Project Proposal: Audio Tagging System

Alexandra Sudomoeva (as5402) and Steve McClain (sdm2171)

1 Description

Goal: Develop an automatic, general-purpose audio tagging system capable of accurately classifying sound collections for a wide range of real-world environments.
Data: The original dataset is taken from Kaggle [2]. The samples (20,000 WAV files) are generated from Freesound's library and include things like musical instruments, domestic sounds, and animals [3]. Each input represents a WAV file with a corresponding annotative label. There are 41 labels overall, each generated from Google's AudioSet ontology. The dataset also includes a boolean column indicating whether the label was manually verified.

2 Proposal

To achieve the goal, we will be cycling through Box's loop [1]. Due to the complexity of the task, we propose two separate stages to address both the model performance (given a fixed number of labels) as well as generalizing to the complexity of real-world data (e.g. classifying sounds that were not in the training set).
Stage 1: This stage will focus on tuning the model for the highest possible performance given a fixed number of labels. The test will be performed on a subset of the data with only training labels in place.

Modeling	Inference	Criticism
Our approach is to use a modified Latent Dirichlet Allocation (LDA) model, with the following equivalences: - The "topics" will be sounds (e.g. hi-hat, laughter) - The "documents" will be WAV files - The "words" will be WAV chunks/frames As each WAV chunk will contain multiple sounds, we will sample from the sound distribution more than once.	For the inference stage, we will use Coordinate Ascent Variational Inference to estimate the posterior distribution. Alternative inference approaches, such as Gibbs sampling, may also be considered.	The performance will be evaluated by computing the mean average precision on the test set. We will consider using a semi- supervised approach to utilize the labels.

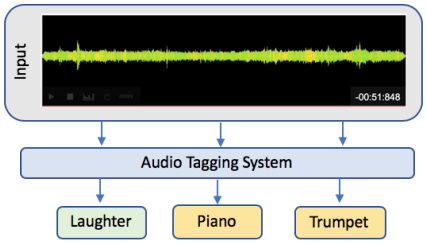

Figure 1: Overview of the multi-tagging system

Having achieved a high performing model during stage 1, it would still not be representative of the real world (expected poor performance on sounds outside of the original labeling).

Stage 2: This stage will focus on using Google's AudioSet ontology tree to improve the model performance on new sounds (i.e. sounds whose labels were not part of original learning)

Modeling	Inference	Criticism
We aim to extend the work done in Stage 1 by connecting each training label to its corresponding position in the tree. Therefore, allowing the model to tackle and classify new labels by assuming parental belonging. These labels can be used to "guide" our choice of prior on the per-wave-file sound distributions.	Same as above.	The performance will be evaluated by computing the mean average test set precision for proper "parent" classification within the anthology tree for never seen test labels.

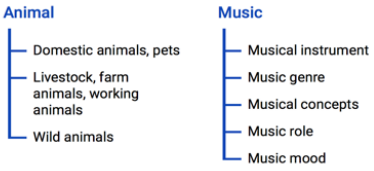

Figure 2: Snapshot of the AudioSet ontology tree (research.google.com)

References

[1] G. E. Box. *Science and statistics.* Journal of the American Statistical Association, 1976.

[2] Frederic Font Daniel P. W. Ellis Xavier Favory Jordi Pons Xavier Serra Eduardo Fonseca, Manoj Plakal. *General-purpose Tagging of Freesound Audio with AudioSet Labels: Task Description, Dataset, and Baseline.* Submitted to DCASE2018 Workshop, 2018.

[3] Xavier Favory Frederic Font Dmitry Bogdanov Andrés Ferraro Sergio Oramas Alastair Porter Eduardo Fonseca, Jordi Pons and Xavier Serra. *In Proceedings of the 18th International Society for Music Information Retrieval Conference.* Suzhou, China, 2017.