Small Language Modelling to Generate Playlists from Song Sequences

Mark Green
Data Science Program
Indiana University
Bloomington, IN 47405
margree@iu.edu

Sahil Dhingra
Data Science Program
Indiana University
Bloomington, IN 47405
sahdhin@iu.edu

Rahul Jain
Data Science Program
Indiana University
Bloomington, IN 47405
rjdharmc@iu.edu

Abstract

Large Language Models (LLMs) such as GPT and BERT have led to revolutionary advances in language applications such as machine translation, sentiment analysis, and text generation. However, these models rely on enormous datasets of texts derived from spoken language which are prohibitively resource-intensive to train. Although pre-trained models solve this to a degree, their scope is limited to the pre-trained spoken language so they cannot be applied to more niche language modeling tasks. This paper explores the application of neural networks for small language modelling where the dataset is small in both size and vocabulary and is composed of non-spoken language "words". Specifically, this approach explores the music of the band Phish, leveraging historical song appearances and setlists from their shows to model the band's performance tendencies. By decoding intricate patterns from their extensive performance history using a targeted small language model, this approach offers a unique solution for personalized music recommendations. The research aims to enhance the user experience for dedicated fans with more artist-centric music recommendation systems and to test the performance of neural network language models for small, niche applications.

1 Introduction

As of 2022, music streaming is an $18 Billion dollar industry[1]. Playlist generation is a key differentiator for streaming platforms and previous works have used reinforcement learning [1], convolutional neural networks (CNNs) [2], and recurrent neural networks (RNNs) [3] as the generative mechanisms. These works have generally spanned a broad scope of artists and genres to generate playlists composed of multiple bands or band-specific playlists ordered on an arbitrary learned feature. While this may work well for a generic music listener, this approach may not be effective for "dedicated fans" - a subset of users whose listening habits tend to be more focused around an individual band and their compositional tendencies.

One such type of "dedicated fan" are those followers of Jam bands. Since the 1960s, Jam bands have captivated audiences with their improvisational Rock and Roll style, reminiscent of Jazz. Notably, these groups do not pre-plan setlists for their live performances and instead weave together a pseudo-random selection of songs such that no two shows are alike. The choice of songs for each show is partially influenced by spontaneous notions such as crowd energy or a band member's mood, but it is also influenced by constraints like temporal frequency and positional sequencing. These shows also generally follow a standard pattern composed of two sets separated by an intermission and a brief encore composed of just a song or two. Beyond the music itself, the sequencing of songs in

[1] https://www.weforum.org/agenda/2023/03/charted-the-impact-of-streaming-on-the-music-industry/

Preprint. Under review.

shows is reminiscent of natural language – song names are analogous to words, setlists to sentences, show starts and stops to parts-of-speech – with the repertoire of each band forming its own unique language full of syntactical intricacies.

We purport that constructing small language models from the song sequences of live performances to generate playlists for a given band could increase the engagement of "dedicated fan" type users by generating playlists that more closely mimic a band's musical tendencies than a more generalized algorithm. In this report, we explore the music of the Jam band Phish for this purpose – who has not only a large catalog of live shows and songs, but also a very dedicated fanbase. A previous endeavor[2] employed a Long Short-Term Memory (LSTM) model with Word2Vec [4] algorithm for token embeddings to achieve a 21% accuracy for next song prediction given a sequence of Phish songs. This paper builds upon this work by recreating it four years later and exploring the prediction task with alternate embedding methods, and RNN model architectures, and attention model architectures.

2 Related Work

the work[3]

We attempted to reproduce the original experiment from Nov 2019 with latest dataset (from Dec 2023). The best model (with 2 LSTM layers) identified by the author was selected and performed a grid search with 50 and 150 sequence length. None of the models from grid search performed well with the Dec 2023 dataset. The validation loss after a few epochs vanished which indicates the model was cursed with gradient exploding problem. Another interesting observation was that the model would saturate in a few epochs even with different LSTM units and dropout rates. This indicates that the model used in original experiment was overfitting the data then hence not able to generalize when tried with latest data. The highest validation accuracy observed was approx. 12% which is less than the originally reported accuracy of 21%.

In summary, the lesson learnt was just the sequence of songs was not enough to predict the song sequences, how more context and data is needed to obtain meaningful inference.

