Project Proposal

Automated Multi-Label Music Genre Classification + AI Music Generation

Group Members: Zhuoyuan Jiang, Zhehao Xing, Changjun Li, Shuodong Wang

1. What is the problem you are investigating? Why is it interesting?

This project addresses two interconnected tasks in music information retrieval: multi-label genre classification and personalized music generation. Music often blends multiple genres (e.g., Pop-Rock, Jazz-Fusion, Pop, Blue, and etc), making genre classification inherently a multi-label problem. Simultaneously, the ability to generate music tailored to specific genre combinations represents a significant advancement in creative AI applications.

Specifically, our goal is to first develop a deep learning model capable of accurately assigning multiple genre labels to songs (For example, a music can be labeled “rock, pop” instead of simple “rock” or “pop”). Then, using these generated genre labels, we aim to conditionally generate original, personalized music compositions reflecting user-specified combinations of genres. We introduce a **novel classifier-in-the-loop generation approach**, where generated music is re-classified by our model, and discrepancies between intended and predicted labels are used to iteratively improve generation quality. This feedback loop ensures both classification accuracy and authentic music generation.

This approach is especially interesting because it mirrors real-world scenarios where music genres aren't clearly separated, while offering exciting practical applications. For instance, music platforms like Spotify and Apple Music could enhance their services by not only improving genre classification accuracy (even small improvements in classification can significantly impact recommendation quality and generate a significant amount of revenue) but also **introducing AI-generated personalized playlists feature** based on users' unique taste patterns. Imagine telling the system "I like this song" and receiving not just similar existing songs, but new, AI-generated music that captures both the genre elements and lyrical themes you enjoy. The system could analyze both genre tags and lyrical patterns from favorite songs to generate music that resonates with personal preferences, effectively letting users enjoy their preferred music styles for longer periods.

The project's impact extends beyond just improving recommendation systems. By combining accurate multi-label classification with personalized generation, we're creating a versatile tool that could serve various applications - from generating background music to creating custom soundtracks. Our approach not only empowers creative AI to better capture and replicate the nuanced complexity of musical styles but also opens new possibilities for how users interact with and consume music. The ultimate vision is simple yet powerful: "Tell us the music you like, and we'll create a tailored musical experience you can enjoy indefinitely."

2. What dataset will you be using?

For this project, we will primarily utilize two comprehensive multi-label music datasets: the **FMA (Free Music Archive) dataset** and the **MuMu Dataset.**

The FMA dataset is particularly valuable for our multi-label classification task as it provides rich genre annotations where each track is associated with multiple genre labels in the form of numerical codes (e.g., genres_all = [322,5]), with each number corresponding to a specific genre. This structured labeling system allows for precise genre relationship modeling and hierarchical classification. A glimpse of FMA dataset is below:

Out[11]:

	track					album			artist	
	title	genres_all	genre_top	duration	listens	title	listens	tags	name	location
track_id										
150073	Welcome to Asia	[2, 79]	International	81	683	Reprise	4091	[world music, dubtronica, fusion, dub]	DubRaJah	Russia
140943	Sleepless Nights	[322, 5]	Classical	246	1777	Creative Commons Vol. 7	28900	[classical, alternate, soundtrack, piano, cont...	Dexter Britain	United Kingdom
64604	i dont want to die alone	[32, 38, 456]	Experimental	138	830	Summer Gut String	7408	[improvised, minimalist, noise, field recordin...	Buildings and Mountains	Oneonta, NY
23500	A Life In A Day	[236, 286, 15]	Electronic	264	1149	A Life in a Day	6691	[idm, candlestick, romanian, candle, candlesti...	Candlestickmaker	Romania
131150	Yeti-Bo-Betty	[25, 12, 85]	Rock	124	183	No Life After Crypts	3594	[richmond, fredericksburg, trash rock, virgini...	The Crypts!	Fredericksburg, VA 22401

The MuMu Dataset complements our analysis by providing explicit text-based multi-genre labels for each track. For example, a track with ID TRJIKJU128F930BF28 might be labeled with multiple genres such as "Vocal Jazz, Jazz, Traditional Vocal Pop, Pop, Modern Postbebop, Broadway & Vocalists". One of the key advantages of MuMu is its modular structure, where multiple subset datasets can be linked through track IDs, allowing us to query and merge various musical features and metadata as needed. This natural language representation of genre combinations, coupled with the dataset's flexible architecture, makes it particularly suitable for both our multi-label classification and music generation tasks. A glimpse of MuMu Dataset is below:

	amazon_id	album_mb	MSD_track	recording_	artist_mbi	genres
2	B00005YQ	77944b8c-	TRJIKJU12	68c38213-	0a6f37da-	Vocal Jazz,Jazz,Traditional Vocal Pop,Pop,Modern Postbebop,Broadway & Vocalists,Vocal Pop,Bebop
3	B00005YQ	77944b8c-	TRLCVKT1	f5c25488-	0a6f37da-	Vocal Jazz,Jazz,Traditional Vocal Pop,Pop,Modern Postbebop,Broadway & Vocalists,Vocal Pop,Bebop
4	B00005YQ	77944b8c-	TRBQSIG1	e76bbfcc-	0a6f37da-	Vocal Jazz,Jazz,Traditional Vocal Pop,Pop,Modern Postbebop,Broadway & Vocalists,Vocal Pop,Bebop
5	B00005YQ	77944b8c-	TRGQLER1	9def0715-	0a6f37da-	Vocal Jazz,Jazz,Traditional Vocal Pop,Pop,Modern Postbebop,Broadway & Vocalists,Vocal Pop,Bebop
6	B00005YQ	77944b8c-	TRQPGRM	9fe52229-	0a6f37da-	Vocal Jazz,Jazz,Traditional Vocal Pop,Pop,Modern Postbebop,Broadway & Vocalists,Vocal Pop,Bebop
7	B000G04U	4520281f-	TRWSCXI1	b4518a84-	d352f5dd-	Oldies,Pop,Rock
8	B000G04U	4520281f-	TRDPEVK1	d7d40efd-	d352f5dd-	Oldies,Pop,Rock

