The Accuracy of Recurrent Neural Networks for Lyric Generation

Josue Espinosa Godinez
ID 814109824

Department of Computer Science
The University of Auckland

Supervisors: Dr. Gillian Dobbie & Dr. David Huang

COMPSCI 380 Project Report
Tuesday 23 October 2018

The Accuracy of Recurrent Neural Networks for Lyric Generation*

Josue Espinosa Godinez
University of Auckland
jesp142@aucklanduni.ac.nz

ABSTRACT

Machine learning use has been steadily increasing over the past decade and is finding a growing presence in the generation of numerous types of art. This brings many questions regarding the optimal neural network configurations to achieve the highest quality content possible. The search query "generate lyrics recurrent neural networks" yields less than 2500 results while "generate text recurrent neural network" yields nearly 100,000 results on Google Scholar. Generating text using recurrent neural networks is a well-researched problem, with many proof of concepts and variations. However, a more niche application is that of song lyrics. More specifically, finding the smallest possible sized training data set for artists that still generates recognisable song lyrics. This paper examines optimal training data set size and at what point larger data sets provide diminishing returns in terms of lyric quality through two tests performed on the generated lyrics to determine their quality and recognisability. First, an objective, empirical analysis is performed using Lempel-Ziv compression to assess patterns and repetition followed by a second, subjective multiple choice survey to determine artist/lyric recognisability. The results show that data sets with only two albums generally produce very similar and sometimes better results than data sets with three albums. Google's open source machine learning framework, TensorFlow, was used to generate lyrics trained using a specific musical artist.

ACM Reference Format:
Josue Espinosa Godinez. 2018. The Accuracy of Recurrent Neural Networks for Lyric Generation. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

There are many instances of machine learning use in regards to text generation, however, using recurrent neural networks to generate lyrics requires further specialisation compared to generalised text. Due to the inherently artistic nature of music and lyrics, there are additional considerations to producing realistic lyrics than regular text generation. A well-optimised neural network for producing news articles will likely not work as well for writing lyrics in the style of P!NK. Songs generally feature an overarching theme which normally contains callbacks between stanzas, a consistent emotional tone, artist-specific vocabulary, repetition, vocal melodies, and instrumental portions. When compared to the less artistic and more logical, methodical structure of general writing, it is evident one must develop and optimise a network model that keeps these factors in mind to generate the highest possible quality lyrics.

1.1 Long Short-Term Memory

It is fairly common to exploit the pattern-recognising abilities of recurrent neural networks to generate text samples from a set of articles, for example. The standard approach in using machine learning to generate text is to utilise some form of a recurrent neural network (RNN), typically using long short-term memory units (LSTM).

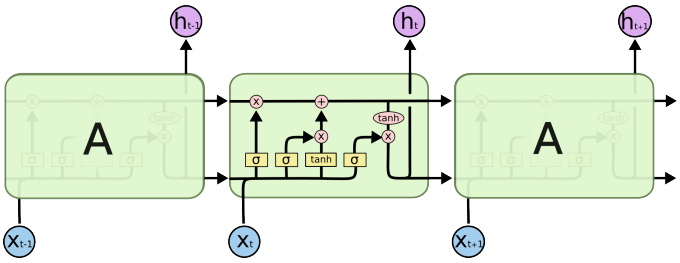

Figure 1: Figure depicting the interacting layers in a LSTM unit. [1]

Most writing is contextual; it depends on the prior words and sentences to establish meaning. Traditional convolutional neural networks classify inputs and are trained using backpropagation algorithms but do not persist sequences, making it impractical for things like identifying music, determining key events in films, and expanding on a topical thought in an article. Conveniently, recurrent neural networks allow output from a previous step to be used as input for the next step of the network in a continuous loop. The RNN repeatedly tries to find a weight matrix to maximise the probability of correct output matching the training set using the input sequences, similar to a multi-dimensional simulated annealing optimisation problem. Having the ability to collectively be the result of all prior inputs as well as the last input given, RNNs are able to remember important information and use it to modify the output. There exist many types of cells to compose our RNNs, traditional RNN cells, gated recurrent united (GRU) cells, neural architecture search (NAS) cells, the appeal of our selected choice, LSTM cells, lies in the gaps of information for context. When sequences are too long, the gradients (the simulated annealing analogy) vanish (small

*Research conducted for COMPSCI 380 for the Department of Computer Science at the University of Auckland under the supervision of Professor Gillian Dobbie and Dr. David Tse Jung Huang.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

values) or explode (large values), making it very difficult to train in these scenarios using traditional RNN units [2]. By expanding the single tanh unit as demonstrated in Figure 1, LSTMs resolve the gradient issue.

