View Reviews

Paper ID
7317

Paper Title
FUTGA-MIR: Enhancing Fine-grained and Temporally-aware Music Understanding with Music Information Retrieval

Track Name
ICASSP 2025 Main Tracks

Reviewer #2

Questions

2. Importance/Relevance
3. Of sufficient interest

5. Originality/Novelty
3. Moderately original; provides limited new insights or understanding

6. Justification of Originality/Novelty Score (required)
In this paper, authors proposed a music information retrieval augmentation approach on music LLMs. By bootstrapping existing music captions aligned with MIR feature distribution, authors create a synthetic pre-training dataset that simulates the ground-truth data distribution, while significantly increasing the music caption length and audio length.In addition, they further
reduce the sim-to-real gap by aligning the model with human annotated MIR features as feedback.

7. Theoretical Development
4. Correct; provides important new insights or theoretical understanding

9. Experimental Validation
3. Limited but convincing

11. Clarity of Presentation
3. Clear enough

13. Reference to Prior Work
3. References adequate

15. Overall evaluation of this paper
4. Definite accept

16. Justification of Overall evaluation of this paper (required)

In this paper, authors proposed a method which try to bridge the gap between recent music LLMs and conventional music information retrieval tasks, authors propose FUTGA-MIR to enhance the existing music LLMs by augmenting them with MIR features and aligned with human feedback.It is interesting topic and authors did a solid work.

Reviewer #4

Questions

2. Importance/Relevance

3. Of sufficient interest

3. Justification of Importance/Relevance Score (required if score is 1 or 2).

The authors identify a key limitation of current music LLMs: they are trained primarily on music caption datasets that lack MIR features, leading to a gap between these models and conventional MIR tasks.

5. Originality/Novelty

3. Moderately original; provides limited new insights or understanding

6. Justification of Originality/Novelty Score (required)

The paper demonstrates a potential limitation of current music LLMs, in that they are trained on music caption datasets that lack MIR features, leading to a gap between these models and conventional MIR tasks. The paper proposes a model called FUTGA-MIR that appears capable of generating fine-grained music captions conditioned on given time boundaries and MIR features. The idea in itself is interesting and potentially novel, but 1) the definition of "MIR features" is at best vague; 2) we never learn what FUTGA stands for; and 3) reported performance metrics are not accompanied by some statistical significance indicator (e.g., standard error of the mean) to get a more accurate sense of the proposed model's capabilities.

7. Theoretical Development

3. Probably correct; provides limited new insights or understanding

9. Experimental Validation

3. Limited but convincing

11. Clarity of Presentation

3. Clear enough

12. Justification of Clarity of Presentation Score (required if score is 1 or 2).

In addition to never explaining what FUTGA stands for, and a possibly oversimplified usage of the term "MIR features", Table captions are generally dense and repeat what's in the main text already. Figure 2 is not referenced in the text.

SALMONN is used as the backbone music LLM but there is no justification as to why this

was selected over, say, LLARK. Moreover, the approach would be easier to understand via a figure illustrating the pipeline involving synthetic data generation, pre-training, alignment with human annotations, and fine-tuning.

13. Reference to Prior Work

3. References adequate

15. Overall evaluation of this paper

3. Marginal accept

16. Justification of Overall evaluation of this paper (required)

The work is substantial and timely. The paper could benefit from a more defined structure and better organisation/flow of ideas, experiments, and results. A discussion of the latter is also lacking. See further comments above.

Reviewer #5

Questions

2. Importance/Relevance

4. Of broad interest

5. Originality/Novelty

4. Very original; provides important new insights or theoretical understanding

6. Justification of Originality/Novelty Score (required)

The integration of MIR features in music LLMs fills a notable gap, aligning synthetic data with realistic MIR distributions.
Extensive evaluations show FUTGA-MIR's effectiveness, particularly in generating temporally structured, high-quality music captions that support various MIR tasks.
The code and demo are open source.

7. Theoretical Development

4. Correct; provides important new insights or theoretical understanding

9. Experimental Validation

4. Theoretical paper: sufficient validation; Empirical paper: rigorous validation

11. Clarity of Presentation

3. Clear enough

13. Reference to Prior Work

4. Excellent references

15. Overall evaluation of this paper

3. Marginal accept

16. Justification of Overall evaluation of this paper (required)

This paper presents FUTGA-MIR, an innovative approach to improve music LLMs by integrating music information retrieval features, which are often overlooked in existing datasets. The model leverages synthetic and human-annotated data to create a more realistic MIR feature distribution, enhancing fine-grained and temporally-aware music captioning. FUTGA-MIR achieves improvements across downstream tasks like classification, retrieval, and generation, demonstrating substantial advancements over prior music captioning models.