Harmonic Recomposition using Conditional Autoregressive Modeling

Kyle Kastner [1] **Rithesh Kumar** [1] **Tim Cooijmans** [1] **Aaron Courville** [1 2]

Abstract

We demonstrate a conditional autoregressive pipeline for efficient music recomposition, based on methods presented in van den Oord et al. (2017). Recomposition (Casal & Casey, 2010) focuses on reworking existing musical pieces, adhering to structure at a high level while also reimagining other aspects of the work. This can involve reuse of pre-existing themes or parts of the original piece, while also requiring the flexibility to generate new content at different levels of granularity. Applying the aforementioned modeling pipeline to recomposition, we show diverse and structured generation conditioned on chord sequence annotations.

1. Introduction

Since the early days of computation, composers have explored methods of combining aleatoric music and algorithmic composition with generic computing devices (Agon et al., 2003; Cope, 1989). Authors have taken a wide variety of data driven approaches to "creative generation" in various domains (Barbieri et al., 2012; Ha & Eck, 2017; Graves, 2013), with extensive application to music modeling (Briot et al., 2017; Roberts et al., 2018; Eck & Schmidhuber, 2002; Sturm et al., 2015; Hadjeres et al., 2016; Boulanger-Lewandowski et al., 2012; Bretan et al., 2017; Roberts et al., 2018).

In this paper, we focus on the task of *harmonic recomposition* (Casal & Casey, 2010). Melody generation and evaluation is a difficult task, even in monophonic music (Jaques et al., 2016), so we use the term *harmonic* recomposition to reference our focus on aspects of agreement and structure between voices. Our pipeline is also applicable to purely sequential and iterative generation, as has been shown in prior work (Huang et al., 2017; van den Oord et al., 2017).

[1]Montréal Institute for Learning Algorithms (MILA), Université de Montréal, Montréal, Québec, Canada [2]CIFAR Fellow. Correspondence to: Kyle Kastner <kyle.kastner@umontreal.ca>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

1.1. Related Work

Autoregressive models have proven to be powerful distribution estimators for images and sequence data, showing excellent results in generative settings (van den Oord et al., 2016a). They have also performed well in related prior work for polyphonic music generation (Briot et al., 2017). Most related to the work described in this paper is CoCoNet (Huang et al., 2017; 2018), which also uses an autoregressive convolutional model over image-like structures for polyphonic music generation and was a direct inspiration for our approach. One key difference of our approach is our utilization of a two stage pipeline (first seen in the work of van den Oord et al. (2017)) which greatly improves training and generation speed as well as creating an implicit separation between local voice agreement (first stage) and global consistency over measures (second stage).

2. Implementation Details

In this section, we describe the data, model, and training details for our recomposition approach. An open source implementation of our setup (including audio samples) is available online[1].

2.1. Data

We use a subset of the scores associated with the composer Josquin des Prez as compiled by the Josquin project [2]. Only pieces with 4 parts are considered, resulting in a dataset with 103 pieces comprising 5568 measures. We hold a contiguous 500 measures out for use as a source of harmonic chord sequences during conditional generation.

After extracting individual measures, we convert to a "piano roll" style multichannel image, with each measure having 16 quantized timesteps (regardless of time-signature) on the horizontal axis, and one of 49 possible tones on the vertical axis, where 49 comes from the set of all possible notes used in the key normalized data (Hadjeres et al., 2016). These 49 values are padded to 52 for compatibility with the convolutional strided layers used in the VQ-VAE, and each voice assigned its own channel in an image-like container of

[1]https://github.com/kastnerkyle/harmonic_recomposition_workshop

[2]http://josquin.stanford.edu/

size $(N, 52, 16, 4)$ described in examples, height, width, and channels format (NHWC). The overall result can be seen in Fig. 1, where each color represents a separate channel.

2.2. Conditional Information

We extract the chord function and voicing of all measures using the music21 software package (Cuthbert & Ariza, 2010), and form "function triplets" of the previous, current, and next measure. A measure group of I, ii_2^4, V_5^6 chords would form triplet groupings of I, I, ii_2^4 (we repeat chords to handle border issues), then I, ii_2^4, V_5^6, and finally ii_2^4, V_5^6, V_5^6.

