

View Reviews

Paper ID

30

Paper Title

Futga: Towards Fine-grained Music Understanding through Temporally-enhanced Generative Augmentation

Track Name

Papers

Reviewer #1

Questions

2. I am an expert on the topic of the paper.

Strongly disagree

3. The title and abstract reflect the content of the paper.

Strongly agree

4. The paper discusses, cites and compares with all relevant related work

Agree

5. Please justify the previous choice (Required if “Strongly Disagree” or “Disagree” is chosen, otherwise write "n/a")

n/a

6. Readability and paper organization: The writing and language are clear and structured in a logical manner.

Agree

7. The paper adheres to ISMIR 2024 submission guidelines (uses the ISMIR 2024 template, has at most 6 pages of technical content followed by “n” pages of references or ethical considerations, references are well formatted). If you selected “No”, please explain the issue in your comments.

Yes

8. Relevance of the topic to ISMIR: The topic of the paper is relevant to the ISMIR community. Note that submissions of novel music-related topics, tasks, and applications are highly encouraged. If you think that the paper has merit but does not exactly match the topics of ISMIR, please do not simply reject the paper but instead communicate this to the Program Committee Chairs. Please do not penalize the paper when the proposed method can also be applied to non-music domains if it is shown to be useful in music domains.

Agree

9. Scholarly/scientific quality: The content is scientifically correct.

Agree

10. Please justify the previous choice (Required if "Strongly Disagree" or "Disagree" is chosen, otherwise write "n/a")

n/a

11. Novelty of the paper: The paper provides novel methods, applications, findings or results. Please do not narrowly view "novelty" as only new methods or theories. Papers proposing novel musical applications of existing methods from other research fields are considered novel at ISMIR conferences.

Agree

12. The paper provides all the necessary details or material to reproduce the results described in the paper. Keep in mind that ISMIR respects the diversity of academic disciplines, backgrounds, and approaches. Although ISMIR has a tradition of publishing open datasets and open-source projects to enhance the scientific reproducibility, ISMIR accepts submissions using proprietary datasets and implementations that are not sharable. Please do not simply reject the paper when proprietary datasets or implementations are used.

Agree

13. Pioneering proposals: This paper proposes a novel topic, task or application. Since this is intended to encourage brave new ideas and challenges, papers rated "Strongly Agree" and "Agree" can be highlighted, but please do not penalize papers rated "Disagree" or "Strongly Disagree". Keep in mind that it is often difficult to provide baseline comparisons for novel topics, tasks, or applications. If you think that the novelty is high but the evaluation is weak, please do not simply reject the paper but carefully assess the value of the paper for the community.

Agree (Novel topic, task, or application)

14. Reusable insights: The paper provides reusable insights (i.e. the capacity to gain an accurate and deep understanding). Such insights may go beyond the scope of the paper, domain or application, in order to build up consistent knowledge across the MIR community.

Strongly agree

15. Please explain your assessment of reusable insights in the paper.

1)The proposed data augmentation technique allows for generating long music tracks with their descriptions, using a similarity measure on these descriptions.

2) Evaluating complex tasks with other deep learning models demonstrates the strength of the proposed models, even though their evaluation is challenging for humans.

16. Write ONE line (in your own words) with the main take-home message from the paper.

By using LLM, and fine tuning them, it is possible to get fine grained music description, leveraging music boundaries and MIR descriptors.

19. Potential to generate discourse: The paper will generate discourse at the ISMIR conference or have a large influence/impact on the future of the ISMIR community.

Agree

20. Overall evaluation: Keep in mind that minor flaws can be corrected, and should not be a reason to reject a paper. Please familiarize yourself with the reviewer guidelines at <https://ismir.net/reviewer-guidelines>

Strong accept

21. Main review and comments for the authors. Please summarize strengths and weaknesses of the paper. It is essential that you justify the reason for the overall evaluation score in detail. Keep in mind that belittling or sarcastic comments are not appropriate.

