View Reviews

Paper ID
237

Paper Title
An Empirical Evaluation of End-to-End Polyphonic Optical Music Recognition

Track Name
Papers

Reviewer #1

Questions

2. The title and abstract reflect the content of the paper.
Agree

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

4. The writing and language are clear and structured in a logical manner.
Strongly Agree

5. The paper adheres to ISMIR 2021 submission guidelines (uses the ISMIR 2021 template, has at most 6 pages of technical content followed by "n" pages of references, references are well formatted). If you selected "No", please explain the issue in your comments.
Yes

6. The topic of the paper is relevant to the ISMIR community.
Strongly Agree

7. The content is scientifically correct.
Strongly Agree

8. The paper provides novel methods, findings or results.
Agree

9. The paper provides all the necessary details or material to reproduce the results described in the paper.
Strongly Agree

10. 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.
Agree

11. Please explain your assessment of reusable insights in the paper.
The paper provides a potential solution to end-to-end polyphonic OMR; however, the model may not work well on other datasets as stated by the authors.

15. The paper will have a large influence/impact on the future of the ISMIR community.
Disagree

17. Overall evaluation
Weak Accept

18. Main review and comments for the authors
This paper is focused on end-to-end polyphonic OMR task, and provides details about creating a dataset for the task using MuseScore files as well as two decoder strategies for the task. The paper is well organized and related background and details about the workflow is provided.

Several questions that arose after reading the paper are as below:

- How would other approaches to polyphonic OMR perform using this specific dataset? It would be nice to include related works on polyphonic OMR and include them as baselines in this work.

- Why was the number of 900 chosen as the size of the MSPD-Hard? Would it make more sense to do stratified sampling for train/val/test split using density, so that we could make sure there would be enough "hard" training examples for the model to learn from?

- Is there any assumption on what kind of information could have been captured by the hidden states of RNN? Would attention mechanism also work or work even better than RNN?

Reviewer #2

Questions

2. The title and abstract reflect the content of the paper.
Strongly Agree

3. The paper discusses, cites and compares with all relevant related work.
Strongly Agree

4. The writing and language are clear and structured in a logical manner.
Strongly Agree

5. The paper adheres to ISMIR 2021 submission guidelines (uses the ISMIR 2021 template, has at most 6 pages of technical content followed by “n” pages of references, references are well formatted). If you selected “No”, please explain the issue in your comments.
Yes

6. The topic of the paper is relevant to the ISMIR community.
Agree

7. The content is scientifically correct.
Strongly Agree

8. The paper provides novel methods, findings or results.
Strongly Agree

9. The paper provides all the necessary details or material to reproduce the results described in the paper.
Strongly Agree

10. 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.
Agree

11. Please explain your assessment of reusable insights in the paper.
two interesting new models applicable to polyphonic OMR

15. The paper will have a large influence/impact on the future of the ISMIR community.
Agree

17. Overall evaluation
Weak Accept

18. Main review and comments for the authors
This paper expands on a previous encoder-decoder approach to OMR for monophonic scores and propose two new methods, both of which outperform the previous method on polyphonic material. While the encoder is the same for all three methods (CNN and bidirectional LSTM), the original decoder which uses two parallel fully connected layers (FCL), one for pitch, one for rhythm, is replaced with a model separating a pitch/rhythm matrix

(two FCL) and a binary vector (one FCL) for staff symbols, and an RNN, respectively. Since previous datasets are either limited in size or specialized on monophonic scores, the authors also generate a new dataset based on MuseScore, which they then use in their evaluation. While both models outperform the baseline, the RNN reduces the error rate of the reference model by half or even a third. The authors also provide some qualitative results which illustrate the differences in results well.

Overall, the methods and results are presented in a clear way and the performance of the methods is impressive, especially for more complex examples, which were separated into a subset of the dataset, and especially given the noisiness of the examples arising due to incomplete cropping, which the authors which the authors argue may lead to more robust results.

abstract:
-availlable -> available

intro:
-inputted -> input
-'thus is where..' -> 'this is where'

3:
-interesting point about the more realistic environment provided by improper cropping. is there any way for you to measure how much this affects the success rate of the methods? or an estimate of the proportion of improperly cropped images? if so, these may be useful additions to the final version

3.2:
-'less measures' -> fewer measures

5:
-the qualitative examples contain errors that may be successfully corrected with post-processing methods, which could be done in future work. for example, in the RNN result in Figure 8, the quarter rest and the following grouping of 32nd notes could be identified as impossible

Reviewer #3

Questions

2. The title and abstract reflect the content of the paper.
Disagree

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

4. The writing and language are clear and structured in a logical manner.
Disagree

5. The paper adheres to ISMIR 2021 submission guidelines (uses the ISMIR 2021 template, has at most 6 pages of technical content followed by “n” pages of references, references are well formatted). If you selected “No”, please explain the issue in your comments.
Yes

6. The topic of the paper is relevant to the ISMIR community.
Agree

7. The content is scientifically correct.
Agree

8. The paper provides novel methods, findings or results.

Agree

9. The paper provides all the necessary details or material to reproduce the results described in the paper.
Strongly Disagree

10. 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

11. Please explain your assessment of reusable insights in the paper.
This is a paper about OMR. Within OMR, it provides reusable insights but not beyond the scope of the paper.

