==========
ISMIR 2018 Reviews for Submission #218
==========

Title: Convolutional Generative Adversarial Networks with Stochastic Binary Neurons for Polyphonic Music Generation

Authors: Hao-Wen Dong and Yi-Hsuan Yang
==========
REVIEWER #1
==========

Reviewer's Scores

Scholarly/scientific quality: medium high
Novelty: high
Relevance of topic: high
Importance: high
Readability and paper organisation: medium low
Title and abstract: yes
Bibliography: yes
Make Review Publicly Accessible: No

Comments

Summary: the authors propose a generative adversarial network for polyphonic music generation capable to "compose" binary-valued piano-rolls. They investigate if via (directly) generating binary-valued piano-rolls one can overcome the problem of generating overly-fragmented piano-rolls (caused by hard thresholding or Bernoulli sampling in current techniques).

Title: since deterministic binary neurons are also used, maybe the title should be "Convolutional generative adversarial networks with binary neurons for polyphonic music generation"?

The discussion about the idea that real-valued outputs by G my lead to difficulties in training D (and Figure 2) is interesting, but there is no evidence backing these speculations. Further, I'm not sure if a real-valued piano-roll can be "easily" classified as real data by the D -- because if G outputs real-valued piano-rolls and true data is binary, D just needs to check if its inputs are real-valued or binary data (what would result in a simple and accurate D that did not learn anything that was musically relevant).

Although I'm closely following the deep learning literature, my field of

expertise is not GANs and I have never used networks having binary-outputs. I guess that for this reason I'm not familiar with the following terms: straight-through estimator, or the slope annealing trick. If these terms are not widely used by the ISMIR community, I would recommend to better explain those.

Generator: I like the idea of the private networks. However, I miss an study comparing shared and shared/private design (similar to what is done with the discriminator).

Refiner: the authors propose using residual networks. Have you considered using U-nets? It would be nice to compare the residual networks based refiner against another having no residual connections.

Discriminator: I like the idea of the private networks, and I also enjoyed the study considering the ablated version. Related to that, maybe it would be nice to mention that this would be studied when the private networks are introduced. I did not understand the last paragraph of section 3.4 (intra-bar, inter-bar, etc.), I suggest to better explain this paragraph.

Dataset: Why authors are only considering songs with the alternative tag? Due to this choice, the training dataset is now much smaller. This needs to be further
justified. Related to that, is this the reason why the performance of the model is not compared to other models? It would be nice to see how this model compares with other state-of-the-art models.

In Figure 7 and 8, you use the "raw" wording, is this R? I would suggest to be consistent with the annotation.

I did not understand what the red dashed lines in Figure 9 mean. Besides "metric value calculated from the training data", I could not find any other explanation. How did you compute that? Or, more importantly, what is this line representing?

When regarding the results, it is clear that the two-steps training is very useful. For that reason, it would be nice to clarify this idea already in the introduction (i.e.: where you explain for first time the refiner network) -- because, otherwise, one can think: why is this two step needed?

This work is interesting, novel and addresses a relevant issue. Interestingly, the problem this paper aims to address is not only relevant for the polyphonic music generation field, but also for the music transcription field -- since the sigmoidal outputs (generally used for deep transcription models) are a challenge since one desires (coherent and not over-segmented) binary outputs.

Finally, I'm curious to know if it is hard for the model to generate "zeros" (empty matrices denoting silence)?

Minor comments:

As a result, G does not need to "learn hard" to generate *A* realistic..

The latter was successfully employed in training *sophisticate networks* -> I suggest to change the wording.

but we like this design for it reflects the intuition -> I suggest to rephrase this sentence, it is not clear.

cleansed -> cleaned (section 4.1)

==========
REVIEWER #2
==========

Reviewer's Scores

```
            Scholarly/scientific quality: medium high
                                 Novelty: medium low
                       Relevance of topic: high
                              Importance: medium low
      Readability and paper organisation: high
                       Title and abstract: yes
                             Bibliography: yes
         Make Review Publicly Accessible: Yes
```

Comments

The paper proposes a solution to the problem of learning (to sample) binary data using GANs for polyphonic music generation. It is well written and well structured, and figures support understanding of the architecture and results. Audio examples are given, even though they are not very pleasing.