The paper explains a novel temporally-enhanced music caption augmentation in order to obtain music caption at local and global times.

A lot of work is presented in the paper. Some fine tuning of SALMONN model (a large audio language models) is done using 1) synthetic data, generated to match, as best as possible, real life music and 2) human annotation and MIR features.

The evaluation process is really thorough and interesting. First, a caption generation evaluation, based on numerous metrics (BLEU score, ROUGE score ...) is performed. The cosine similarity between CLAP score of music and caption is also studied. Finally a music generation task is performed.

All the evaluations help understanding the FUTGA capacities.

The paper is well written and very well explained. I know that the editing process is sometime hard, but I found that it could be better if the tables 1,2,3,4 were better incorporated in the text, as the reading of the last section of the paper is difficult, having to do back and forth between text and tables.

Some additional materials showcasing the synthetic data could help understand better how realistic they are. A t-sne representation is provided, but some more insights could have been given on just how realistic those music tracks are.

typo : 391 Thus, In order

Questions

2. I am an expert on the topic of the paper.

Strongly agree

3. The title and abstract reflect the content of the paper.

Agree

4. The paper discusses, cites and compares with all relevant related work

Agree

5. Please justify the previous choice (Required if “Strongly Disagree” or “Disagree” is chosen, otherwise write "n/a")

Could cite:

McKee, D., Salamon, J., Sivic, J., & Russell, B. (2023). YouTube8M-MusicTextClips (1.0.0) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.8040754>

6. Readability and paper organization: The writing and language are clear and structured in a logical manner.

Disagree

7. The paper adheres to ISMIR 2024 submission guidelines (uses the ISMIR 2024 template, has at most 6 pages of technical content followed by “n” pages of references or ethical considerations, references are well formatted). If you selected “No”, please explain the issue in your comments.

Yes

8. Relevance of the topic to ISMIR: The topic of the paper is relevant to the ISMIR community. Note that submissions of novel music-related topics, tasks, and applications are highly encouraged. If you think that the paper has merit but does not exactly match the topics of ISMIR, please do not simply reject the paper but instead communicate this to the Program Committee Chairs. Please do not penalize the paper when the proposed method can also be applied to non-music domains if it is shown to be useful in music domains.

Strongly agree

9. Scholarly/scientific quality: The content is scientifically correct.

Strongly disagree

10. Please justify the previous choice (Required if "Strongly Disagree" or "Disagree" is chosen, otherwise write "n/a")

1. All evaluation results are computed on the MusicCaps and SongDescriber datasets while both are the source of the training data. Strictly separating training and test data to prevent leakage is a strong requirement for a sound scientific evaluation.

2. The retrieval & text-to-music generation are hard to interpret and provide limited insight into the quality of the proposed captioning paradigm.

3. The paper provides little detail about the quality of the automatic segmentation while it being an integral part of the proposed idea.

11. Novelty of the paper: The paper provides novel methods, applications, findings or results. Please do not narrowly view "novelty" as only new methods or theories. Papers proposing novel musical applications of existing methods from other research fields are considered novel at ISMIR conferences.

Agree

12. The paper provides all the necessary details or material to reproduce the results described in the paper. Keep in mind that ISMIR respects the diversity of academic disciplines, backgrounds, and approaches. Although ISMIR has a tradition of publishing open datasets and open-source projects to enhance the scientific reproducibility, ISMIR accepts submissions using proprietary datasets and implementations that are not sharable. Please do not simply reject the paper when proprietary datasets or implementations are used.

Agree

13. Pioneering proposals: This paper proposes a novel topic, task or application. Since this is intended to encourage brave new ideas and challenges, papers rated "Strongly Agree" and "Agree" can be highlighted, but please do not penalize papers rated "Disagree" or "Strongly Disagree". Keep in mind that it is often difficult to provide baseline comparisons for novel topics, tasks, or applications. If you think that the novelty is high but the evaluation is weak, please do not simply reject the paper but carefully assess the value of the paper for the community.