3 Data Preparation

The `https://phish.net` fan forum curates an API[4] cataloguing every setlist of songs for every show the band has played. First queried from this endpoint were song-related details encompassing song name, ID, artist, play frequency, last performance date, and debut date. Further, we meticulously filter shows attributed exclusively to Phish, forming a chronological compendium of their extensive concert history. Next queried were setlist data such as showdate, set number, song positions, ID, name, transition markers, song gaps since the last performance, and jam categorization. Incomplete information were eliminated to ensure the dataset's integrity, resulting in a comprehensive dataframe. This curated dataset, exclusively featuring "full" shows with two sets and an encore, serves as the foundation for the predictive modeling task and is saved as a csv file[5] in the project's code repository.

In the initial exploratory data analysis of the Phish concert dataset, it was revealed the frequency of songs played resembled a power-law distribution whereby almost half of the songs were only played once or twice throughout the band's entire history. To address the challenge posed by rarely performed songs, we consolidated these one-off and two-off songs into a unified song category termed the "wildcard". For prediction, this wildcard song is treated like any other song Phish plays but it represents an intractable uncertainty to relax the problem for new song debuts or unpredictably random cover songs.

Furthermore, for predictive modeling, we systematically aggregated songs into sequences by grouping them based on show date and set number. These sequences were then concatenated into a unified string with additional separators in the string to indicate a song's position in the show. For example, the beginning of every string starts with 'set-1'. Then after all the 1st set songs appear in the string, the

[2] `https://towardsdatascience.com/predicting-what-song-phish-will-play-next-with-deep-learning-947ccce3824`

[3] `https://github.com/andrewrreed/phish-setlist-modeling`

[4] `https://docs.phish.net/`

[5] `https://github.iu.edu/rjdharmc/dlProject/blob/main/data/allphishsets.csv`

'set-2' separator appears, then the 'set-e' separator before the encore, and finally the 'eos' separator at the end of the string after all the songs.

These preprocessing steps refined the dataset and established a structured framework for developing models focused on predicting Phish setlists. Following this modification and the original data cleaning, the dataset is composed of 1,550 shows, 33,533 total songs played, with 482 unique songs and separators.

3.1 Interpretting Song Sequences as Language

In order to mathematically interpret the song names as a language, it was necessary to tokenize the song names and separators into numerical values suitable for input into a neural network model. This was predominantly done using unigram tokens, or integer representations mapped to each full song name. In this way, each setlist is similarly decomposed into a list of integer values.

Recall, the aim of this project is to generate song sequences (IE playlists) which are representative of the style in which the band plays. This problem can be framed in terms of language modelling in either a *sequence-to-token* perspective, using a sequence to predict the next token, or a *sequence-to-sequence* perspective, using a sequence to predict the next sequence. Since the models will have to either predict the next n^{th} song or the next n songs, multiple data splitting techniques for input and output data were applied to evaluate the model performance under both of these paradigms. These are described below.

Method 1 The *sequential split* captures the immediate temporal dependencies within the songs. This approach involves a sequential splitting of the song data, where each setlist sequence is transformed into input and target sequences for training. The sequences are then padded to ensure uniform length. This method enables the model to learn and predict the next song in a sequence, essentially forecasting the playlist evolution

Method 2 The *time-based split* considers the evolving nature of song sequences over time. In this approach, the dataset is chronologically split into training and testing sets based on a specified date, for example '2015-12-31'. This method reflects a time-based split, capturing the temporal dynamics of song sequences.

Method 3 The *unified encoding approach* offers insights into the global patterns and relationships across the entire song repertoire. In this approach, the training data is like [n], [n, n+1], [n, n+1, n+2] and training labels are [n+1], [n+2], [n+3] and so on. Basically, the next song to the incoming sequence is our prediction label.

4 Models

Since we are dealing with the sequential aspect of the data, we are using the RNN LSTM model so that model can retain the sequential information for prediction. For experiments, we used both Uni and Bi-directional models.

4.1 Unidirectional LSTM with N-Dimensional Embedding

The initial model architectures includes an Embedding layer to represent songs numerically, two Long Short-Term Memory (LSTM) layers for capturing sequential dependencies, dropout for regularization, and a Dense layer with softmax activation for multi-class classification. The model is trained on variable-length input sequences of songs, and the Embedding layer allows it to learn meaningful representations for each song. The use of masking ensures proper handling of sequences with varying lengths. This RNN architecture aims to effectively capture the temporal patterns in the dataset, enabling it to make accurate predictions about the next song in each sequence.