A problem I see, is that the impact of using the refiner network and binary units is not shown well in the evaluation, as the baseline model has a different architecture (i.e. no "private" networks). It is not clear, if the improvement is due to these private networks, or due to the stochastic units. That is a major problem, as the whole motivation is based on those binary units.
Furthermore, the use of binary units with the Hinton method is quite ad-hoc and strictly not correct (proposed rather as a regularization than as a method for probabilistic sampling).

Further remarks:

- Why is a refiner network necessary at all? Why cannot one sample directly from the output of the generator and use the unit probabilities for the backward pass? When pre-training improves results, then why not train the former architecture and then turn on sampling in the forward pass?

- The evaluation metrics seem to be a bit ad-hoc and not really musically plausible (except tonal distance).

- "In the backward pass, ST simply treats BNs as identify functions and ignores their gradients."
Is this right (if *identity* functions are meant)? I don't think, it ignores the gradients, but uses the probability which led to the binary input (i.e. forward -> sample, backward -> use probability).

==========
REVIEWER #3
==========

Reviewer's Scores

```
            Scholarly/scientific quality: high
                                 Novelty: medium low
                       Relevance of topic: high
                              Importance: medium high
      Readability and paper organisation: high
                       Title and abstract: yes
                             Bibliography: yes
         Make Review Publicly Accessible: Indifferent
```

Comments

The authors detail a modification to the previous state of the art for piano roll generation with GANs. They attempt to overcome the errors introduced with non-differentiable binary decisions for each note/frame by including the decision function explicitly in the network during training. The use a sigmoid adjusted straight-through estimator and compare several variations of deterministic and stochastic output neurons. They further find that adding a second "refinement" network is beneficial for performance, and that separately training it allows for a curriculum of training that improves stability and performance.

The experiments are well done, with ablations to justify each element, and both quantitative and qualitiative results. The paper is clearly written and cites

appropriate literature.

The impact of the paper is perhaps limited by the quality of the results, which still leave much to be desired, but are an improvement over the previous SOTA for discrete music GAN generation. Further examination and explanation of the limitations of these techniques and areas for future progress would help strengthen the paper.

==========
REVIEWER #4
==========

Reviewer's Scores

Scholarly/scientific quality	medium low
Novelty	medium low
Relevance of topic	medium high
Importance	medium low
Readability and paper organisation	medium low
Title and abstract	yes
Bibliography	yes
Make Review Publicly Accessible	Yes

Comments

This is a meta-review of the paper.

This paper describes an automatic music generation method based on a generative adversarial network (GAN) with binary neurons for directly generating a binary piano-roll representation. Since the gradients of the proposed network cannot be calculated, the sigmoid-adjusted straight-through (ST) estimator is used for the forward pass in the training phase. The proposed network consists of a generator that takes as input Gaussian noise and outputs a real-valued piano-roll representation, a refiner that converts the real-valued piano-roll representation into a binary piano-roll representation, and a discriminator that judges the genuineness of the binary piano-roll representation. The generator and discriminator are trained jointly, and the the refiner and discriminator are further trained jointly.

Pros:
This paper reports premature but interesting generation results. An idea of using binary neurons seems promising for both generation and transcription of music.

Cons:
The improvement of using binary neurons was not evaluated experimentally (the whole method was evaluated). Some claims were not justified and the paper writing needs to be improved.

Section 1:
I could not understand from Fig. 2 why the binary representation is easier to discriminate than the real-valued representation.

Section 3:
Please explain more carefully the data representation using a figure.

Why did you not train all the three subnetworks jointly? This should be theoretically discussed here, not in Section 4.5.

Section 4:
The proposed method was compared with GANs with Bernoulli sampling and hard thresholding. However, the implementation details of those methods are not described at all. So, I could not evaluate the comparative results.

Figure 7 makes no sense. I could not understand the difference between the five outputs.

Please show more clear figures with high resolution. The score representation would be better.