View Reviews

Paper ID
149

Paper Title
AugmentedNet: A Roman Numeral Analysis Network with Synthetic Training Examples and Additional Tonal Tasks

Track Name
Papers

Reviewer #1

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.
Strongly 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.
Strongly Agree

11. Please explain your assessment of reusable insights in the paper.
The performance of roman numeral analysis can be improved by multi-task learning and even adding a new task.

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

17. Overall evaluation
Strong Accept

18. Main review and comments for the authors
This paper presents the improved performance of functional harmony analysis by convolutional recurrent neural networks. The improvement has been achieved by combining synthetic training examples for data augmentation and solving a multi-task estimation. The paper makes an original and substantial contribution to the field.

Although both strategies using synthetic data and multi-task learning have been often used, this paper has a novelty in highlighting that symbolic music analysis has the property of multi-task on its own. Creating a new task for improving performance is also interesting.

The work can produce stimuli to other symbolic music analysis tasks. Symbolic music processing, including roman numeral analysis, suffers from a lack of data with human annotations. This work indicates one possible solution for the problem.

Minor points:
l. 218: "The output of the network follows a MTL approach with hard parameter sharing": this seems the weights and biases in the final dense layers are the same for all tasks, which is hard to understand. The "hard parameter"s should be the ones in the networks in the front, such as CNNs and RNNs. The reviewer recommends rewriting the sentence to avoid misunderstanding of the readers.

l. 289: "in the target range of key signatures": the actual target range is unclear. Although the readers could understand the notes are in the range that does not require too many lines, it would be clearer if you could indicate the range for each voice.

The performance of roman numeral analysis can be improved by multi-task learning and even adding a new task.

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.
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.
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.
Strongly Agree

11. Please explain your assessment of reusable insights in the paper.
The archive provided by the authors has already some data. The authors also pledge to publicly release the code. More over, the ideas of synthetic examples / more tasks could also be reused by others.

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

The authors present AugmentedNet, a convolutional neural network to predict functional harmony in the form of Roman Numerals (RN) starting from a score in symbolic format. Several recent papers (2019-2021) worked on the same task, predicting 6 outputs (local key, chord root, quality, primary degree, secondary degree, inversion). Although similar, the input encoding and the network architecture are here different from what was in (Micchi, 2020), in particular within the layout of the convolutional layers. Moreover, the authors bring two key ideas. The first one is to augment data, not only through transposition, but also by crafting synthetic scores with a combination of block chords and some “texturization“ of the same chords. The second one is to train the model to solve at once *more* “tonal tasks” than the 6 outputs, hoping to improve the performance. These additional tasks are somehow included in the other ones but may help to network to learn more high-level concepts.

Altogether, the authors compare favorably against the recent papers with a model of similar size (about 90,000 weights). In particular, they have about 3% improvement on each task over (Micchi, 2020). They also introduce a different output layer, focusing on the 75 most present RN that account for 98% of the RN. They report between 46% and 61% accuracy with these "75RN". The authors already provided an archive with data and pledge to publicly release the code

This study is interesting: The field is very active and gaining even a few percent of performance requires improving our understanding of how these models work with harmony. The two key ideas (synthetic examples and additional tonal tasks) are very appealing. My major remark is that I would like to better know whether and how these two ideas improve the performance. Indeed, improvement may also have come from the network layout and the encoding. I would thus like to see somewhere a comparison of the results of AugNet against itself, with and without the synthetic examples and/or the additional tonal tasks. Note that even if one of these points may not bring improvement, the study is still interesting in understanding what happens.

Some minor remarks

- In the results of AlltRN, how are handled the 2% of “non-75RN” chords? If they are removed from the experiment, such results are not directly comparable to the RN column (and AltRN should not be put in bold in this case).

- 407/408: Is it really “their available training data” against “our available training data”? (It is not exactly fair if it’s the case).