Let us take for example Queen's legendary song, "Bohemian Rhapsody". The song opens with a questioning of reality with the first verse concluding with the line "Any way the wind blows, doesn't really matter to me, to me". This is followed by five minutes of highly varied music and lyrics. The song concludes with the lines "Nothing really matters, nothing really matters to me. Any way the wind blows..." In the first example, the gap between "doesn't really matter to me" and the subsequent "to me" is immediate and RNNs can easily learn this form of repetition. However, as the size between the repetition grows such as in this example with the nihilistic lyrics that appear at the very beginning and at the very end of the song, traditional RNNs increasingly struggle to identify this pattern so we use LSTMs.

Although it was introduced 20+ years ago by a Swiss and a German researcher, LSTM solves the issue of remembering information over extended time intervals and is cited in over 13,000 articles, still in widespread use today for many applications [3]. Christopher Olah presents a thorough explanation of LSTM networks, comparing the standard RNN module's single tanh layer to the LSTM version which has several gates to conditionally let information flow to update the cell state as shown in Figure 1 [1].

1.2 Specialising recurrent neural network for lyric generation

To begin the endeavour of mimicking artists via artificially generated lyrics, it is important to first obtain lyrical samples to train the recurrent neural network on the patterns that those specific artists exhibit. The artists chosen to sample are irrelevant to the overarching optimisation problem, however, since there is a later exam on recognisability, a pre-survey is conducted on participants to determine which artists they are already familiar with to allow us to later identify at what data size the lyrics begin to resemble these artists.

After conducting a pre-survey collecting demographic information about the participants and their musical tastes, 4 artists were selected for their relevance to the participants: Ed Sheeran, Bruno Mars, Red Hot Chili Peppers, and Guns N' Roses. Due to the lack of a preexisting data set for these artists, song lyrics from their albums were all compiled in the same format in a plain text file. No special formatting was used besides the standard line spacing and denoting song parts such as "[Chorus]". To ensure the artists were all on a level playing field in terms of data set sizes, there were three separate levels of data sizes for each artist: 1 album, 2 albums, and 3 albums from each artist's respective discography.

Once we have got the training data sorted and sanitised, we need to tweak the recurrent neural network's parameters to be optimised for the task at hand. With the influence of personal taste in music, it is difficult to objectively assess the quality of lyrics, however, there are many similarities found throughout music that can be used as markers for quality. For the first test, repetition and artist-specific vocabulary were selected as a simple yet objective measure of how "song-like" the generated lyrics were. For the this particular test, the Lempel-Ziv compression algorithm was utilised to compare the amount of repetition in the hallmark songs of the relevant artists versus the artificially generated songs.

On top of the objective numerical measurement, a second, more subjective test was conducted on participants from the Department of Computer Science at the University of Auckland. 12 songs were generated by the recurrent neural network: the single, double, and triple album data sets for each of the 4 artists specified. The same songs that were evaluated by the earlier compression analysis were listed in a multiple-choice questionnaire asking what artist the lyrics most closely resemble as well as a confidence ranking on the participant's answer and their reasoning for choosing that particular artist.

Through these two tests, it is possible to analyse the relationship between how "good" the generated songs are and the size of the training data set on the optimised recurrent neural network, as well as the point at which the improvement gradually tapers into diminishing gains.

1.3 Main Contributions

There are 3 main contributions provided by this work: (1) Establishing the baseline for the smallest possible data set that still provides reasonably high quality generation to allow for quick training when no prior data set exists like in this case. If you want to quickly analyse an artist whose discography spans several decades, it is very time-consuming to manually create and format a proper data set, however, knowing that two albums is sufficiently large enough to produce unique but recognisable lyrics, we save time testing different data set sizes or constructing an unnecessarily large data set to model an experiment. (2) Establishing baseline parameters for the RNN also eliminates excess time tweaking parameters and allows for good initial parameters for further optimisation for specialised problems. (3) Analysing the results of two different approaches for quality assessment of the generated lyrics allows the reader to determine the practicality and pros/cons of the different approaches, providing a starting point for how to interpret the resulting content generated by their RNN.

2 RELATED WORK

There have been many examples of using machine learning to generate artificially created content resembling human work. A very relevant project that demonstrates using recurrent neural networks to generate text content, is Karpathy's work on training a neural network on Shakespeare's work to establish a similar structure and style through artificially generated content resembling the legendary poet's writings [4]. One of the most interesting and more complex examples of artificially generated content comes through the use of attentional generative adversarial networks [5]. Using an attentional generative adversarial network, researchers were able to create a system that attempts to produce an image described by text (testable at http://t2i.cvalenzuelab.com). Another related example is the Neural Doodle project, which allows you to draw a quick doodle, upload a source image to extract style from, and produce an image similar to the source image using your doodle as a guide with very good results if you annotate patches from the style image [6].

