EFFICIENT TRAINING AND LABELING FOR INSTRUMENT RECOGNITION USING ACTIVE LEARNING

Ana Elisa Mendez Mendez
Music and Audio Research Lab
New York University, USA
anaelisamendez@nyu.edu

Yu Wang
Music and Audio Research Lab
New York University, USA
wangyu@nyu.edu

Juan Pablo Bello
Music and Audio Research Lab
New York University, USA
jpbello@nyu.edu

ABSTRACT

Building a classifier with a large unlabeled and unbalanced dataset can be challenging and requires a huge amount of human resources for annotation. In this study, we propose an active learning (AL) approach to the problem. We trained classifiers with AL on OpenMIC-2018 dataset for instrument recognition (specifically guitar). The best F measure is 0.94 and is achieved with only 88 labeled data points, while the baseline model trained with random sampling has an F measure of 0.73 with same amount of labeled data. The examples queried during AL training process have 47% positive examples, giving a more balanced labeled set, while random sampling returns 36% positive examples. The results indicate that AL can be a good tool for a much more efficient training and labeling process with the least possible human resources.

1. INTRODUCTION

Supervised learning is a machine learning method which works well on training a binary classifier. It is widely used for music information retrieval tasks such as instrument recognition, genre categorization, etc. However, supervised learning requires a large amount of labeled data for training and testing to produce a robust model. When working with a large dataset that is unlabeled, and even more challenging: unbalanced, labeling would take a considerable amount of time and annotators.

In this study we propose an efficient method for labeling a dataset and training a classifier at the same time by using active learning (AL). AL is a semi-supervised machine learning method that queries for the label of the most informative instances to increase classification performance. We apply the proposed AL training framework to OpenMIC-2018 dataset, which includes a large amount of unlabeled audio data for instrument recognition. We also apply the framework to SONYC (Sounds of New York City) dataset, a dataset with urban sounds, to show how AL works in different domains.

2. RELATED WORK

Previous works by [4] and [1] talk about the benefits of active learning in sound classification for reducing annotation resources in the cases where datasets are too large to be labeled by humans. Zhao et al. [4] proposed a novel active learning method to save annotation efforts when preparing training data. Han et al. [1] used active learning and self-training with the same purpose. The combination of both approaches greatly reduces human efforts in data annotation.

3. METHODS

3.1 Dataset

The OpenMIC-2018 dataset [3] is a new open access dataset for multi-instrument recognition. The dataset, which contains 20,000 examples in the form of 10-second excerpts, has been partially labeled for the presence or absence of 20 instrument classes. In this study, we focus on labeling and training a classifier for one instrument: guitar, which is relatively easy to annotate for lower annotation error. There are 1650 labeled examples, 1137 positives and 513 negatives. The rest of the examples would serve as the unlabeled data pool. The input representation for the classifier is the mean and standard deviation of VGGish [2] features over the 10-second long recordings.

One-third of the labeled data is held out as the test set, and the rest is split for 5-fold cross validation. From the training set, we randomly draw two data points, one positive and one negative, as the initial training data. We keep the initial training set small to see a clearer trend of how efficient AL can be.

3.2 Training

Figure 1 shows the framework of our proposed training process. A binary random forest classifier is first trained with the initial training data (number of trees = 100, maximum depth of the tree = 8). Then we run 100 AL iterations. In each iteration, the active learner searches in the unlabeled data pool and returns the query which is most informative for learning. We use least confident uncertainty sampling querying strategy. Then a human annotator listens to the queried audio example, labels it and adds it back to the training set. The model then is retrained with the updated training data and pool.

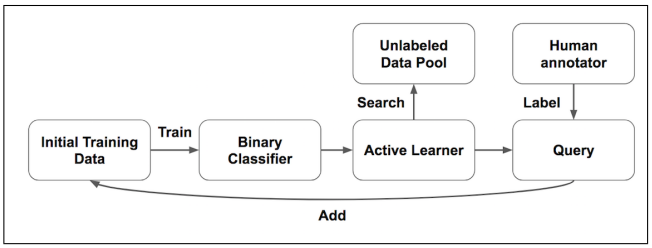

Figure 1. Proposed AL framework.

