Notes on the Music Transformer

Aditya Gomatam
March 21, 2021

1 Introduction

The Music Transformer or Transformer with Relative Self-Attention[1] is an autoregressive sequence model that builds on the Transformer[2] to consider the relative distances between different elements of the sequence rather than / along with their absolute positions in the sequence. This consideration was designed to model the fact that music relies heavily on repetition to construct meaning and maintain long-term structure. The Music Transformer, from my experience with RNNs, appears to be the first neural network capable of reasonably capturing any sort of meaning to a piece. Google achieved extremely compelling results that truly demonstrate this capability of the Transformer architecture with their Piano Transformer. This essay presents my thoughts on why the Self-Attention and Relative Self-Attention algorithms are so effective.

2 Relative Self-Attention

References

[1] C. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, C. Hawthorne, A. M. Dai, M. D. Hoffman, and D. Eck, "Music Transformer: Generating music with long-term structure," *arXiv preprint arXiv:1809.04281*, 2018.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *arXiv preprint arXiv:1706.03762*, 2017.