Representation Learning for the Automatic Indexing of Sound Effects Libraries

Alison B. Ma, Alexander Lerch

ama67@gatech.edu, alexander.lerch@gatech.edu

Georgia Tech Center for Music Technology
College of Design

Introduction

Problem Statement

Too much time is spent (re-)labeling and performing quality assurance on databases that continually grow in size and undergo taxonomy changes.

Non-uniform metadata makes successful sound search and taxonomy creation difficult for both users & vendors.

Potential Solutions

Deep learning to automate sound classification.

Universal Category System (UCS)
New industry-proposed solution to standardize taxonomies, designed by & for sound designers/editors.

Challenges with Potential Solutions

Training models on every sound library that exists...

is time-consuming.

might be unsuccessful: many libraries are too small, have messy labels, and have imbalanced class distributions.

Preliminary results reveal that UCS is used in a non-uniform way.

Method to address all aforementioned challenges

Representation Learning: Train a model to act as an intelligent feature extractor, which learns an embedding space that can generalize to any taxonomy of sound. Supported by existing literature, e.g. VGGish, OpenL3.

Overarching Research Question

What can we do to improve **learning generalized representations** for categorizing sounds?

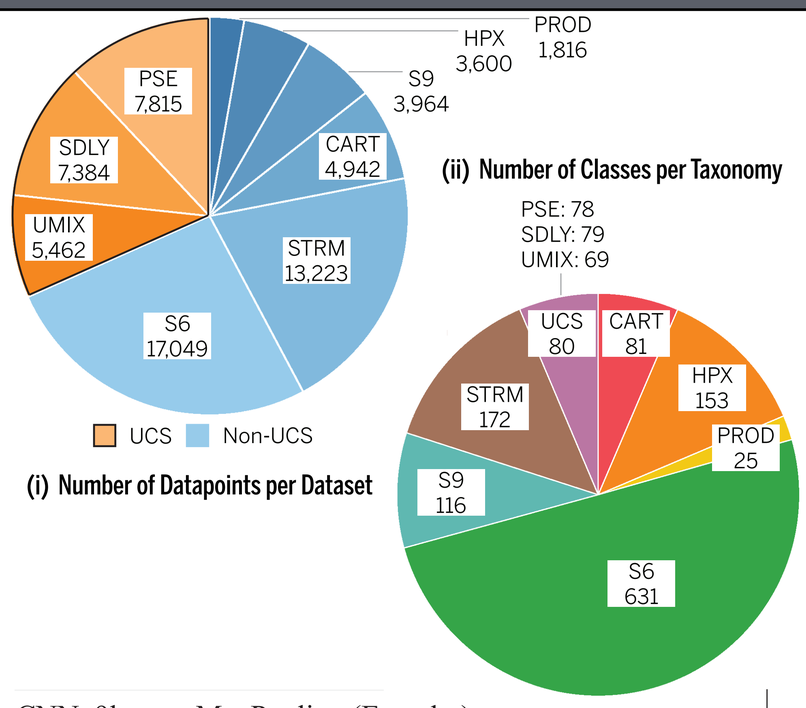
Specific Research Questions

Apply metric learning loss functions instead of standard cross-entropy?

Use cross-dataset training?

Use different cross-dataset training methods?