Agree (Novel topic, task, or application)

14. Reusable insights: The paper provides reusable insights (i.e. the capacity to gain an accurate and deep understanding). Such insights may go beyond the scope of the paper, domain or application, in order to build up consistent knowledge across the MIR community.

Disagree

15. Please explain your assessment of reusable insights in the paper.

The paper lacks some deeper analysis of the proposed method For example it could share more details on segmentation detection or on what additional information gets generated by the fine-tuned model

16. Write ONE line (in your own words) with the main take-home message from the paper.

Stitching together music clips and re-writing their associated captions into coherent text can serve as a synthetic training data source for long-form music captioning.

19. Potential to generate discourse: The paper will generate discourse at the ISMIR conference or have a large influence/impact on the future of the ISMIR community.

Agree

20. Overall evaluation: Keep in mind that minor flaws can be corrected, and should not be a reason to reject a paper. Please familiarize yourself with the reviewer guidelines at <https://ismir.net/reviewer-guidelines>

Strong reject

21. Main review and comments for the authors. Please summarize strengths and weaknesses of the paper. It is essential that you justify the reason for the overall evaluation score in detail. Keep in mind that belittling or sarcastic comments are not appropriate.

The paper presents an interesting approach at generating long-form music captions that provide detailed descriptions over the full-length of a track by identifying prominent sections in the music.

I believe that this indeed a very interesting problem to work on. Further, I am convinced that the paper does a good job at explaining all steps involved in building a synthetic training set by cleverly reusing existing music caption datasets. Using Harmonixset as an additional dataset is a nice idea. Also, the architecture choice seems reasonable.

Nevertheless I suggest to reject this paper.

The primary reason for my decision is the inadequate choice of evaluation for the proposed approach which makes it almost impossible to judge the merit of the work. The synthetic dataset (FUTGA) is based on MusicCaps and SongDescriber which again are used as evaluation dataset. The evaluation data and training data should be different and have no overlap. Instead of using SongDescriber as a training source it might be better to set it aside completely for evaluation. Alternatively, YouTube8M-MusicTextClips might be a suitable evaluation dataset.

In addition to my main concern there are some other aspects that could be improved to make the paper clearer:

- * The paper sometimes talks about "complete songs" when it refers to the patchwork of MusicCaps clips. While this is a nice idea for creating synthetic data of longer 'music' recordings, it is questionable whether those recordings exhibit a structure similar to actual songs. For example, those combined clips will often differ in tempo and harmonic content while those will usually be more consistent in full-length tracks.

Looking at the example in Table 1 it seems that the second half of the 'song' seems somewhat unrelated to the first.

- * FUTGA: The choice of MusicCaps and SongDescriber as source for the synthetic dataset is not obvious to me. If I understand correctly, at the stage of inference no music captions should be required or does the model still need a 'global' caption as input to write the complete caption? Why not use a dataset like MTG-Jamendo/FMA or similar as source which would also give you much more training data?

- * Instead of evaluating in retrieval and generation which both are dependent on the performance of the CLAP model, it might be more interesting to study the segmentation performance of the model.

- * Table 3: The results reported for LP-MusicCaps are reported for the non-finetuned version and thus lower than the top ones reported in [14]. I'm not convinced that this is a fair comparison as MusicCaps data was accessible to the proposed model indirectly through Futga. Additionally, the results on SongDescriber do not match the ones presented in [22].
- * Sec 2.1: When querying for the most similar clips it might be better to (additionally) use a model for music similarity than just a text model.
- * Figure 2: It's not clear what model is used to obtain the embeddings for the representation. Also MiniLM-L12? Or a music embedding model?
- * Sec 2.1, line 196: This part is very hard to follow. I am assuming that clips are repeated consecutively?
- * Sec: 1, line 39: what do you mean with multi-tasking
- * Sec 1, line 50: Did you mean music tagging datasets?
- * Sec 3, 279: More details on the annotators would be nice.
- * Sec 4, 300: It's not clear to me what "musical changes are"
- * Sec 5.2: A few steps are unclear here: How reasonable is it to have segmented descriptions for MusicCaps clips that are only 10 seconds long? What 'audio clip time segments' are used for Song Describer dataset (they are not pre-segmented)? Is the process applied to the entire datasets?
- * Sec 5.2, line 377 and following: Might it not simply be the case that human-annotated captions are better?
- * Table 2: The caption should explain what the columns starting with '#' mean.