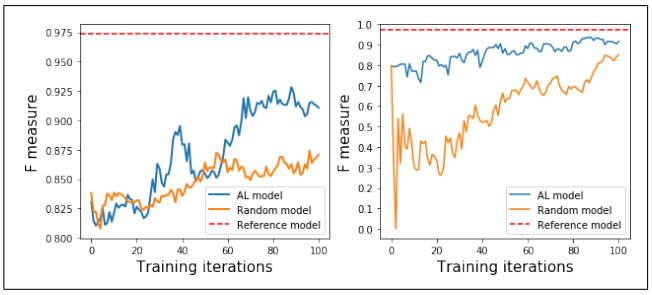

Figure 2. Left: Learning curve during training; Right: Model performance on test set, for AL models, random models and reference model.

For comparison, we trained baseline models with random sampling. Starting with the same framework as AL: two initial training data points and 100 training iterations. At each iteration, the query is randomly sampled from the pool instead of uncertainty-based sampling. A reference model is also trained with the entire labeled training set to show the best possible performance with all the labeled data available.

4. EXPERIMENTS

Figure 2 shows the resulted learning curve and performance measured on test set for the guitar classifier. AL model performance during training outperforms random model after 60 training iterations. The test performance difference shows that AL model generalizes better compared to random model. The best performance of AL models is 0.94 at training iteration 86. This best model is only trained with 88 labeled training data points in total, and reaches a performance close to reference model (0.97), which is trained with 955 labeled data points. The performance of the random model at same iteration is 0.73.

Same experiment is also applied to SONYC to classify a specific inference noise. The resulted performance in Figure 3 shows the same trend, and the AL model even outperforms the reference model.

Table 1 reflects the percentage of positive examples queried by each sampling method during training. For guitar on the OpenMIC-2018 dataset AL returns 47% of positive examples, while random sampling only provides 36%. On the other hand, for the SONYC dataset, AL returns 47% of positive examples compared to 8% returned by random sampling.

	OpenMIC-2018	SONYC
Active Learning	47%	47%
Random Sampling	36%	8%

Table 1. Percentage of positive queries returned from AL and random sampling.

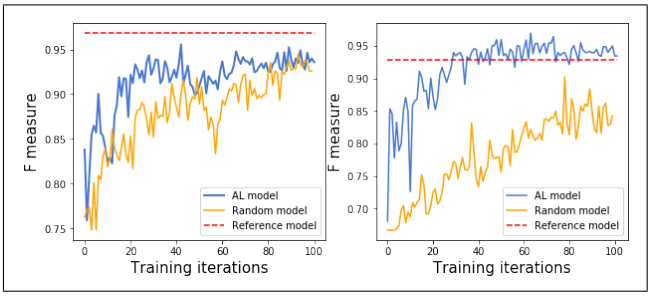

Figure 3. Left: Learning curve during training; Right: Model performance on test set for AL models, random models and reference model on SONYC data.

5. DISCUSSION AND FUTURE WORK

The experiment results show that AL provides more efficient labeling and training process when building a classifier on large unlabeled datasets. We also showed that AL can be applied to different domains, music and urban sounds.

For future work, the proposed AL framework will be applied to more instruments in the OpenMIC-2018 dataset. We also plan to try multi-class and multi-label training with AL, as well as modify the querying strategy for better performance.

6. REFERENCES

[1] Wenjing Han, Eduardo Coutinho, Huabin Ruan, Haifeng Li, Bjrn Schuller, Xiaojie Yu, and Xuan Zhu. Semi-supervised active learning for sound classification in hybrid learning environments. *PLoS ONE*, 11(9):1 – 23, 2016.

[2] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017.

[3] E.J. Humphrey, S. Durand, and B. McFee. Openmic-2018: An open dataset for multiple instrument recognition. In *19th International Society for Music Information Retrieval Conference*, ISMIR, 2018.

[4] Z. Shuyang, T. Heittola, and T. Virtanen. Active learning for sound event classification by clustering unlabeled data. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 751–755, March 2017.