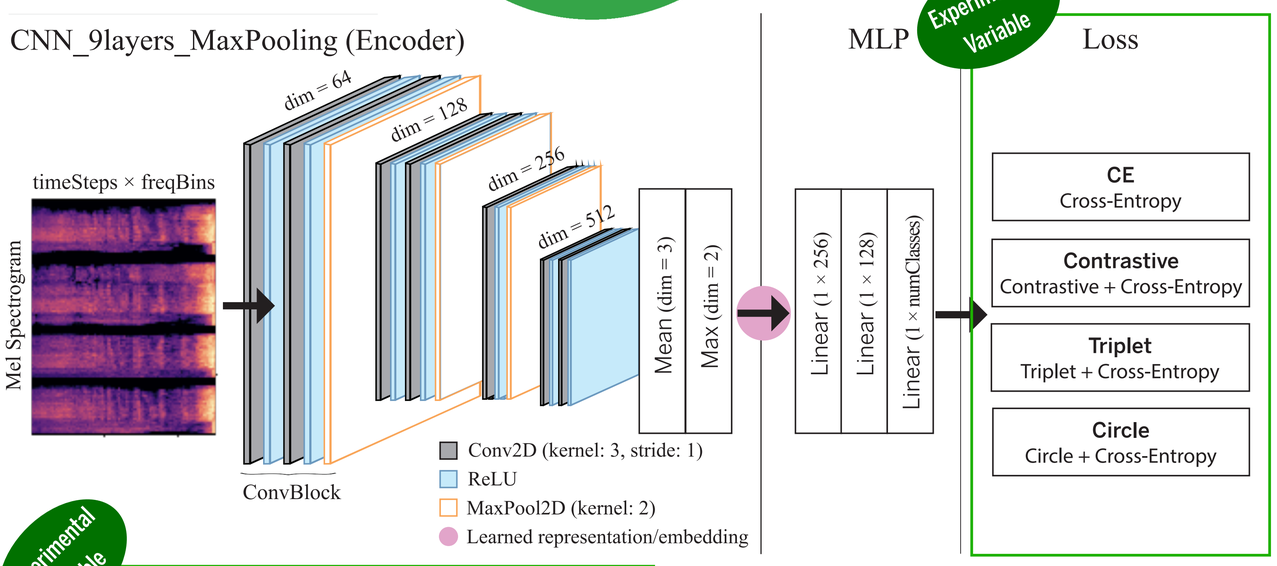
Experimental Setup (Data Statistics, Model, Variables)



Experimental Variable

Training Scenarios

- **Within-Dataset**: Pre-train/ evaluate encoder on same dataset.
- **UCS-Transfer**: Pre-train encoder on a UCS-compliant dataset, evaluate on other datasets in a transfer learning scenario.
- **Cross-Dataset**: Pre-train encoder on all datasets & taxonomies, evaluate encoder on any dataset.



Experimental Variable

Cross-Dataset Training Methods

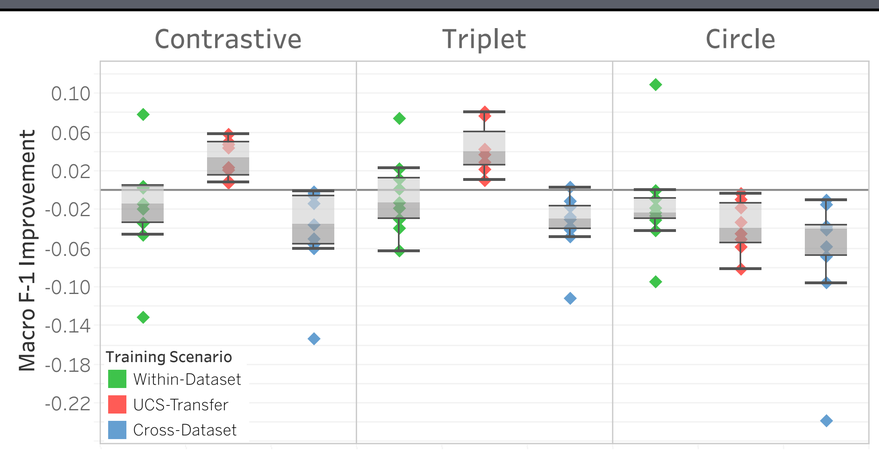
- **Data Mixing (Sequential, Joint)**: The order of datasets for pre-training.
- **'Focal' Dataset Regularization (FDR)**: Re-weighting datasets by "difficulty".
- **Dataset-Independent BatchNorm Layers (BN)**: BatchNorm layers that correspond to each dataset.