- It would be worth to have more music discussion on the results, for example by detailing which of the 75RN are correctly/badly predicted

Reviewer #3

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.
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.
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.
Disagree

11. Please explain your assessment of reusable insights in the paper.
While this paper makes contributions beyond the SoTA in automatic roman numeral analysis, the contributions are fairly specific to this problem and do not offer any insights which might be more generally applicable. Ideas like data augmentation using synthetic training examples and multi-task learning are already known to be useful techniques. However, the variety in texturization procedures might be useful in other areas.

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

17. Overall evaluation
Weak Reject

18. Main review and comments for the authors
This paper presents a neural network architecture and training strategy for automatic roman numeral analysis. Building on previous work by Micchi et al. in ISMIR 2020, they propose several strategies to improve the network architecture and training procedure, and demonstrate an improvement over prior results.

While the results are impressive, it is unclear to me why the improvements over the work of Micchi et al. are significant enough to constitute a contribution. The network architecture, multi-task and transposition-based augmentation seem to be a slight modification of Micchi's approach, and the synthetic training examples seem to be a modification of the approach from Nápoles López and Fuginaga. While combining these elements is novel, it is unclear how much each modification is contributing to the increase in performance, and whether they are even all helpful.

While I appreciate detailed description, this paper has a little too much "what" and not enough "why", which makes it difficult to identify elements of this work which are applicable to problems beyond roman numeral analysis. Moving some of the implementation details (e.g. section 4.2) to the supplemental materials and providing more discussion of the reasons for different modeling decisions would help a lot. For instance, why did you arrive at the specific four

texturization procedures you used? are they based on training experiments, observations about the common practice period composers in the datasets, or just your intuition?

Lastly, as someone unfamiliar with the specific chord representations used, I find Table 3 very difficult to read. Are these numbers all accuracy values? How many classes does the model have to choose from in each column? What is the difference between 75RN and AltRN?

For these reasons I'm inclined to reject, though I could be persuaded to accept if the authors commit to revisions.

View Meta-Reviews

Paper ID
149

Paper Title
AugmentedNet: A Roman Numeral Analysis Network with Synthetic Training Examples and Additional Tonal Tasks

Track Name
Papers

META-REVIEWER #1

META-REVIEW 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.
Strongly 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.
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.
Disagree

11. Please explain your assessment of reusable insights in the paper.
The authors promise to offer the code in an open-access repository, which is great. It's possible that some of the methods and findings here could be extended to other problems and tasks in the MIR community. That said, the authors don't evaluate various architectures for the model (instead simply choosing to compare the best version of the model to models from previous publications), so readers won't know if (and how much) strategies like texturizing contributed to the model's performance. From this perspective, I can't be sure how relevant these improvements to the model really are.

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

17. Overall evaluation

Strong Accept

18. (Initial) Main review and comments for the authors

Summary

The authors present AugmentedNet, a convolutional recurrent neural network that predicts functional harmony (i.e., Roman numeral) annotation labels. Relative to previous studies ([4,5] in their references list), the authors examine the model using additional tonal tasks and augment the data sets in new ways, including by creating synthetic versions of the symbolic surface by "texturizing" the surface using well-known patterns (e.g., Alberti bass). The authors also evaluate the model using data sets reported in the studies mentioned above, as well as using several other recently published data sets (e.g., TAVERN, ABC, etc.). Their results outperform the previously mentioned models, though the improvements relative to Mi20 [5] appear to be meager for most data sets.

Comments

Strengths

The paper is beautifully written and well organized. The introduction and related work clearly demonstrate the authors' knowledge of functional harmony, and the suggestion to use multi-task learning by predicting several sub-tasks is a clever solution to a complex problem (following [4], of course). I also appreciated the additional approach to data augmentation (not just transposition, but texturizing, as well). The authors offer snippets of code and promise to publish the code and data in an open-access repository, which is great (and necessary, given that this is a comparison study). Finally, the model, based on my limited knowledge, is clearly presented and produces publishable results.