If we have time, we will try to enhance the dataset further, data augmentation techniques such as pitch shifting, time stretching, and noise injection will be applied. This will help the model generalize better across different styles and recording conditions. Additionally, spectral transformations (e.g., mel-spectrograms and MFCCs) will be extracted to serve as input features for the deep learning models.

3. What Deep Learning approach will you use or develop?

Our approach leverages state-of-the-art transformer architectures for both classification and generation tasks, moving away from traditional CNN/RNN hybrid models or GANs. The system consists of two main components:

Classification Model: We will implement the Audio Spectrogram Transformer (AST) architecture, which has demonstrated superior performance in audio classification tasks. We might also try BERT-based transformers. Depending on which model we use in the end, the model will be initialized with pre-trained weights from wav2vec2-base-960h available on HuggingFace, allowing us to benefit from transfer learning. Instead of using traditional single-label classification, we'll implement true multi-label classification by treating each genre as an independent binary classification task, enabling the model to capture the nuanced reality of songs belonging to multiple genres simultaneously. This approach is particularly suitable for our dataset where songs often span multiple genres (e.g., "Vocal Jazz, Traditional Vocal Pop, Broadway & Vocalists").

For audio representation and tokenization, we'll utilize Music2vec-V1, a specialized music audio representation model that can effectively capture musical features and structures.

Generation Model: For the generation component, we'll employ MusicGen, a transformer-based architecture specifically designed for music generation conditioned on textual descriptions. Our novel approach involves using the classification model's output probabilities as structured prompts for MusicGen. For example, if our classifier predicts "Pop Rock (0.8), Indie Rock (0.3), Alternative Rock (0.6)", we'll develop prompt engineering techniques to translate these probability distributions into effective conditioning signals for music generation.
The integration between classification and generation creates a feedback loop where:

- The classification model provides detailed genre probability distributions (and also binary decisions using sigmoid with threshold - one layer for binary decision oand one layer for probability distribution)
- These probabilities are formatted into structured prompts (e.g., "Create a song that is strongly Pop Rock (0.8), with subtle Indie Rock influences (0.3) and moderate Alternative Rock elements (0.6)" followed by original text prompt like “I want a song that tells the story of two strangers meeting in a rainy city, their eyes meeting across a crowded coffee shop, leading to an unexpected connection”)
- MusicGen generates music based on these nuanced prompts
- The generated music can be re-classified to verify genre adherence and adjust the generation process if needed

This approach allows for fine-grained control over the generated music's stylistic elements while maintaining musical coherence. The transformer-based architecture's ability to handle long-range dependencies and complex patterns makes it particularly suitable for capturing the subtle interplay between different genre characteristics in both classification and generation tasks.

4. What are the anticipated challenges of this project?

1) Data Challenges

- Limited Multi-label Resources: High-quality datasets with consistent multi-genre annotations are scarce. Also, different datasets have different label for genres, which leads to a labeling/annotation inconsistency across datasets.
- Genre Imbalance: Popular genres significantly outnumber niche genres, requiring targeted sampling and augmentation strategies

2) Technical Challenges

- Computational Demands: Transformer architectures (AST, MusicGen) require substantial computing resources
- Dual-Output Classification: Balancing binary decisions and probability distributions might require custom loss functions

3) Integration Challenges

- Feedback Loop Efficiency: Classifier-in-the-loop approach adds complexity and latency to the generation process
- System Optimization: End-to-end pipeline must be optimized to achieve reasonable inference speeds
- Evaluation Complexity: Multi-label scenarios require specialized metrics (hamming loss, per-class F1)
- Quality Assessment: Objective metrics must be complemented by subjective human evaluation of musical coherence

5. How will you evaluate your results?

Music Classification Task Evaluation

- **Quantitative Metrics**
    - Multi-label Classification Performance: Hamming Loss, Jaccard Index, and per-class F1 scores to measure accuracy across genres
    - Confusion Matrix Analysis: Identification of commonly confused genre pairs to improve model discrimination
    - Threshold Optimization: Performance at various confidence thresholds to determine optimal binary decision boundaries
- **Qualitative Analysis**
    - Embedding Visualization: t-SNE/UMAP projections to verify meaningful genre clustering in feature space

Music Generation Task Evaluation:

- **Quantitative Metrics**
    - Fréchet Audio Distance (FAD): Quantitative assessment of realism and fidelity compared to original genre samples
    - Musical Feature Analysis: Objective measurements of tempo stability, harmonic consistency, and rhythmic complexity
    - Classification-Generation Alignment: Agreement between input genre conditions and re-classified output using Jaccard Index and Hamming Loss
- **Qualitative Analysis**
    - Human Listening Tests: Expert and non-expert participants rating generated samples on:Genre authenticity (how well samples match intended genres), Musical coherence (particularly for multi-genre combinations), and overall audio quality and appeal.