its contents, its faults, their effects on evaluation, and its future use," *http://arxiv.org/abs/1306.1461*, 2013.

[4] G. Tzanetakis and P. Cook, "Musical Genre Classification of Audio Signals," *IEEE. Trans. Speech and Audio Process.*, vol. 10, no. 5, pp. 293–302, July 2002.