Modeling Expressive Musical Performance with Transformers: An Empirical Error Study

Richard W Timpson
University of Utah

UUCS-21-008

School of Computing
University of Utah
Salt Lake City, UT 84112 USA

26 April 2021

Abstract

Current state-of-the-art modeling of expressive musical performance (EMP) is based on hierarchical Recurrent Neural Networks (RNN). The Transformer is a recent Neural Network (NN) sequence modeling architecture that has led to significant research improvement in Natural Language Processing (NLP) and other related fields. To date, there has been no application of the Transformer in music performance modeling – we present the first study that attempts to do so. The results indicate that our encoder-only Transformer model outperforms a similar encoder-only RNN but does not outperform the existing hierarchical RNN state-of-the-art. However, an analysis of current evaluation methods for EMP generation models reveals that the current metrics for “objective” quantitative evaluation do not correctly capture the essence of musical performance and are significant bottlenecks in developing robust EMP models. Therefore, we cannot draw any meaningful conclusions about our models’ performance when evaluating them quantitatively. We present an error analysis of the current evaluation methods and provide suggestions for future efforts in building better models and finding better evaluation metrics.