3 OBTAINING ACCURATE MODELS

Recurrent neural networks can be notoriously challenging to obtain accurate models for with particular data sets [7]. Generally it is preferable to make networks larger if you have the computational power and a reasonably sized data set. Since the data sets I was working with were tiny compared to standardised data sets, it was initially difficult to avoid overfitting. After experimenting with parameters and dropout values, I found a model that had good validation performance and reasonable output. This section is important because it determines the quality of the lyrics that are generated. The same parameters were utilised across all of the artists except where otherwise noted (the final data set for Guns N' Roses had a small amount of dropout added). The only thing that changes are the internal weights on the hidden states within the RNN during training for each individual artist.

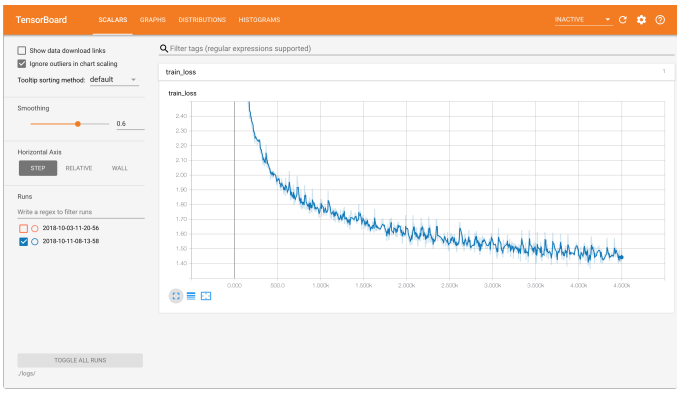

Figure 2: Graph depicting decreased error on the training set of data over time

3.1 TensorBoard

TensorBoard is a visualisation toolkit provided by Google to debug and optimise TensorFlow programs. It was used to visually assess the performance of the recurrent neural network training and tweak parameters to obtain the optimal configuration for lyric generation. Figure 2 is TensorBoard displaying the decrease of training loss over time for the training set with 3 Red Hot Chili Peppers albums.

3.2 Input Sanitisation

In order to minimise the difference between varying artists and data sizes, a consistent data format is maintained for the different training and validation data sets. Dictated by practicality, the data sets consist of a plain text file with no formatting except standard spacing, line breaks for stanzas, and structural song parts denoted in square brackets, e.g. [Chorus].

3.3 RNN size

Choosing the hidden state size of RNNs does not have a formulaic method for deciding the value and is largely based on experience. I alternated between using high and low extremes for testing. Large sizes were computationally expensive and also excessive for the comparatively tiny data sets I was working with. On the other hand, the conversely low sizes were not effective at learning so I expanded the sizes exponentially by factors of 2 before arriving at 256 for the RNN size.

3.4 Number of Layers

Due to the relatively small size of the data sets, 2 layers were sufficient for the small RNN. For the 3 album training data set with Guns N' Roses, there was a low error on the training set and a higher error on the validation set. To deal with this overfitting, a small amount of dropout was added.

3.5 Sequence Length

Generally songs have specific themes and a tone consistent throughout. Contextually, all songs vary in their composition and storytelling, especially regarding their level of complexity; it could be a simple, catchy pop song or it could be a complex suite with several sections like Bohemian Rhapsody which features no chorus, an intro, a ballad segment, an operatic passage, a hard rock section, and a reflective coda [8]. I found two stanzas to be sufficient for establishing context and remaining relevant yet interesting so 25 was chosen as the number of time steps to unroll for. The LSTM cells can remember longer than 25 but the effect falls off for longer sequences.

3.6 Dropout

Adding dropout was unnecessary a majority of the time since the optimised network worked without overfitting. For the 3 album training set of Guns N' Roses, the probability of keeping weights in the output layer was dropped to 90% and 95% for the input layer.

4 IMPLEMENTATION

The setup of the project is a TensorFlow Python project that sets up the recurrent neural network using long short-term memory units. It was given varying sizes of artist's discographies, learning a language model with words being the smallest unit of data.

4.1 Training/Validation Data Sets

The data sets were plain text, unformatted lyrics compiled into a file, with albums averaging a file size of around 20 KB. The data sets were divided at a 1 to 10 ratio for validation compared to training data sets due to their relatively small sizes.