Minor things:

- * Sec 3: Mention the length of tracks in Harmonixset.
- * Prefer full conference references over arXiv pre-prints, e.g. use the ISMIR reference for [14]
- * Fix typesetting of reference [19]
- * Sec. 5.3, line 391: typo
- * Sec 1, line 66: It's a bit unclear what you mean with MIR features here, maybe introduce your approach again here before listing details.
- * Sec 1, line 68: a reader unfamiliar with MusicCaps dataset will have trouble following here
- * Sec 1, line 79 and following: It might make sense to move this paragraph up and insert it after the sentence in line 66.
- * "sim-to-real gap" could be introduced as an acronym.
- * Sec 1, line 119: Explicitly mention SALMONN
- * Sec 1, line 123: "*our* music understanding model" might be better.
- * Move table 1 up so that the other tables are close to the text that reference them.

Questions

2. I am an expert on the topic of the paper.

Agree

3. The title and abstract reflect the content of the paper.

Agree

4. The paper discusses, cites and compares with all relevant related work

Agree

5. Please justify the previous choice (Required if “Strongly Disagree” or “Disagree” is chosen, otherwise write "n/a")

n/a

6. Readability and paper organization: The writing and language are clear and structured in a logical manner.

Agree

7. The paper adheres to ISMIR 2024 submission guidelines (uses the ISMIR 2024 template, has at most 6 pages of technical content followed by “n” pages of references or ethical considerations, references are well formatted). If you selected “No”, please explain the issue in your comments.

Yes

8. Relevance of the topic to ISMIR: The topic of the paper is relevant to the ISMIR community. Note that submissions of novel music-related topics, tasks, and applications are highly encouraged. If you think that the paper has merit but does not exactly match the topics of ISMIR, please do not simply reject the paper but instead communicate this to the Program Committee Chairs. Please do not penalize the paper when the proposed method can also be applied to non-music domains if it is shown to be useful in music domains.

Strongly agree

9. Scholarly/scientific quality: The content is scientifically correct.

Agree

10. Please justify the previous choice (Required if "Strongly Disagree" or "Disagree" is chosen, otherwise write "n/a")

n/a

11. Novelty of the paper: The paper provides novel methods, applications, findings or results. Please do not narrowly view "novelty" as only new methods or theories. Papers proposing novel musical applications of existing methods from other research fields are considered novel at ISMIR conferences.

Agree

12. The paper provides all the necessary details or material to reproduce the results described in the paper. Keep in mind that ISMIR respects the diversity of academic disciplines, backgrounds, and approaches. Although ISMIR has a tradition of publishing open datasets and open-source projects to enhance the scientific reproducibility, ISMIR accepts submissions using proprietary datasets and implementations that are not sharable. Please do not simply reject the paper when proprietary datasets or implementations are used.

Agree

13. Pioneering proposals: This paper proposes a novel topic, task or application. Since this is intended to encourage brave new ideas and challenges, papers rated "Strongly Agree" and "Agree" can be highlighted, but please do not penalize papers rated "Disagree" or "Strongly Disagree". Keep in mind that it is often difficult to provide baseline comparisons for novel topics, tasks, or applications. If you think that the novelty is high but the evaluation is weak, please do not simply reject the paper but carefully assess the value of the paper for the community.

Agree (Novel topic, task, or application)

14. Reusable insights: The paper provides reusable insights (i.e. the capacity to gain an accurate and deep understanding). Such insights may go beyond the scope of the paper, domain or application, in order to build up consistent knowledge across the MIR community.

