Coupled Recurrent Models for Polyphonic Music Composition

John Thickstun, Zaid Harchaoui, Dean P. Foster and Sham M. Kakade

Learning a Distribution over Scores

Train a model to compose music by estimating a distribution p_θ using scores from a dataset of compositions $\mathcal{D}$:

$$\max_\theta \sum_{\mathcal{S}\in\mathcal{D}} p_\theta(\mathcal{S}), \text{ where } p_\theta(\mathcal{S}) = \prod_{i=1}^{m} p_{\theta,i}(\mathcal{S}_i|\mathcal{S}_{<i}).$$

- How to order the content of a score $\mathcal{S}_1,\ldots,\mathcal{S}_m$?
- How to featurize the history $\mathbf{e}\equiv\mathcal{S}_{<i}$?
- How to parameterize the conditional distributions $p_{\theta,i}$?

Ordering the Content of a Score

Many variants, at least two main approaches:

- **raster**: discretize a score $\mathcal{S}$ into fine time-slices. Order these slices temporally, and factor the distribution over slices.
- **note-based**: assign an order to notes in a score (e.g. temporally, based on the time when the note begins) and factor the distribution over notes.

We take the note-based approach.

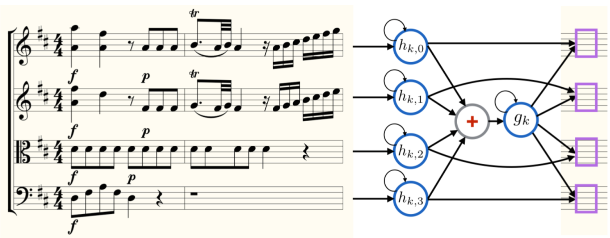
Featurizing the History

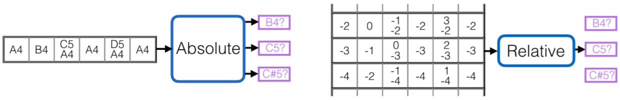
Again many variants, at least three high-level approaches:

- **raster**: can exploit pitch-domain structure via convolution, but requires a large history tensor.
- **note-based**: compact history tensor (list of notes) but cannot easily exploit structure.
- **run-length encoding of raster**: can exploit pitch-domain structure with a compact history tensor.

We use run-length encoding of the history.

Parameterizing the Conditional Distributions With Coupled Voice Models





We build a recurrent estimate $h_{k,v}$ of the state of each voice v at index k. We couple these estimates to construct a global estimate g_k of the state of the full score at position k:

$$h_{k,v}(\mathbf{e}) \equiv \mathbf{a}\left(W_v^\top h_{k-1,v}(\mathbf{e}) + W_e^\top \mathbf{c}\left(\mathbf{e}_{k,v}\right)\right),$$

$$g_k(\mathbf{e}) \equiv \mathbf{a}\left(W_g^\top g_{k-1}(\mathbf{e}) + W_{hv}^\top \sum_u h_{k,u}(\mathbf{e})\right).$$

We relativize the pitch predictor: instead of building an m-way classifier for each of the m possible pitch classes, we build a single classifier that sees a shifted view of the history tensor $\mathbf{e}$.

A Computer-Generated Score

Qualitative Results (user study)

Clip Length	10	20	30	40	50
Average	5.3	5.7	6.6	6.7	6.8

- Can listeners tell the difference between clips of computer-generated scores and human compositions?
- How does the length of the clip affect listeners' ability to discriminate?
- An average of 5.0 indicates random guessing: no ability to discriminate human compositions from the computer.

Quantitative Results (log-loss)

#	History (voice/global)	Architecture	Loss (total)	Loss_t (time)	Loss_n (notes)
1	3 / 3	hierarchical	14.05	5.65	8.40
2	5 / 5	hierarchical	13.40	5.35	8.04
3	5	distributed	13.82	5.41	8.41
4	10 / 1	hierarchical	13.20	5.22	7.98
5	10 / 5	hierarchical	12.94	5.13	7.81
6	10 / 10	hierarchical	12.87	5.12	7.75
7	20 / 20	hierarchical	12.78	5.01	7.76
8	10	independent	18.63	6.56	12.08

http://homes.cs.washington.edu/~thickstn/ismir2019composition/