4.2 Bidirectional LSTM with N-grams and Adjacency Matrix Embedding

An N-gram sequencing approach was attempted to focus the LSTM model on the long-term sequencing of songs. Although similar to the previous sequence-to-token approach, the N-gram tokenization

Table 1: Sample table title

	Part	
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

splits a given setlist into as many vectors as it contains songs and set markers. Each of these vectors is composed of a padded sequence of 0s and an increasing number of songs and set markers from the set list. For example: the first vector is composed of just "set-1", the next also contains the 1st song, and the next also contains the first and second song – etc. In this way, each sequence is composed of N tokens (the N-grams) which are used to predict the next target token.

In addition to using the N-gram pre-processing technique, this method embeds the tokens using their co-occurrence matrix. This co-occurrence matrix is given by the LaPlace-smoothed adjacency matrix of the directed graph created by overlaying all the setlists as walks. In this graph, each song is a node, and each song transition contributes to a directed weighted edge.

Bi-directional LSTM is also employed – although the setlists are unidirectional sequences, this methodology looks for patterns more wholistically. A feed forward layer is used to compute the final output, with softmax activation.

4.3 Self-Attention Model with N-d + Positional Embedding

In contrast to the sequence-to-token models attempted with the LSTM architectures, sequence-to-sequence models are attempted with multi-head attention architectures. These models process the song sequence inputs and targets into offset pairs, where the original sequence is broken down into 1..N-1 tokens for the input sequence and 2..N tokens for the output sequence. The token sequences are encoded into an embedding layer composed of the sum of a learned N-dimensional vector embedding layer and a positional embedding layer. The positional embedding layer applies sine and cosine functions for fixed positional encodings.

The model architecture itself is similar to the decoder-only transformer used in the GPT-2 model. It consists of a causal self-attention layer using masked multi-head attention to attend to the sequencing. This layer is added and normalized, then passed to a feed forward layer to form the Transformer block. These transformer blocks are then layered sequentially and an additional feed forward layer computes the final output, then activated by softmax.

4.4 Self-Attention + Cross-Attention Model with N-d + Positional Embedding

This final model is setup similarly to the original transformer model proposed in "Attention is all you need". The sequence of songs concatenated from the previous 5 setlists are used as the context for the global self-attention layer, which is passed to the decoder via the cross-attention layer.

5 Results

Summary of the Results.

6 Discussion

Discussion of what was accomplished.

7 Conclusion

Some concluding words.

The code we used to acquire and process the dataset, train and evaluate our models is available at `https://github.iu.edu/rjdharmc/dlProject`. Note that access to this codebase will require an active account on the IU Enterprise Github instance.

References

[1] F. Tomasi, J. Cauteruccio, S. Kanoria, K. Ciosek, M. Rinaldi, and Z. Dai, "Automatic Music Playlist Generation via Simulation-based Reinforcement Learning," Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, Aug. 04, 2023. doi: 10.1145/3580305.3599777.

[2] R. T. Irene, C. Borrelli, M. Zanoni, M. Buccoli, and A. Sarti, "Automatic playlist generation using Convolutional Neural Networks and Recurrent Neural Networks," 2019 27th European Signal Processing Conference (EUSIPCO). IEEE, Sep. 2019. doi: 10.23919/eusipco.2019.8903002.

[3] K. Choi, G. Fazekas, and M. Sandler, "Towards Playlist Generation Algorithms Using RNNs Trained on Within-Track Transitions." arXiv, 2016. doi: 10.48550/ARXIV.1606.02096.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space." arXiv, 2013. doi: 10.48550/ARXIV.1301.3781.

8 Submission of papers to NeurIPS 2023

Please read the instructions below carefully and follow them faithfully. **Important:** This year the checklist will be submitted separately from the main paper in OpenReview, please review it well ahead of the submission deadline: `https://neurips.cc/public/guides/PaperChecklist`.

8.1 Style

Papers to be submitted to NeurIPS 2023 must be prepared according to the instructions presented here. Papers may only be up to **nine** pages long, including figures. Additional pages *containing only acknowledgments and references* are allowed. Papers that exceed the page limit will not be reviewed, or in any other way considered for presentation at the conference.

The margins in 2023 are the same as those in previous years.

Authors are required to use the NeurIPS LaTeX style files obtainable at the NeurIPS website as indicated below. Please make sure you use the current files and not previous versions. Tweaking the style files may be grounds for rejection.