2.3. Models

The model pipeline is a two-stage generative setup, as described by van den Oord et al. (2017), wherein an initial stage (denoted VQ-VAE) is unconditionally trained to compress inputs to a spatially reduced, discrete representation (which we call Z), and uncompress. Once the VQ-VAE stage is trained, we use it to generate a compressed Z for each element in the dataset, and train an autoregressive generative "prior model" on this representation. The prior model learns to generate components of Z (which takes the form of a $(13, 4)$ spatial map in this work), denoted $Z_{i,j}$, one at a time conditioning the next generation step on all previously generated $Z_{<i,<j}$.

The prior model may also take conditioning as one or multiple vectors (separate embeddings for each previous, current, and next chord, indexed by a chord integer), a spatial map (a Z from some previous measure), or a combination of both during the generation process. The effect of conditioning type can be seen in Fig 2.

2.4. Experiment Details

The first stage VQ-VAE has 2 strided convolutional layers of kernel size $(4, 4)$ and a stride of 2 on both spatial axes, followed by an additional 2 layers of $(4, 4)$ convolution with stride 1, using rectified linear activations (Glorot et al., 2011) and batch normalization (Ioffe & Szegedy, 2015). These layers have sizes 64, 128, 257, and 256 leading to a VQ codebook size of 256, which results in a latent Z dimension of $(13, 4)$ for size $(52, 16, 4)$ input. This procedure is inverted using transpose convolution for the decoder and combined with a binary cross-entropy loss alongside codebook and commitment losses for the VQ-VAE, averaged over all channels and spatial dimensions. Training was performed over 50000 minibatches of size 64 with an Adam optimizer with $\alpha = 0.0002$, $\beta 1 = 0.9$, $\beta 2 = 0.999$, $\epsilon = 1E - 8$ (Kingma & Ba, 2014).

In the second stage, a gated conditional PixelCNN (van den Oord et al., 2016b) is configured with 15 layers of 64 projection channels. The first layer has a kernel size of $(5, 5)$ and no residual connection. Layers after the first utilize residual connections (He et al., 2016) and have a kernel size of $(3, 3)$, followed by $(1, 1)$ convolution, rectified linear activation, and a final convolution of size $(1, 1)$ and 256 channels (due to the VQ codebook size of 256). Training was performed over 100000 minibatches of size 50 with an Adam optimizer (configured as before) with categorical cross-entropy loss averaged over the output.

3. Results

We experiment with two types of conditioning combined with the aforementioned architecture. The higher level information contained in the chord sequences alone seems sufficient to produce directed, coherent trajectories, without need for spatial conditioning information. When spatial conditioning from the previous timestep is included, the resulting generations are punctuated by dissonant intervals or long silent gaps. Finding better ways to combine local note level information with chord annotation will be an important step to improving this pipeline.

Figure 1. Two 8 bar sequences generated over the conditioning sequence (III),III, v, iv, iv, v, i, iv, III, (III) using different random seeds

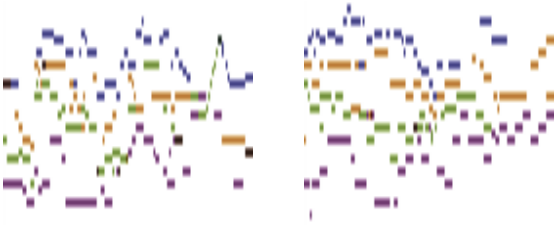

Figure 2. 8 bar sequences generated over conditioning sequence (ii),I,V,vi,ii,V,ii,ii,$i5$,(I). *Left*: only chord triplet conditioning. *Right*: chord triplet conditioning along with previous generation as conditioning spatial map.

4. Conclusion

Chord-conditional generative models are an ideal fit for harmonic recomposition. We find that a two-stage pipeline reminiscent of van den Oord et al. (2017) and Huang et al. (2017) captures musical structure, and allows for chord conditional generation. Our work demonstrates note-level realizations of given chordal sequences and provides an open-source implementation with examples.

References

Agon, Carlos, Andreatta, Moreno, Assayag, Gérard, and Schaub, Stephan. Formal aspects of Iannis Xenakis' Symbolic Music: a computer-aided exploration of some compositional processes. *Journal of New Music Research*, 2003. cote interne IRCAM: Agon03a.