4.2 Parameters

One-hot encodings were used for the input/output of the model. The model was trained via minimisation of the cross-entropy loss using AdamOptimizer. The following are the values for the hyperparameters.

```
Model: LSTM
RNN size: 256
Number of Layers: 2
Sequence Length: 25
Batch size: 50
Number of Epochs: 50
Gradient Clip: 5
Learning Rate: 0.002
```

```
Decay Rate: 0.97
Output Layer Keep Probability: 1
Input Layer Keep Probability: 1
```

5 EMPIRICAL ACCURACY ANALYSIS

Due to the artistic nature of music, it is difficult to assess the "quality" of generated lyrics because they have an inherently subjective value assessment individual to personal taste. The goal is to develop a consistent measurement for quality between input data sets and outputs. Firstly, an objective test based on repetition and compressibility is performed with varying weights placed in a basic formula-based format. Following this test, a subjective test with human participants is conducted, where we compare similarities and trends between the two results.

5.1 Repetitiveness/Structure

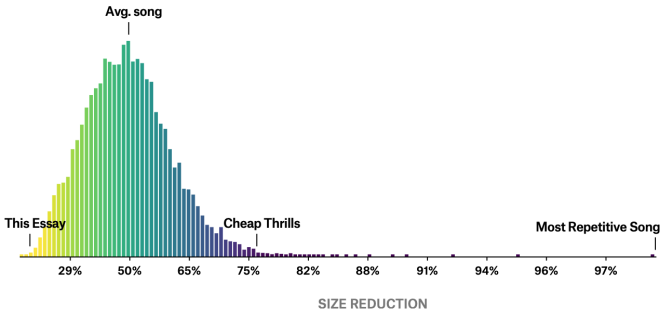

Figure 3: Graph depicting average song compression from 15,000 songs excluding the 20 most repetitive. [11]

One could argue repetition is the most important element in music; it can be found throughout every genre, culture, and time period [12]. Colin Morris conducted an excellent lyrical analysis on the repetitiveness of 15,000 of the most popular pop songs from the past 50 years, finding that the average pop hit has a size reduction of about 50% using the Lempel-Ziv algorithm shown in Figure 3 [11]. The Lempel-Ziv algorithm is a lossless compression algorithm that works by exploiting repeated sequences and powers `gzip`. The compression rate is directly related to the amount of repetition in the text [13]. Using the Lempel-Ziv algorithm, which is heavily based on repeated sequences, we can say that the higher the compression rate, the higher the repetitiveness in the song lyrics. I measured repetitiveness by comparing compression rates between a particular artist's averaged signature hits as determined by Spotify's top 10 songs for the particular artist, e.g. taking the averaged compression rate in "Here Comes The Sun", "Come Together", "Let It Be" for the Beatles and a recurrent neural network generated song trained using the Beatles. The reasoning behind choosing the top 10 songs is to develop a repertoire representative of the most characteristic songs for that particular artist. The average repetition will likely be skewed to be higher than an average of an artist's entire discography due to hit songs normally featuring higher levels of repetition, but these songs represent the best of the particular artist and ultimately represent what we want the RNN to emulate.

Another potential approach would be to randomly select 9 songs, 3 from 3 different albums, representing a more equally distributed data set likely including non-hits and deep cuts, presenting a look at a truer "average" song for that artist, however, since those songs are likely not as recognisable or characteristic as the top 10 hits, the alternative approach was selected.

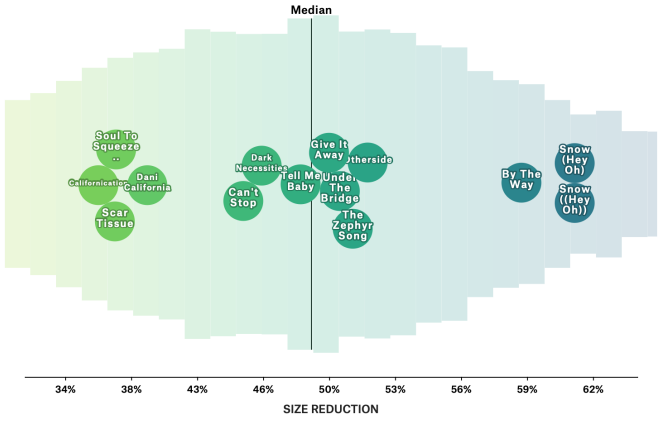

Figure 4: Graph depicting Lempel-Ziv compression rates across the Red Hot Chili Peppers discography. [11]

The `gzip` utility conveniently uses Lempel-Ziv (LZ77) compression, so it was used to carry out this comparison. Averaging the compression between the top 10 songs for an artist on Spotify and comparing the file size reduction to the artificially RNN-generated songs for each of our 4 artists with all 3 data set sizes, we achieved the results displayed in Table 1.

	Ed Sheeran	Bruno Mars	RHCP	GNR
Top 10 avg.	69%	72%	72%	70%
1 album RNN	40%	38%	33%	38%
2 album RNN	35%	41%	38%	40%
3 album RNN	36%	36%	34%	42%

Table 1: Table depicting file size reduction using Lempel-Ziv compression for top 10 Spotify songs averaged and the RNN-generated songs from 1, 2, and 3 album data sets.