Disagree

15. Please explain your assessment of reusable insights in the paper.

The authors propose a synthetic data augmentation method to produce temporally fine-grained captions and fine tune a LLM model on the augmented caption dataset. The output of the LLM is evaluated as a dataset used as input to other downstream models, e.g., captioning, generation or similarity based retrieval. Unfortunately, the dataset is not publicly available nor the fine-tuned model. For this reason I think the reusable insights are quite limited.

16. Write ONE line (in your own words) with the main take-home message from the paper.

It is possible to concatenate audio segments and captions, further processed with an LLM, for data augmentation.

19. Potential to generate discourse: The paper will generate discourse at the ISMIR conference or have a large influence/impact on the future of the ISMIR community.

Agree

20. Overall evaluation: Keep in mind that minor flaws can be corrected, and should not be a reason to reject a paper. Please familiarize yourself with the reviewer guidelines at <https://ismir.net/reviewer-guidelines>

Weak accept

21. Main review and comments for the authors. Please summarize strengths and weaknesses of the paper. It is essential that you justify the reason for the overall evaluation score in detail. Keep in mind that belittling or sarcastic comments are not appropriate.

The paper presents two main contributions: a method to augment a captioned music dataset and a fine-tuned model to caption full-length songs.

According to the proposed data augmentation strategy, a new audio-caption pair is produced by concatenating short audio excerpts and their captions, respectively. Captions are embedded with information related to the temporal starting point of the relative audio segment, and further paraphrased with an LLM.

A music captioning system is then fine-tuned using the built augmented dataset, with the addition of some music tags (genre and BPM). The model is then able to caption a full song on a fine temporal grid, highlighting possible changes in time.

The proposed method for data augmentation and fine-tuning sounds reasonable, even though not all choices are particularly well motivated. As an example, similar audio tracks to be concatenated are chosen based on their caption similarity. This makes sense, but could also lead to unpleasant situations in which tracks with strongly different BPMs are concatenated. Why not using an audio signal information then?

However, in my opinion, the main issue is related to the presentation of the results and the performed analysis. The authors perform a series of different tests on different tasks using differently fine-tuned version of the proposed model. This makes very difficult to understand which is the final version of the proposed model, as well as which tests are significant. I would recommend making the results analysis more clear.

Moreover, the authors are not publicly distributing the augmented dataset or the fine-tuned model. Hence the reusable insights are limited to the augmentation strategy.

View Meta-Reviews

Paper ID

30

Paper Title

Futga: Towards Fine-grained Music Understanding through Temporally-enhanced Generative Augmentation

Track Name

Papers

META-REVIEWER #1

META-REVIEW QUESTIONS

2. I am an expert on the topic of the paper.

Agree

3. The title and abstract reflect the content of the paper.

Agree

4. The paper discusses, cites and compares with all relevant related work.

Agree

5. Please justify the previous choice (Required if “Strongly Disagree” or “Disagree” is chosen, otherwise write "n/a")

n/a

6. Readability and paper organization: The writing and language are clear and structured in a logical manner.

Strongly disagree

7. The paper adheres to ISMIR 2024 submission guidelines (uses the ISMIR 2024 template, has at most 6 pages of technical content followed by “n” pages of references or ethical considerations, references are well formatted). If you selected “No”, please explain the issue in your comments.

Yes

8. Relevance of the topic to ISMIR: The topic of the paper is relevant to the ISMIR community. Note that submissions of novel music-related topics, tasks, and applications are highly encouraged. If you think that the paper has merit but does not exactly match the topics of ISMIR, please do not simply reject the paper but instead communicate this to the Program Committee Chairs. Please do not penalize the paper when the proposed

method can also be applied to non-music domains if it is shown to be useful in music domains.

Strongly agree

9. Scholarly/scientific quality: The content is scientifically correct.

Disagree

10. Please justify the previous choice (Required if “Strongly Disagree” or “Disagree” is chose, otherwise write "n/a")

While evaluation results are presented, evaluation of the overall system and evaluation of individual components lacks a systematic approach, justification, and convincing analysis. After reading the paper, I could not summarize the findings except that the new approach is much better.