Barbieri, Gabriele, Pachet, François, Roy, Pierre, and Esposti, Mirko Degli. Markov constraints for generating lyrics with style. In *Proceedings of the 20th European Conference on Artificial Intelligence*, ECAI'12, 2012.

Boulanger-Lewandowski, Nicolas, Bengio, Yoshua, and Vincent, Pascal. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. *arXiv preprint arXiv:1206.6392*, 2012.

Bretan, Mason, Oore, Sageev, Engel, Jesse, Eck, Douglas, and Heck, Larry P. Deep music: Towards musical dialogue. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence,*, 2017.

Briot, Jean-Pierre, Hadjeres, Gaëtan, and Pachet, François. Deep learning techniques for music generation - A survey. *CoRR*, abs/1709.01620, 2017.

Casal, David Plans and Casey, Michael. Decomposing autumn: A component-wise recomposition. In *ICMC*, 2010.

Cope, David. Experiments in musical intelligence (emi): Nonlinear linguisticbased composition. *Interface*, 1989.

Cuthbert, Michael Scott and Ariza, Christopher. Music21: A toolkit for computer-aided musicology and symbolic music data. In *Proceedings of the 11th International Society for Music Information Retrieval Conference*, 2010.

Eck, Douglas and Schmidhuber, Juergen. Learning the long-term structure of the blues. In Dorronsoro, J. (ed.), *Artificial Neural Networks – ICANN 2002 (Proceedings)*, 2002.

Glorot, Xavier, Bordes, Antoine, and Bengio, Yoshua. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 315–323, 2011.

Graves, A. Generating sequences with recurrent neural networks. *arXiv:*`1308.0850 [cs.NE]`, August 2013.

Ha, David and Eck, Douglas. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*, 2017.

Hadjeres, Gaëtan, Pachet, François, and Nielsen, Frank. Deepbach: a steerable model for bach chorales generation. *arXiv preprint arXiv:1612.01010*, 2016.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016.

Huang, Anna, Chen, Sherol, Nelson, Mark, and Eck, Douglas. Towards mixed-initiative generation of multichannel sequential structure. 2018.

Huang, Cheng-Zhi Anna, Cooijmans, Tim, Roberts, Adam, Courville, Aaron, and Eck, Douglas. Counterpoint by convolution. In *Proceedings of ISMIR 2017*, 2017.

Ioffe, Sergey and Szegedy, Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, 2015.

Jaques, Natasha, Gu, Shixiang, Bahdanau, Dzmitry, Hernández-Lobato, José Miguel, Turner, Richard E, and Eck, Douglas. Sequence tutor: Conservative fine-tuning of sequence generation models with kl-control. *arXiv preprint arXiv:1611.02796*, 2016.

Kingma, Diederik P and Ba, Jimmy. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Raffel, Colin and Ellis, Daniel PW. Large-scale content-based matching of midi and audio files. In *ISMIR*, pp. 234–240, 2015.

Roberts, Adam, Engel, Jesse, Raffel, Colin, Hawthorne, Curtis, and Eck, Douglas. A hierarchical latent vector model for learning long-term structure in music. *CoRR*, abs/1803.05428, 2018.

Sturm, Bob, Santos, João Felipe, and Korshunova, Iryna. Folk music style modelling by recurrent neural networks with long short term memory units. In *16th International Society for Music Information Retrieval Conference, late-breaking demo session*, pp. 2, 2015.

van den Oord, Aaron, Dieleman, Sander, Zen, Heiga, Simonyan, Karen, Vinyals, Oriol, Graves, Alex, Kalchbrenner, Nal, Senior, Andrew, and Kavukcuoglu, Koray. Wavenet: A generative model for raw audio. 2016a.

van den Oord, Aaron, Kalchbrenner, Nal, Espeholt, Lasse, kavukcuoglu, koray, Vinyals, Oriol, and Graves, Alex. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems 29*. 2016b.

van den Oord, Aaron, Vinyals, Oriol, and kavukcuoglu, koray. Neural discrete representation learning. In *Advances in Neural Information Processing Systems 30*, pp. 6306–6315. 2017.