It is important to keep in mind that because the top 10 most popular songs were used as comparison, there may be a higher-than-average amount of repetition which could present mildly inflated file reduction rates that are not necessarily indicative of the artist's music in general, simply the hits since we earlier discussed that hit songs were found to display higher levels of repetition. Had we used 10 random songs that are more indicative of the training set data, we likely would be comparing our RNN-generated lyrics to file compression rates that reduced file size by a lower amount which would be closer to the file reduction rates in our artificially generated lyrics.

As is evident in Table 1, the training data sets with 2 albums performed best on 50% of the tests and performed within 2% of the

best on the remaining 50% of the tests. 2 albums appear to be the ideal amount of training data since 3 album training data sets only performs better 50% of the time, and even then it is only by a margin of 2% or less. The additional work in creating the larger data set and training the network is not worth the time investment based on the results. We can also see that the recurrent neural network generates songs that on average have 33% less compressibility compared to the top 10 hits of the artists.

Another characteristic of high-quality lyrics is poetic structure, for example the number of words per line. Comparing the RNN-generated songs that displayed the highest level of compressibility with that artist's top song on Spotify, we achieve the following results:

	Ed Sheeran	Bruno Mars	RHCP	GNR
Top 10 avg.	7	7	5	5
1 album wpl	6	6	4	4
2 album wpl	5	7	6	7
3 album wpl	6	6	7	5

Table 2: Table depicting the average amount of words per line for the top 10 Spotify songs of each artist and the RNN-generated songs.

The general poetic structure of all the lyrics generated by the recurrent neural network was very similar to the original songs by the artists [7]. Overall, the performance of the recurrent neural network was positive. Based solely off this metric, one might believe that the artificially generated songs should be easily confused with original songs, however, it is important to keep in mind, NLP and working to have proper grammar and parts of speech was not implemented so while the tone is similar, the lyrics generally do not make sense grammatically. However, for further research and implementation, adding another RNN to focus on parts of speech along with NLP on the weighted sampling would improve this area significantly.

5.2 Word-based

An important note regarding training data size and quality of lyrics involves artist-specific vocabulary. For example, in "I Am The Walrus" by The Beatles, the lyrics "goo goo g'joob" appear. This is an incredibly distinctive trait of this song and the artist in general because this phrase is very rare in music in general, one could say it may very well be exclusive to The Beatles. Someone familiar with The Beatles might not recognise an RNN-generated song from the other lyrics, but that artist-specific vocabulary occurring could give it away, biasing the results. On the third question of every song, "What made you guess this particular artist?", 33% of the survey participants wrote they guessed Red Hot Chili Peppers due to the mention of California since many of their songs contain themes relating to California. However, this is not necessarily a negative response; California is characteristic of the Red Hot Chili Peppers and ultimately the goal is to resemble the artist as closely as possible, so in that sense, the recurrent neural network succeeded. The Red Hot Chili Peppers are not the only artist with a particular tone or repeated theme. Take for example the first stanza from the first song generated by the recurrent neural network trained using a single Ed Sheeran album:

```
Real love right get you like
on we there little of ya on my guitar,
preach jukebox when I was and I people,
deserve her you takes you wait
```

Although it does not make grammatical sense, there is a general theme that could be argued resembles Ed Sheeran's writing. It discusses real love, a guitar, preaching, a jukebox, deserving a girl, waiting, etc. The words found even at the very beginning of the song establish a feeling that is reminiscent of Ed Sheeran and participants understandably mentioned choosing him due to the mention of a guitar and English slang found later on in the lyrics. Sex, alcohol, and women were themes that were mentioned for Bruno Mars by participants. A specifically denoted guitar solo section and unique words particular to songs by Guns N' Roses such as Brownstone were described as primary reasons for selecting Guns N' Roses as the probable training artist in some RNN-generated lyrics.

5.3 Results

The songs that were generated by the recurrent neural network demonstrated similar but not comparable levels of repetition. On average, the RNN-generated lyrics lagged behind by 30% compared to the average file reduction rate of the top 10 Spotify songs for each artist.

Beyond just common repetition, structured repetition such as in "Shape of You" by Ed Sheeran where a line is repeated several times such as the choral phrase, "I'm in love with the shape of you", was highly uncommon. Small phrases were sometimes repeated or words appeared twice in a row, but never to the level of repeating a line four times in a chorus. In the RNN-generated song "Dangerous" with 1 Bruno Mars album as a training data set, the pattern "girl" is repeated as the final word at least once in a stanza 4 times. In the corresponding 1 album Guns N' Roses RNN-generated song, the phrase "n-n-n-n-n-n-n-n" is presented after the intro and before the outro. These minor levels of structured repetition elicit some similarities between the source material but not quite on the same level used by artists today with recurring motifs and entire repeated choruses.

