WEAK MULTI-LABEL AUDIO-TAGGING WITH CLASS NOISE

Katharina Prinz and Arthur Flexer
Austrian Research Institute for Artificial Intelligence (OFAI), Vienna, Austria
{name.surname}@ofai.at

EXTENDED ABSTRACT

The necessity of annotated data for supervised learning often contrasts with the cost of obtaining reliable ground-truth in a manual fashion. Automated methods, on the other hand, simplify the annotation process and result in greater quantities of data with possibly noisy labels. Task 2 of the DCASE2019 Challenge, titled "Audio tagging with noisy labels and minimal supervision", tried to answer the question whether such data can be incorporated into an audio-tagging learning process in a meaningful manner [1].

This work builds upon a submission to the DCASE2019 Challenge [3], with a particular focus on audio samples with musical labels. More precisely, out of the 80 labels in the original setup, our experiments are restricted to samples with one or a multiple of 12 different classes, containing musical instruments and male/female singing. In total, 3967 audio clips remain of which roughly 20% (825) are curated, and 80% (3142) are noisy samples. Curated audio clips are denoted as such due to the manual data labelling process used for obtaining annotations [2]. On the other hand, labels of noisy audio clips are the result of automated heuristics [1]. As annotations are only available at clip-level without time stamps, we speak of weak labels.

In our experiments, we use a Convolutional Neural Network (CNN) with eight convolutional- and three average-pooling layers as proposed in [3]. Input features of the CNN are half-overlapping windows of perceptually weighted 128-bin Mel-spectrograms, which are normalised to zero mean and unit variance over each of the frequency bins according to the available data. The learning procedure uses an Adam optimizer with an initial learning rate of 0.001. After 90 epochs, the learning rate is reduced by a factor of 10, and training is continued for 30 more epochs. In order to prevent overfitting, mix-up data augmentation [5] with $\alpha = 0.3$, batch normalisation and drop-out are used.

As opposed to finding a way to maximise the performance of an audio-tagging system on weakly multi-labelled data by exploiting both curated and noisy data, we focus on the impact these two types of data have individually. In particular, we take a closer look at the contribution of noisy data in the task of predicting musical tags. To evaluate the performance of different training data, we use multi-label stratified 4-fold cross validation [4] on the curated dataset. In other words, validation sets always contain one fold of curated data only; training data either consists of the remaining three curated folds, or of noisy data.

As in the DCASE2019 Challenge, the Label-Weighted Label-Ranking Average Precision[1] (lwlrap) is used to measure how well a particular model fits the curated audio samples within a validation set. We compute 5 runs with different random CNN initializations, which yields 5 times 4-fold cross validations altogether. Table 1 shows mean and standard deviation of the validation lwlrap for three different scenarios over all 5 runs, after averaging the lwlrap of a model over four folds. To make these results more comparable, all four folds are computed once, and used in every setting.

RANDOM	CURATED	NOISY
0.268 ± 0.009	0.947 ± 0.003	0.654 ± 0.012

Table 1. Mean $\pm$ standard deviation over 5 runs of different models, averaged over 4 folds.

[1] http://dcase.community/challenge2019/task-audio-tagging#evaluation-metric

The first column in Table 1 shows the performance of our baseline. As we work with almost equally distributed labels in a multi-label environment, our baseline predicts random probabilities for each class, resulting in what can be seen as a random ranking of all 12 classes. The performance of models trained solely on curated or noisy data is compared to this baseline subsequently. In addition to comparing mean and standard deviation of different models over five runs of cross validation, paired sample t-tests are performed to determine the significance of the difference between these means.

Training our CNN with curated audio clips results in a high mean validation lwlrap of 0.947, as can be seen in the second column of Table 1. This is a statistically significant difference compared to the mean baseline performance of 0.268 ($|t| = |156.320| > t_{95,df=4} = 2.776$). Note here, that the validation set is one fold of curated data, with the remaining three curated folds being used for training.

Similarly, the difference between the model trained solely on noisy data (with a mean lwlrap of 0.654) and our baseline is statistically significant ($|t| = |48.004| > t_{95,df=4} = 2.776$). Instead of using the entire set of noisy audio clips to train the CNN, we take a random subset with the same cardinality as the curated training sets in the prior scenario. This noisy training subset is re-sampled each epoch, and for evaluation, we once more use curated validation folds. All other training settings are as for our curated CNN. Comparing the mean lwlrap of our CNNs trained on curated and noisy data respectively, we find that also this difference is statistically significant ($|t| = |46.952| > t_{95,df=4} = 2.776$).

In conclusion, we see that using the set of curated audio samples to train a model for audio-tagging of musical data leads to a considerably higher validation lwlrap than when working with noisy data only. On the other hand, a model trained on noisy audio samples still performs significantly better than our random baseline, but with room for improvement compared to the curated scenario. This is particularly interesting, as noisy and curated clips have the same set of labels, but originate from different sources [1]. Being based on Freesound (`www.freesound.org`) content, curated audio files often contain isolated sound samples of a class; on the opposite, noisy files tend to be of a more composite nature, as they consist of Flickr video soundtracks.

The unknown class noise ratio within the data complicates further conclusions. We could argue, that the performance in our third setting is a result of the model being able to extract some kind of useful information for audio-tagging from noisy data. Nevertheless, it could also be a consequence of a possibly high amount of correct labels within the 'noisy' dataset. Still, we saw that for given data, a model solely trained on noisy data can perform much better than a random predictor in the task of weak multi-label audio-tagging, even in the case of a domain mismatch.

ACKNOWLEDGMENTS

A special thanks to Fabian Paischer and everyone involved with our DCASE2019 Challenge submission. This work is supported by the Austrian National Science Foundation (FWF, P 31988).

REFERENCES

[1] Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel PW Ellis, and Xavier Serra. Audio tagging with noisy labels and minimal supervision. *arXiv preprint arXiv:1906.02975*, 2019.

[2] Eduardo Fonseca, Jordi Pons Puig, Xavier Favory, Frederic Font Corbera, Dmitry Bogdanov, Andres Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. Freesound datasets: a platform for the creation of open audio datasets. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, pages 486–93, Suzhou, China, 2017.

[3] Fabian Paischer, Katharina Prinz, and Gerhard Widmer. Audio tagging with convolutional neural networks trained with noisy data. June 2019. `http://dcase.community/documents/challenge2019/technical_reports/DCASE2019_Paischer_37_t2.pdf`.

[4] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. On the stratification of multi-label data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 145–158, Athens, Greece, 2011.

[5] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.