8.2 Retrieval of style files

The style files for NeurIPS and other conference information are available on the website at

`http://www.neurips.cc/`

The file `neurips_2023.pdf` contains these instructions and illustrates the various formatting requirements your NeurIPS paper must satisfy.

The only supported style file for NeurIPS 2023 is `neurips_2023.sty`, rewritten for LaTeX 2_ε. **Previous style files for LaTeX 2.09, Microsoft Word, and RTF are no longer supported!**

The LaTeX style file contains three optional arguments: `final`, which creates a camera-ready copy, `preprint`, which creates a preprint for submission to, e.g., arXiv, and `nonatbib`, which will not load the `natbib` package for you in case of package clash.

Preprint option If you wish to post a preprint of your work online, e.g., on arXiv, using the NeurIPS style, please use the `preprint` option. This will create a nonanonymized version of your work with the text "Preprint. Work in progress." in the footer. This version may be distributed as you see fit, as long as you do not say which conference it was submitted to. Please **do not** use the `final` option, which should **only** be used for papers accepted to NeurIPS.

At submission time, please omit the `final` and `preprint` options. This will anonymize your submission and add line numbers to aid review. Please do *not* refer to these line numbers in your paper as they will be removed during generation of camera-ready copies.

The file `neurips_2023.tex` may be used as a "shell" for writing your paper. All you have to do is replace the author, title, abstract, and text of the paper with your own.

The formatting instructions contained in these style files are summarized in Sections 9, 10, and 11 below.

9 General formatting instructions

The text must be confined within a rectangle 5.5 inches (33 picas) wide and 9 inches (54 picas) long. The left margin is 1.5 inch (9 picas). Use 10 point type with a vertical spacing (leading) of 11 points. Times New Roman is the preferred typeface throughout, and will be selected for you by default. Paragraphs are separated by ½ line space (5.5 points), with no indentation.

The paper title should be 17 point, initial caps/lower case, bold, centered between two horizontal rules. The top rule should be 4 points thick and the bottom rule should be 1 point thick. Allow ¼ inch space above and below the title to rules. All pages should start at 1 inch (6 picas) from the top of the page.

For the final version, authors' names are set in boldface, and each name is centered above the corresponding address. The lead author's name is to be listed first (left-most), and the co-authors' names (if different address) are set to follow. If there is only one co-author, list both author and co-author side by side.

Please pay special attention to the instructions in Section 11 regarding figures, tables, acknowledgments, and references.

10 Headings: first level

All headings should be lower case (except for first word and proper nouns), flush left, and bold.

First-level headings should be in 12-point type.

10.1 Headings: second level

Second-level headings should be in 10-point type.

10.1.1 Headings: third level

Third-level headings should be in 10-point type.

Paragraphs There is also a `\paragraph` command available, which sets the heading in bold, flush left, and inline with the text, with the heading followed by 1 em of space.

11 Citations, figures, tables, references

These instructions apply to everyone.

11.1 Citations within the text

The `natbib` package will be loaded for you by default. Citations may be author/year or numeric, as long as you maintain internal consistency. As to the format of the references themselves, any style is acceptable as long as it is used consistently.

The documentation for `natbib` may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

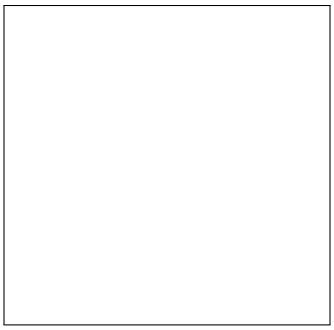

Figure 1: Sample figure caption.

Of note is the command `\citet`, which produces citations appropriate for use in inline text. For example,

```
\citet{hasselmo} investigated\dots
```

produces

> Hasselmo, et al. (1995) investigated…

If you wish to load the `natbib` package with options, you may add the following before loading the `neurips_2023` package:

```
\PassOptionsToPackage{options}{natbib}
```

If `natbib` clashes with another package you load, you can add the optional argument `nonatbib` when loading the style file:

```
\usepackage[nonatbib]{neurips_2023}
```

As submission is double blind, refer to your own published work in the third person. That is, use "In the previous work of Jones et al. [4]," not "In our previous work [4]." If you cite your other papers that are not widely available (e.g., a journal paper under review), use anonymous author names in the citation, e.g., an author of the form "A. Anonymous" and include a copy of the anonymized paper in the supplementary material.

