# CS 4644/7643: Deep Learning
# Project Proposal

Instructor: Danfei Xu
TAs: Mihir Bafna (Head TA), Krishanu Agarwal, Manav Agrawal,
Anshul Ahluwalia, Aditya Akula, Matthew Bronars,
William Held, Vikranth Keerthipati, Renzhi Wu, Wei Zhou
Discussions: https://piazza.com/gatech/fall2023/cs46447643

Release: September 26, 2023
Due Date: September 26, 2023

- Write 2-3 paragraphs per section for full credit. The document should not exceed 800 words.
- There is no late deadline for this assignment. Submissions submitted after the due date will receive a 0.
- Submit this LaTeX document as a PDF to Gradescope.
- Submission: Please submit your proposal as a PDF on Gradescope. Only one person on your team should submit. Please have this person add the rest of your team as collaborators as a "Group Submission".

**Project Title**: Multi-Modal Approaches to Music Genre Classification
**Team Members**: Anthony Wong, Austin Chemelli

## 1 Project Summary

Music genre classification has, for a long time, been a popular task in music information retrieval (MIR). Many approaches have been explored, using various modes of information about the songs. These include analyzing raw audio clips, song lyrics, and extracted audio features. Predicting a genre of music based on only one of these modes, however, has proven to be a difficult task. Music genres themselves are not deterministic and are defined by the musical elements, lyrical content, and thematic concepts. Thus, predicting a genre solely based on a single mode of information seems insufficient, and this is where multi-modal learning comes in.
We are seeking to combine the spatial information given by a song's audio with the textual context provided by its lyrics to more accurately predict a song's genre. This process has been explored in other research papers, with limited amounts of success, barely surpassing a 50% accuracy level on the upper end). [1] Our goal is to make improvements on existing approaches and to explore various approaches in search of a well-performing multi-modal network.

## 2 Proposed Method

To implement this, we will train two "separate" models that deal with lyrics and audio data individually. For lyrics interpretation, we plan on doing some sort of reduction such that we keep only principal words, using an algorithm like ELMo (Embeddings from Language Models)) to vectorize the remaining lyrics, and using this to develop a deep neural network consisting of several LSTM (Long Short-Term Memory) layers. For audio data, we plan on taking .wav files and converting them to spectrograms to use in the development of a convolutional neural network. We can combine these two branches of networks using Multi-Modal Fusion and concatenating their outputs into a dense layer before classification. Our tentative loss function is categorical cross-entropy, though this may change depending on our findings. We will evaluate our model simply based on how well it classifies lyric-audio pairings into genres.
We believe this can work due to the features of its components. For lyrics, ELMo has characteristics that can capture word order and contextual detail. LSTMs similarly can recognize and "remember" contextual information, making it well-suited for lyric interpretation. For the audio, spectrograms have emerged as great ways to succinctly represent long-form audio. It pairs well with CNNs, as spectrograms have similarities to images, something CNNs are particularly suited for. We think we can combine these models in some way. Concatenating them seems like a simple way to do this, such that we can properly combine both networks. It is vaguely similar to the methods taken in the paper: *Hybrid Attention based Multimodal Network for Spoken Language Classification* [2]. However, there may be better ways this can be done, as we will find out. Regardless, we believe that this system can work.

## 3 Related Work

1. Improving Music Genre Classification from Multi-Modal Properties of Music and Genre Correlations Perspective [1]
   This research paper is a very recent (2023) approach to multi-modal learning using audio and lyric information. They utilize the same dataset (Music4All) that we will be using, and provide specific loss and accuracy calculations for this specific problem. This paper will be an important benchmark to compare our results to and to reference implementation methods used in there.

2. Music4All: A New Music Database and Its Applications [3]
   This research paper introduces the Music4All dataset that we will be using, describing the attributes it includes and how the data was collected. This will be useful in understanding the origin of the dataset while evaluating results. Additionally, the paper includes a basic genre classification implementation using Mel-spectrograms whose results are referenced in [1].

3. Multi-modal, Multi-task and Multi-label for Music Genre Classification and Emotion Regression [4]
   This research paper focuses on multi-level music genre classification, which will be important to consider and implement when we are making predictions. The results of their multi-modal implementation are also referenced in [1]

4. Music genre classification with the million song dataset [5]
   This research paper is one of the earlier implementations of multi-modal learning combining audio and lyrics, but on a different dataset (Million songs dataset). However, the paper goes

into detail on how the two modes are represented and implemented into the model, which will be valuable to reference and to see the improvements made over a long time period (12 years).

5. Multi-modal Network for Spoken Language classification [2]
   This paper aids us in building our knowledge of multi-modal fusion. Similarly to us, their network takes in both audio and textual data separately and has a specific layer for model fusion. This could contain useful information to help our model be as accurate as possible.

## 4 Datasets / Environments

The dataset we will be using is called Music4All, which " contains 15,602 anonymous users, their listening histories, and 109,269 songs represented by their audio clips, lyrics, and 16 other metadata/attributes". [3] We are confident that this dataset is large enough, and could potentially need to be sampled from since it contains a lot of songs and by extension, audio samples. Due to the potential volume of the data, we will likely train/run our model using a desktop with a dedicated GPU.

## References


[1] G. Ru, X. Zhang, J. Wang, N. Cheng, and J. Xiao, "Improving music genre classification from multi-modal properties of music and genre correlations perspective," 2023.

[2] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Hybrid attention based multimodal network for spoken language classification," Aug 2018.

[3] I. A. Pegoraro Santana, F. Pinhelli, J. Donini, L. Catharin, R. B. Mangolin, Y. M. e. G. da Costa, V. Delisandra Feltrim, and M. A. Domingues, "Music4all: A new music database and its applications," in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pp. 399–404, 2020.

[4] Y. R. Pandeya, J. You, B. Bhattarai, and J. Lee, "Multi-modal, multi-task and multi-label for music genre classification and emotion regression," in *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 1042–1045, 2021.

[5] D. Liang, H. Gu, and B. OâConnor, "Music genre classification with the million song dataset," *Machine Learning Department, CMU*, 2011.