Weaknesses

I understand why the authors wanted to compare the best version of the model to previous publications, but given the meager improvements in performance, I was expecting Table 3 to include comparisons of AugmentedNet's various architectures. For example, how did the augmentation strategies (texturizing, transposition, etc.) contribute to model performance? Including comparisons of several architectures for AugmentedNet was necessary here, I think. I also note a little confusion in my specific comments about the evaluation measure (accuracy?) and approach to splitting the data for training and test (train-validation-test vs. cross-validation). Given the meager size of these data sets relative to those in the NLP community, I also wonder if less sophisticated models might do quite well here, too (e.g., prediction by partial match). PPM models could serve as a useful baseline for the deep learning models, for example. Finally, I also suggest an alternative evaluation measure that the authors might consider for future work, particularly given how the NLP community evaluates these sorts of models for natural language corpora (i.e., using information-theoretic measures like corpus cross-entropy).

I offer specific comments below, some of which expand on my suggestions above.

Specific Comments

Line 10 – "resulting [from] applying…"
Line 14 – "transposition[,] which…" (nonrestrictive clauses receive commas)
Line 72 – "Beside[s] the architecture…"
Line 217 – "a[n] MTL approach…"
Line 331 – "To constraint" should be "To constrain"
Section 4.3 – Is there any particular reason why the authors didn't report results for a composite data set that aggregates the individual data sets? As a reader, I would have hung my hat on those accuracy estimates, particularly given some of the strange annotation schemes specific to some corpora (TAVERN sometimes remains in the local key even when the theme or variation modulates to the dominant, producing strings of applied chords that clearly reflect modulation rather than tonicization (e.g., V64/V--V7/V--V).
Section 4.3 – I was expecting to find a model evaluation section, and so had to guess in 4.3 that the authors are simply reporting accuracy estimates (i.e., for each predicted chord onset, did the prediction match the observed label?). This point could have been clearer. Also, if the authors used cross-validation to compare with [4], why not use cross-

validation throughout? I'm not entirely clear on this issue, admittedly, but reporting means and standard deviations across the test folds would allow for better direct comparison (i.e., inferential evidence) that the model significantly outperformed the results in [5], for example.
Section 4.3 – The authors didn't compare several of the model inputs in Table 3, so we don't know how much the various data augmentation strategies contributed to the improved model performance. Texturizing, for example, seems quite clever, but how much did it help the model?
Section 4.3 – No doubt the authors use this evaluation measure for comparison purposes, but I'll suggest an additional measure for the authors to consider, particularly given what I see in NLP (and sometimes at ISMIR as well). Performance measures like accuracy relate to classification tasks (to my mind, anyway), but I'm not sure this is necessarily a classification task. The authors state that they're attempting to "reconstruct" the sequence of labels, which could be viewed as a prediction task. If the model outputs a probability distribution across the alphabet of labels, the authors could instead (or additionally) report evaluation measures related to the *predictive accuracy* of the model. Language models predicting sequences of words often use measures like corpus cross-entropy (Hm; a measure of the average information content across the sequence) for this purpose. Theoretically, if two models produced identical accuracy estimates for a test corpus (i.e., produced the same sequence of symbols, with the same errors between the predicted and observed symbols), we'd privilege the model that minimized its uncertainty for the predicted sequence (i.e., produced larger probability estimates for the correct predictions, and smaller probability estimates for the incorrect predictions). I personally think this is a more nuanced evaluation measure for this task (see, for example, Korzeniowski, Sears, & Widmer, 2018).

19. Meta-review and final comments for authors

The authors tackle a well-known musicological problem (RNA) that has recently attracted the interest of the MIR community (e.g., [4,5] in the references list). Relative to previous studies, the paper offers additional data augmentation strategies (e.g., texturizing), a new model architecture, other vocabulary strategies (e.g., RN75), additional tonal tasks, and model evaluations based on additional data sets. What is more, the models outperform previously published models, though the improvements relative to Mi20 [5] appear to be meager for most data sets.

The reviewers unfortunately did not reach consensus in their overall evaluations, but based on the arguments above, my final recommendation is 'Strong Accept'. That being said, the reviewers raised several concerns about the current approach that the authors should consider going forward. Three reviewers noted, for example, that the paper should evaluate several architectures for AugmentedNet in Table 3 in order to determine how the various proposed modifications (e.g., texturizing) contributed to the final model's performance. We hope that the authors will address criticisms like this one in the final manuscript if the paper is accepted for publication in the proceedings.