11.2 Footnotes

Footnotes should be used sparingly. If you do require a footnote, indicate footnotes with a number[6] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).

Note that footnotes are properly typeset *after* punctuation marks.[7]

11.3 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction. The figure number and caption always appear after the figure. Place one line space before the figure caption and one line space after the figure. The figure caption should be lower case (except for first word and proper nouns); figures are numbered consecutively.

You may use color figures. However, it is best for the figure captions and the paper body to be legible if the paper is printed in either black/white or in color.

[6] Sample of the first footnote.

[7] As in this example.

Table 2: Sample table title

	Part	
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

11.4 Tables

All tables must be centered, neat, clean and legible. The table number and title always appear before the table. See Table 2.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

Note that publication-quality tables *do not contain vertical rules.* We strongly suggest the use of the `booktabs` package, which allows for typesetting high-quality, professional tables:

`https://www.ctan.org/pkg/booktabs`

This package was used to typeset Table 2.

11.5 Math

Note that display math in bare TeX commands will not create correct line numbers for submission. Please use LaTeX (or AMSTeX) commands for unnumbered display math. (You really shouldn't be using $$ anyway; see `https://tex.stackexchange.com/questions/503/why-is-preferable-to` and `https://tex.stackexchange.com/questions/40492/what-are-the-differences-between-align-equation-and-displaymath` for more information.)

11.6 Final instructions

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the **References** section; see below). Please note that pages should be numbered.

12 Preparing PDF files

Please prepare submission files with paper size "US Letter," and not, for example, "A4."

Fonts were the main cause of problems in the past years. Your PDF file must only contain Type 1 or Embedded TrueType fonts. Here are a few instructions to achieve this.

- You should directly generate PDF files using `pdflatex`.
- You can check which fonts a PDF files uses. In Acrobat Reader, select the menu Files>Document Properties>Fonts and select Show All Fonts. You can also use the program `pdffonts` which comes with `xpdf` and is available out-of-the-box on most Linux machines.
- `xfig` "patterned" shapes are implemented with bitmap fonts. Use "solid" shapes instead.
- The `\bbold` package almost always uses bitmap fonts. You should use the equivalent AMS Fonts:

  ```
  \usepackage{amsfonts}
  ```

  followed by, e.g., `\mathbb{R}`, `\mathbb{N}`, or `\mathbb{C}` for $\mathbb{R}$, $\mathbb{N}$ or $\mathbb{C}$. You can also use the following workaround for reals, natural and complex:

```
\newcommand{\RR}{I\!\!R} %real numbers
\newcommand{\Nat}{I\!\!N} %natural numbers
\newcommand{\CC}{I\!\!\!\!C} %complex numbers
```

Note that `amsfonts` is automatically loaded by the `amssymb` package.

If your file contains type 3 fonts or non embedded TrueType fonts, we will ask you to fix it.

12.1 Margins in LaTeX

Most of the margin problems come from figures positioned by hand using `\special` or other commands. We suggest using the command `\includegraphics` from the `graphicx` package. Always specify the figure width as a multiple of the line width as in the example below:

```
\usepackage[pdftex]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.pdf}
```

See Section 4.4 in the graphics bundle documentation (`http://mirrors.ctan.org/macros/latex/required/graphics/grfguide.pdf`)

A number of width problems arise when LaTeX cannot properly hyphenate a line. Please give LaTeX hyphenation hints using the `\-` command when necessary.

Acknowledgments and Disclosure of Funding

Use unnumbered first level headings for the acknowledgments. All acknowledgments go at the end of the paper before the list of references. Moreover, you are required to declare funding (financial activities supporting the submitted work) and competing interests (related financial activities outside the submitted work). More information about this disclosure can be found at: `https://neurips.cc/Conferences/2023/PaperInformation/FundingDisclosure`.

Do **not** include this section in the anonymized submission, only in the final paper. You can use the `ack` environment provided in the style file to autmoatically hide this section in the anonymized submission.

13 Supplementary Material

Authors may wish to optionally include extra information (complete proofs, additional experiments and plots) in the appendix. All such materials should be part of the supplemental material (submitted separately) and should NOT be included in the main submission.

References

References follow the acknowledgments in the camera-ready paper. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. Note that the Reference section does not count towards the page limit.

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.