Song parts in general were recreated very well, with the structure of choruses and verses normally presented very similarly as the source songs in the format "[Chorus]", "[Verse]", etc. The length of lines, along with average words per line were consistently nearly the average of the original songs, regardless of the size of the training data. The main impact training data sizes had was on the compressibility and the standard deviation was normally within 5% with linear growth between the training data sizes.

6 SUBJECTIVE ACCURACY ANALYSIS

As previously mentioned, music is inherently subjective and so naturally, it is important to not only gather a numerical analysis on how training data size impacts the perceived similarity between artificially generated lyrics and real ones, but also human perception to discover what the threshold is for recognising a particular artist's style. How much training data is really necessary before a majority of participants can say "these lyrics really seem like they were

written by this artist"? Using the same songs from the empirical analysis section, multiple choice question were created. A set of similar artists are presented along with a confidence ranking and an explanation for why the participant chose that particular artist. This survey enables us to capture the practical real-world side of the generated lyrics since ultimately how humans perceive the lyrics is the important part since they are the ones to determine whether or not the lyrics actually resemble a particular artist.

6.1 Pre-Survey

In order to produce the most relevant results, it was important to determine the participants' musical knowledge and ability to recognise artists via song lyrics alone. For this purpose, a pre-survey questionnaire was sent out asking 6 participants about their ability to recognise artists exclusively via song lyrics. The survey also asked for their levels of familiarity with particular genres/artists to determine which artists would provide the most relevance when generating lyrics, as well as to increase familiarity with lyrical habits and traits of the participants' personal musical taste. Demographics were also collected to evaluate how their age/gender might influence their musical taste to list possible biases in the results. The demographics heavily influenced the results, with all of the participants being over the age of 25 and a majority being between 35 and 44. There was an even 50% split between men and women. Pop was selected as the most popular genre with 67% of people marking it as their most familiar genre, with Rock selected as second at 33%. 67% of participants indicated average to above average confidence in determining an artist based on song lyrics alone. Specific artists that were mentioned several times included Guns N' Roses, Red Hot Chili Peppers, Bruno Mars, and Ed Sheeran, so those were the artists selected for use in the survey. Conveniently, the 2 most popular genres pop and rock both had 2 artists available for sampling in the survey that followed.

6.2 Survey Design

A test was given to a group of 9 people from the Department of Computer Science at the University of Auckland to determine the minimum amount of training data necessary before a majority of participants would be able to successfully recognise artists. The most popular artists listed in the pre-survey were selected to be used in the survey with 2 artists from the 2 most popular genres. Red Hot Chili Peppers and Guns N' Roses for the rock genre and Bruno Mars and Ed Sheeran for the pop genre. Due to Bruno Mars and Ed Sheeran both only having 3 albums, 3 levels of data sets were tested for all of the artists: 1 album, 2 albums, and 3 albums. All artists had similar sizes between their albums e.g. each of Ed Sheeran's 3 albums had a similar size for total lyrics. The recurrent neural network was fed each of the 3 data sets for each artist, generating lyrics for each of the 4 artists. Consequently 12 songs were generated for the 4 artists at the 3 levels of data sizes.

The survey was 12 pages long, with one page for each "song" with the same questions for all of the lyrics:

```
1) Which artist do these lyrics most closely resemble?
2) How confident is your guess?
3) What made you guess this particular artist?
```

For the first question, a multiple A, B, C, D, choice question was presented with an answer set consisting of artists displayed on the Spotify page for that particular artist under the "Fans Also Like" page to construct an answer set of similar artists to determine if the lyrics generated truly resemble that artist. By populating the answer set with similar artists, you would have to distinguish them from several similar options, proving the lyrics truly resemble the particular artist. This answer set is reused for all of the questions regarding this particular artist, that is, the answer set for all 3 of the Ed Sheeran songs was always the same, regardless of the amount of training data. Overall there were 4 answer sets (one unique answer set for each of the 4 artists). The songs were randomly shuffled as well as the answers to avoid bias/preference for a particular letter e.g. someone blindly selecting "A" for every question. This also forces participants to read all of the options since their order might have changed from what they previously were.

The purpose of question 2 is to reduce the weight placed on the answers of participants that do not feel confident about their familiarity with the artist or participants who are just guessing and do not feel that the lyrics demonstrate a resemblance to any of the listed artists in particular. By knowing the confidence level of participants for a particular question, we can separate our results from the answers with no confidence to eliminate an unfair bias in the findings. The options available were no confidence, low confidence, average confidence, and high confidence. The results analysed normally only take into account the questions answered with average or higher confidence to prevent random guesses from skewing the final results.