15. The paper will have a large influence/impact on the future of the ISMIR community.
Disagree

17. Overall evaluation
Weak Accept

18. Main review and comments for the authors
This work goes one step further on the existing OMR literature with deep learning. For this, in addition to a specific data creation with which to carry out new experiments, authors propose a couple of approaches to decode an image in terms of a sequence of notes that is better adapted to the polyphonic context. Fortunately, this work builds on previous approaches and does not start everything from scratch. Also, they compare with previous work to demonstrate an effective contribution to the state of the art.

One of my main concerns is what the work promises in its title and in much of its wording. The authors speak of polyphonic OMR but it should remain clear that they refer to excerpts depicting a single staff (as opposed to, for example, the typical piano scores consisting of two staves in a single system). In addition, unless I have not understood well, the output is limited to providing the sequence of token/notes in the image but in no case are the voices recovered, that is, there is no continuity. This means that, going from the result to a specific encoding (MusicXML, MEI, Humdrum ...) is very far from trivial and would require advanced algorithms to do so. Although this last step can be considered as another specific problem to deal with, I think authors should be more explicit in what they achieve with their model.

My second concern, that I consider a serious flaw in the manuscript, is the difficulty of understanding how the proposed decoders work and, in particular, how they are trained. I can't quite understand how CTC is integrated into FlagDecoder and RNNDecoder. For the former, the explanation "The main difference with this decoder's implementation is that when optimizing for CTC (...) thus the probabilities of each symbol must be manually calculated as opposed to directly being able to train using the immediate output of the neural network." is vastly insufficient. For the latter, nothing is even mentioned.

I am willing to recommend acceptance of this work despite these two concerns but I would ask the authors to please take these comments into account when preparing their eventual camera-ready version. It's a pity that a good research is overshadowed by issues that are (or should be) easy to fix.

Furthermore, as a minor detail, I think that one of the main problems with this kind of research --as mentioned in the paper-- is the transfer of models trained with synthetic data to real case. Perhaps, instead of just making that critique, authors could have started laying the foundations to address this by providing a baseline result with their model on real data.

View Meta-Reviews

Paper ID
237

Paper Title
An Empirical Evaluation of End-to-End Polyphonic Optical Music Recognition

Track Name
Papers

META-REVIEWER #1

META-REVIEW QUESTIONS

2. The title and abstract reflect the content of the paper.
Agree

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

4. The writing and language are clear and structured in a logical manner.
Agree

5. The paper adheres to ISMIR 2021 submission guidelines (uses the ISMIR 2021 template, has at most 6 pages of technical content followed by “n” pages of references, references are well formatted). If you selected “No”, please explain the issue in your comments.
Yes

6. The topic of the paper is relevant to the ISMIR community.
Strongly Agree

7. The content is scientifically correct.
Agree

8. The paper provides novel methods, findings or results.
Strongly Agree

9. The paper provides all the necessary details or material to reproduce the results described in the paper.
Agree

10. 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.
Agree

11. Please explain your assessment of reusable insights in the paper.
The ideas can be expanded and it looks like the source code will be made public.

15. The paper will have a large influence/impact on the future of the ISMIR community.
Agree

17. Overall evaluation
Strong Accept

18. (Initial) Main review and comments for the authors
The paper proposes end-to-end polyphonic OMR with two formulations: "one treating the problem as a type of multi-task binary classification" (FlagDecoder) and the other treating it as multi-sequence detection" (RNNDecoder).

It seems like an interesting and novel approach, although some details are unclear, such as the explanation of RNNDecooder, which may be due to a lack of space.
I would appreciate a more clear explanation of the handling of accidentals. It looks like the accidentals are

recognized in the key signatures but not elsewhere?
I'm also not understanding "single horizontal slices" (Introduction) and "Figure 4. Horizontally sliced image features"

The created dataset should be a valuable resource for future research.

A clearer explanation of the below would be appreciated:
"To generate the dataset, we first used the MuseScore plugin API to resize the page height of the rendering to allow for generating single staff images. Then we executed a script to remove credit text which covered up some scores."

Overall, the English is fine but one more check with an English-speaking editor should improve it.

Other edits and comments:

Abstract
Why just: "string and orchestral scores"? What about piano or winds scores?

"homophonic music can be described as having a single musical rhythm"
The proper term is "homorhythmic music".

1. Introduction
"The accuracy of systems built using these tools is very exciting" -->
"The improvement in the accuracy of sytems built using these tools is very exciting"

2. Background
"Previous work on polyphonic OMR has been limited." -->
"Previous published work on polyphonic OMR has been limited."

3. MuseScore
"keys" --> "key signatures"

MPSD-Hard examples in Figure 1 don't looks like real music.

"Less measures" --> "fewer measures"

How is number of voices defined?

4.4 FlagDecoder
"each 'symbol'" should use quote signs.

4.5 RNNDecoder
Probably a space issue but this section is difficult to understand,
I don't understand: "Since the largest amount of symbols we observed in a single horizontal position of an image was 10, "

5.3 Experimental results
"Identifying the correct three pitches". What three pitches?

5.4 Error
Figure 7 is unusual; Figure 8 is normal

"required compute." --> "required computing resources,"

7. References
[1] "H." --> "Hajič"

[6] "Hajic Jr" --> "Hajič jr."
[9] "Hajic Jr" --> "Hajič jr."
[20] "Hajič" --> "Hajič jr."
[21] "labelling" --> "Labelling"
[22] "lstm" --> "LSTM"

19. Meta-review and final comments for authors

We find that the "paper is well organized and related background and details about the workflow is provided." It proposes new OMR methodologies with "impressive" results. Although we are accepting this paper, it is essential that, in the camera-ready version of the paper, the authors explain in much more detail how the neural network is optimized.