11. Novelty of the paper: The paper provides novel methods, applications, findings or results. Please do not narrowly view "novelty" as only new methods or theories. Papers proposing novel musical applications of existing methods from other research fields are considered novel at ISMIR conferences.

Agree

12. The paper provides all the necessary details or material to reproduce the results described in the paper. Keep in mind that ISMIR respects the diversity of academic disciplines, backgrounds, and approaches. Although ISMIR has a tradition of publishing open datasets and open-source projects to enhance the scientific reproducibility, ISMIR accepts submissions using proprietary datasets and implementations that are not sharable. Please do not simply reject the paper when proprietary datasets or implementations are used.

Disagree

13. Pioneering proposals: This paper proposes a novel topic, task or application. Since this is intended to encourage brave new ideas and challenges, papers rated “Strongly Agree” and “Agree” can be highlighted, but please do not penalize papers rated “Disagree” or “Strongly Disagree”. Keep in mind that it is often difficult to provide baseline comparisons for novel topics, tasks, or applications. If you think that the novelty is high but the evaluation is weak, please do not simply reject the paper but carefully assess the value of the paper for the community.

Disagree (Standard topic, task, or application)

14. Reusable insights: The paper provides reusable insights (i.e. the capacity to gain an accurate and deep understanding). Such insights may go beyond the scope of the paper, domain or application, in order to build up consistent knowledge across the MIR community.

Disagree

15. Please explain your assessment of reusable insights in the paper.

The combination of different models, datasets, and features as well as training strategies contains undoubtedly useful components. As no detailed analysis of components is performed, true insights are very limited.

16. Write ONE line (in your own words) with the main take-home message from the paper.
More generative models for more subtasks improve the captioning task.

19. Potential to generate discourse: The paper will generate discourse at the ISMIR conference or have a large influence/impact on the future of the ISMIR community.
Agree

20. Overall evaluation (to be completed before the discussion phase): Please first evaluate before the discussion phase. Keep in mind that minor flaws can be corrected, and should not be a reason to reject a paper. Please familiarize yourself with the reviewer guidelines at <https://ismir.net/reviewer-guidelines>.
Weak reject

21. Main review and comments for the authors (to be completed before the discussion phase). Please summarize strengths and weaknesses of the paper. It is essential that you justify the reason for the overall evaluation score in detail. Keep in mind that belittling or sarcastic comments are not appropriate.

I trust that the authors have found an interesting and effective extension of existing music captioning approaches but the paper in itself does not convince me scientifically. As a reader I got the impression, I should know the work already in detail to understand what is going on. Highly complex and specific generative models are floating around this paper, some serving as baseline, data source, and evaluation criteria at the same time, others providing (I'm tempted to say magically) new capabilities. The problem addressed is never explained stepwise and neither are the components introduced. Terms like "music understanding" are also used without critical distance and such capabilities claimed. Especially at ISMIR, such claims should be better reflected upon.

Overall the paper is very confusing. A lot is presented but not a lot is explained.

22. Final recommendation (to be completed after the discussion phase) Please give a final recommendation after the discussion phase. In the final recommendation, please do not simply average the scores of the reviewers. Note that the number of recommendation options for reviewers is different from the number of options here. We encourage you to take a stand, and preferably avoid "weak accepts" or "weak rejects" if possible.
Reject

23. Meta-review and final comments for authors (to be completed after the discussion phase)

Overall the ideas of the paper are seen positive but there are strong reservations expressed in the reviews. During the discussion (more than in the individual reviews), there was consensus that the paper exhibits severe flaws (or fails to clearly describe the settings otherwise),

especially in the use of models for presumably both data and evaluation. I strongly urge the authors to revisit their setup and present a corrected setup in a more lucid manner. If results still hold up, a resubmission is encouraged.