The purpose of question 3 is to establish if there is any consistent pattern that biases results. For example, many participants marked their reasoning for picking Red Hot Chili Peppers as "because the lyrics mention California", perhaps mildly skewing the correct response rate positively (the Red Hot Chili Peppers do reference California and California-related themes in many of their songs).

6.3 Survey Results

	No c.	Low c.	Average c.	High c.
Ed Sheeran 1 album	50%	0%	50%	N/A
Bruno Mars 1 album	25%	33%	100%	0%
RHCP 1 album	33%	100%	100%	100%
GNR 1 album	60%	100%	50%	100%
Ed Sheeran 2 albums	0%	0%	50%	N/A
Bruno Mars 2 albums	75%	67%	50%	N/A
RHCP 2 albums	40%	0%	100%	100%
GNR 2 albums	50%	0%	100%	100%
Ed Sheeran 3 albums	67%	50%	100%	N/A
Bruno Mars 3 albums	100%	50%	50%	100%
RHCP 3 albums	25%	67%	100%	100%
GNR 3 albums	60%	0%	100%	50%

Table 3: Table depicting survey results for all answer confidence levels and training data sizes.

There were 9 participants in the survey. The majority of the participants were male and over the age of 35. Using answers with

Figure 5: Graph depicting % of correct answers with average or higher confidence for all of the artists and the different data set sizes.

all confidence ratings, 53% of answers for the 4 questions with only 1 album of training data were correct. Similarly 47% of questions with 2 albums of training data were correct. Finally 56% of 3-album-trained questions were correct including all confidence levels.

Since these results show low variance, I investigated artist-specific results to determine which artist had the highest scores to use it as a more pointed example and reduce bias with artists that performed more poorly on average such as Ed Sheeran. Including all confidence levels and all 3 questions for every artist, Ed Sheeran scored 37% correct overall, Bruno Mars scored 48%, Red Hot Chili Peppers scored 59%, and Guns N' Roses scored 63%. Given that the participants were mainly 30-40 year old men, the results are understandable since the Red Hot Chili Peppers (RHCP) and Guns N' Roses (GnR) are both roughly 30 years old and were very popular at their peak compared to Ed Sheeran and Bruno Mars who are both modern artists.

Using Guns N' Roses and the Red Hot Chili Peppers, the following results were determined. 80% of Guns N' Roses questions marked with an average or higher confidence level were correct while 100% of the Red Hot Chili Peppers songs were correct at this same confidence level. This may also be due to lyrical content bias as covered in the Pre-Survey explanation for the third question of explaining the reasoning for choosing a particular artist. For example, Red Hot Chili Peppers are easily identifiable due to lyrical themes relating to California. However, since the purpose of this survey is to determine how data size impacts lyrical artist resemblance, let us investigate the relationship between data sizes and correct guesses. Using only one album and all answers regardless of confidence levels, Guns N' Roses songs were correct 67% of the time while the Red Hot Chili Peppers were correct 78% of the time. Using 2 albums and all confidence level answers, Guns N' Roses was correct again 67% of the time while RHCP was correct only 44% of the time. Finally, using 3 albums at all confidence level answers, Guns N' Roses was correct 56% of the time and RHCP was also at 56%.

One possible reason for the discrepancy behind this finding might not be due to the size of the recurrent network losing accuracy but rather the sizeable discography these particular artists have and how long-term artists evolve over time and change styles between albums and eras. When fed a data set with many different types of lyrics and themes, the RNN roughly produces an average that is difficult to differentiate akin to having several bright colours that are easy to identify independently but when mixed together create a non-distinctive brown colour that is difficult to identify. However, it is also likely this is due to people unfamiliar with the artists skewing the results poorly since many of their lyrics were marked with low-to-no confidence. This illustrates a heavy downside to evaluating using human metrics, it is very subjective and there are many external factors that can potentially influence results including but not limited to: artist familiarity, personal skill at recognising artists via lyrics alone, and test-taking skills.

When all artists are taken into account however, with an average or higher confidence level, we see results consistent with the prior empirical analysis: after 2 albums, there is little to no improvement in lyric quality.

Looking at Table 3, we can find more similarities between the earlier objective test and the present subjective test. The 1 album data set has the lowest scores on average/high confidence, 2 album data sets have very similar scores to 3 album data sets with the 3 album data sets performing only slightly better. This reinforces the idea that we experience diminishing returns in regard to training data size and improvements in performance after 2 albums.

	No c.	**Low c.**	**Average c.**	**High c.**
% of Correct Answers	43%	39%	75%	82%

Table 4: Table depicting relationship between answer confidence and correctness.