Evaluating Embeddings

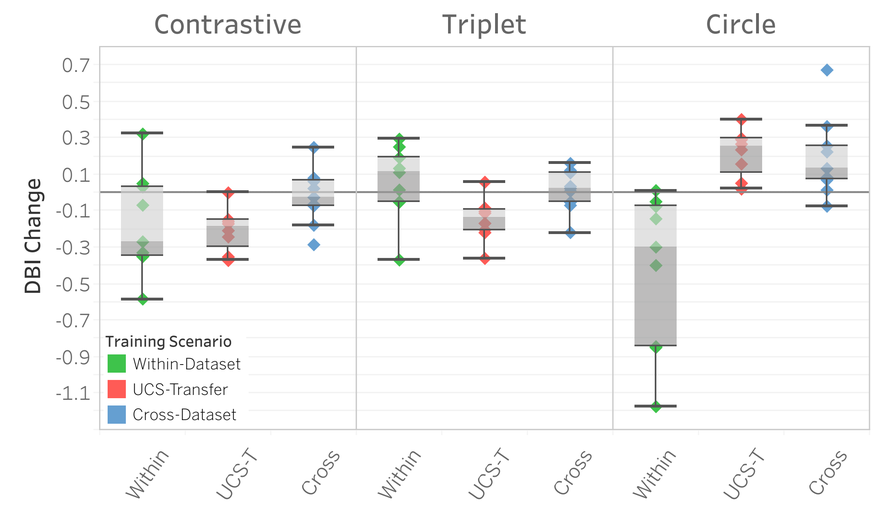
- All plots show results at the 2-second **frame-level**. Though not shown, **global** results may be obtained by performing a majority vote on all frame-level predictions. Our best **global** results are 6-7% higher.
- Classification macro F-1 scores used a **Nearest Neighbors** algorithm.

Representation Learning Experiments

Metric Learning

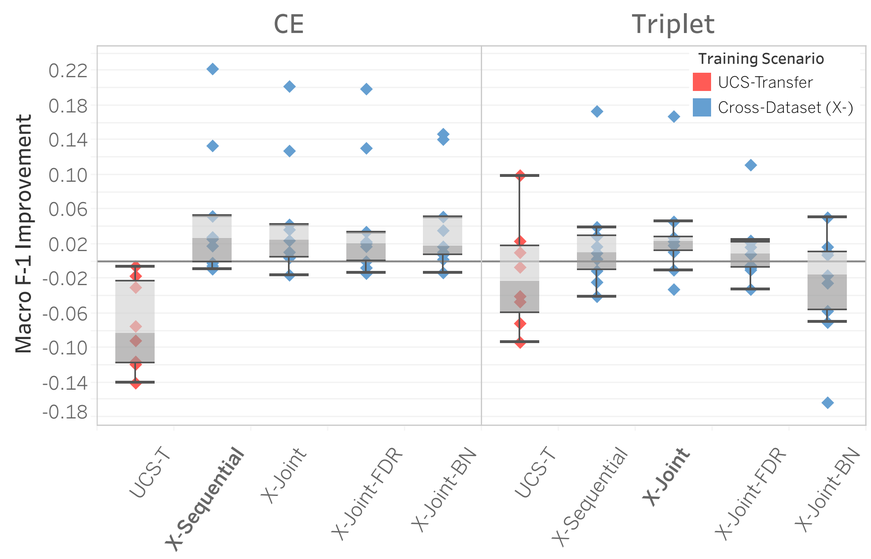


- Macro F-1 score improvement of **metric learning models relative to CE** in different training scenarios.



- Davies-Bouldin Index (DBI) change of **metric learning models relative to CE** in different training scenarios.

