CONLON: A Pseudo-Song Generator Based on a New Pianoroll, Wasserstein Autoencoders, and Optimal Interpolations

Luca Angioloni[1], Tijn Borghuis[2,3], Lorenzo Brusci[3], and Paolo Frasconi[1]

[1] DINFO, Università di Firenze, Italy
first.last@unifi.it
[2] Eindhoven University of Technology, The Netherlands
v.a.j.borghuis@tue.nl
[3] Music-co, Eindhoven, The Netherlands
lorenzo.brusci@musi-co.com

Significant progress in algorithmic music generation has recently resulted from the widespread application of new and powerful methods based on deep generative models, letting this class of data-driven approaches gradually take over more traditional rule-based or probabilistic techniques. The musical quality of the results is still not always sufficient to enable a widespread adoption in realistic professional scenarios. The generation system we present in [1], introduces novelties across three dimensions: the type of data structures that are used to describe MIDI patterns, the nature of the generative learning models, and the strategy used to produce a whole musical piece, whose combination allows us to generate meaningful and professionally usable streams of music. We call our system CONLON, for Channeled Onset of Notes and Length Of Notes, and in honor of Conlon Nancarrow (1912–1997), a pioneer of piano roll compositions.

We introduce a novel pianoroll-like pattern description, PR^{C} , that stores velocities and durations in two separate channels. Our description does not suffer the ambiguity between long notes and repeated occurrences of the same note that is inherent in binary piano roll descriptions (PR). PR^{C} is completely lossless: a quantized MIDI pattern transformed into the corresponding PR^{C} tensor can be recovered exactly. Additionally, it can be perceptually more robust to reconstruction errors. A further advantage is that all the information about a note is local, whereas in the case of PR, a convolutional network requires a wide receptive field to infer the note duration.

As a generative model, we experiment with Wasserstein autoencoders (WAE) [4], a type of autoencoder that is less subject to the "blurriness" problem typically associated with variational autoencoders (VAE), which manifests itself in the case of music patterns as large clusters of notes being played together and sometimes in swarms of short notes that are never present in the training data. WAEs avoid this problem by pushing the expectation inside the divergence, i.e., penalizing a divergence $\mathcal{D}$ between the prior q_z and the *aggregated* posterior $q_z(z) = \mathbb{E}_p\, q(z|x)$, where p is the data distribution. They thus minimize, with respect to the parameters of the decoder, the quantity

$$\min_{q(z|x)} \mathbb{E}_p\, \mathbb{E}_{q(z|x)}\, c(x, G(z)) + \lambda \mathcal{D}(q_z, p_z) \tag{1}$$

where c is a reconstruction loss and λ a hyperparameter to be fixed. In all our experiments we employed the Maximum Mean Discrepancy (MMD) for $\mathcal{D}$ and a Gaussian prior for p_q, and we structured the encoder and the decoder as in the DCGAN [3] architecture.

Our generation strategy is similar to interpolation, where MIDI pseudo-songs are obtained by concatenating patterns decoded from smooth trajectories in the embedding space, but we formulate it as an optimization problem for exploring the autoencoder latent space in a way that prevents abrupt transitions between consecutively generated patterns, as well as regions with little variation. The optimal trajectories are computed as the solution of a widest-path problem.

We tested CONLON on three datasets. ASF-4 is a set of 910 patterns of four bars in three genres: *acid jazz*, *soul* and *funk*. Each pattern has 4 tracks associated with a simple electro-acoustic quartet: drums, bass, Rhodes piano, and Hammond organ. HP-10 is a set of 968 patterns of four bars in two genres: *high-pop* and *progressive trance*. Each pattern has 10 tracks associated with the following instrument set: drums, bass, Rhodes, brass-synth, choir, dark-pad, guitar, lead, pad, and strings. Both ASF-4 and HP-10 have been especially composed by two professional musicians for this study[4]. The third dataset was LPD-5 (cleansed version) derived from the Lakh MIDI dataset by Dong *et al.* [2].

To validate the CONLON approach, we conducted three listening experiments with a group of 69 musicians. These experiments showed that musicians find pseudo-songs generated with WAEs and PR^{C} descriptions more useable in music production than pseudo-songs generated with the MuseGAN model [2] and PR descriptions, find pseudo-songs generated by WAEs with PR^{C} descriptions more useable than pseudo-songs generated by the same WAE with PR descriptions, and find the development over time of pseudo-songs generated with WAEs and PR^{C} description coherent rather than incoherent (with respect to Harmony, Rhythm, Melody, and Interplay of instruments).

References

1. Angioloni, L., Borghuis, T., Brusci, L., Frasconi, P.: Conlon: A pseudo-song generator based on a new pianoroll, wasserstein autoencoders, and optimal interpolations. In: Proceedings of the 21th International Society for Music Information Retrieval Conference ISMIR MTL2020. pp. 876–883 (2020)
2. Dong, H., Hsiao, W., Yang, L., Yang, Y.: Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. pp. 34–41 (2018)
3. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: 4th International Conference on Learning Representations (2016)
4. Tolstikhin, I.O., Bousquet, O., Gelly, S., Schölkopf, B.: Wasserstein auto-encoders. In: 6th International Conference on Learning Representations (2018)

[4] These datasets, along with other additional materials are available at https://paolo-f.github.io/CONLON/.