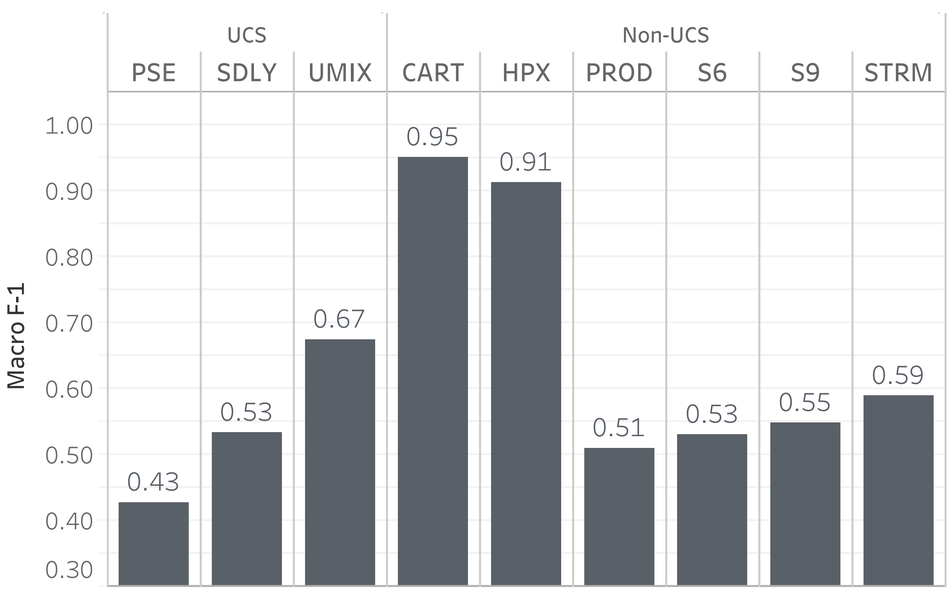
Cross-Dataset Methods



- Macro F-1 score improvement of various training scenarios from **Within-Dataset**.

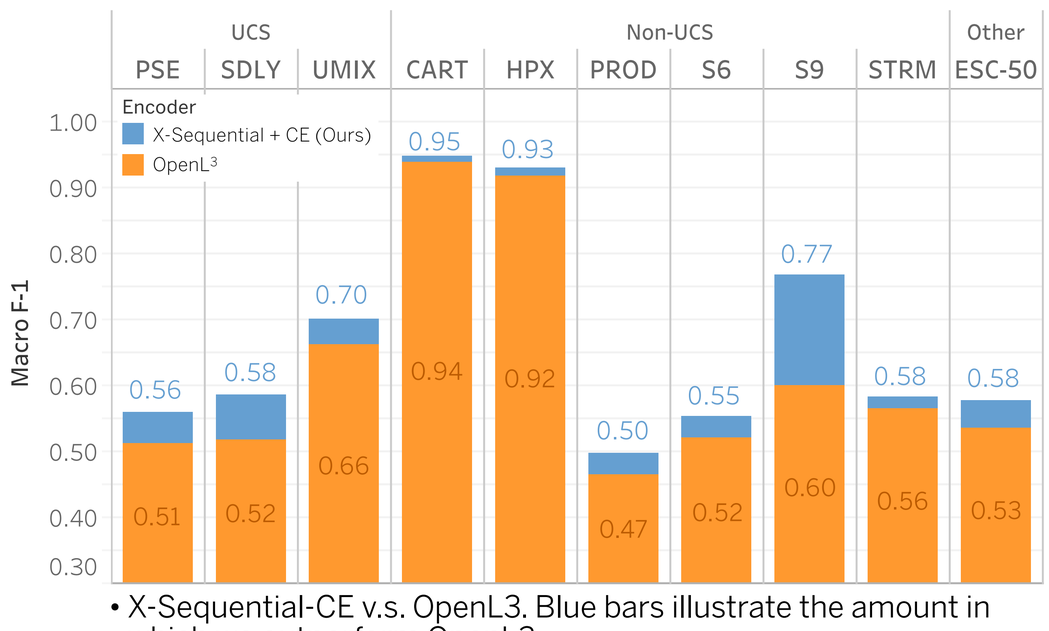
Conclusion

Baseline Results



- Macro F-1 score classification results using Cross-Entropy loss, CE (**Within-Dataset**)

Best Model v.s. SoTA



- X-Sequential-CE v.s. OpenL3. Blue bars illustrate the amount in which we outperform OpenL3.

Links!

- use our best model.
- view examples of classes.
- view t-SNE plots for UCS-compliant datasets generated with our best model.