The confidence/correct relationship is also very intuitive. As Table 4 clearly demonstrates, there is a straightforward linear relationship between answer confidence and correctness.

Utilising only the relevant results, Guns N' Roses single album data set with average or higher confidence guesses were correct 67% of the time. At 2 albums they were correct 100% of the time, and at 3 they were again correct 67% of the time.

A note to keep in mind regarding these results: although they are more useful since the participant feels sure of their guess and/or familiarity, this is coming from a small fraction of an already small data set.

7 CONCLUSION AND FUTURE WORK

There are many approaches to advancing artificial content generation with recurrent neural networks. We examined using a LSTM

RNN on 3 varying data set sizes for 4 artists and established 2 albums as the smallest training data set with the highest relative levels of artist recognition. We were able to eliminate one aspect of investigation when it comes to the ideal setup for the best content output for recurrent neural networks. In this initial version of our RNN, the focus was to demonstrate the feasibility of barebones optimised recurrent neural networks to minimise setup and maximise return on investment of output. In our testing we observed our RNN was successful in achieving decent compressibility scores and similar poetic structure to original lyrics, as well as a majority of correct identification from survey participants with the aforementioned RNN parameters and training data sets.

One limitation of this system is the disregard for grammar; not only are the data sets not large enough to provide an adequate data size to learn proper English structure but there is no collaboration with Natural Language Processing (NLP) algorithms or language parsers. With this limitation, the lyrics generated can feel and resemble an artist but ultimately be grammatically incorrect. Future work includes grammar and NLP development as an area for dramatic improvement of quality of lyrics, as they are currently not present. Despite its lack of formal training in producing proper English phrases, the RNN manages to capture the tone and feel of artists adequately enough for its source training artist to be identifiable by people, which is substantial progress. Additionally, mapping sentence structure with language parsers such as the Stanford Parser would dramatically improve the similarity between generated lyrics and real ones, narrowing the gap between artificial and genuine lyrics.

Additionally, further empirical analysis of generated song lyrics could prove to be interesting and useful in measuring the quality of generated lyrics. Analysing the parts of speech statistics could be used to measure frequency of noun-to-verb ratio for example, to not only have proper grammar but also compare how closely the RNN follows characteristics of artists in parts of speech and grammatical writing habits of the artist.

It may prove to be useful to develop a system to give artists a recognisability score based on how unique their lyrics are or how consistently they present a theme throughout their work. If this were to be developed it would allow us to then select artists with an average score to minimise artist-specific traits from impacting our results in identifying the quality of artificially generated song lyrics.

Using recurrent neural networks to generate artificially created lyrics appears to have significant potential in creating lyrics that closely resemble the original artists with enough optimisation and development. RNN-generated lyrics have many practical applications. For example, it could be used in the event that a lead singer passes away and the new singer is unable to write lyrics that accurately capture the band's style; this could allow for a RNN to assist in the writing process while still allowing human creativity to flow using text priming and varying temperatures for more interesting lyrics. As featured in a New York Times article, artificial intelligence is beginning to assist novelists during the writing process [14]. Ultimately, the applications of utilising optimised recurrent neural networks is nearly limitless in the creative space and there is great potential for future work.

REFERENCES

[1] Olah, C. (2015). Understanding lstm networks, 2015. URL http://colah. github. io/posts/2015-08-Understanding-LSTMs.

[2] Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks, 5(2), 157-166.

[3] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[4] Karpathy, A. (2015). The unreasonable effectiveness of recurrent neural networks. Andrej Karpathy blog.

[5] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. (2018). Attngan: Fine-grained text to image generation with attentional generative adversarial networks. arXiv preprint.

[6] Champandard, A. J. (2016). Semantic style transfer and turning two-bit doodles into fine artworks. arXiv preprint arXiv:1603.01768.

[7] Sutskever, I., Martens, J., & Hinton, G. E. (2011). Generating text with recurrent neural networks. In Proceedings of the 28th International Conference on Machine Learning (ICML-11) (pp. 1017-1024).

[8] McLeod, K. (2001). Bohemian rhapsodies: operatic influences on rock music. Popular Music, 20(2), 189-203.

[9] Graves, A. (2013). Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850.

[10] Mikolov, T. (2012). Statistical language models based on neural networks. Presentation at Google, Mountain View, 2nd April.

[11] Morris, C. (2017). Are Pop Lyrics Getting More Repetitive? Retrieved from https://pudding.cool/2017/05/song-repetition/

[12] Margulis, E. H. (2014). On repeat: How music plays the mind. Oxford University Press.

[13] Ziv, J., & Lempel, A. (1978). Compression of individual sequences via variable-rate coding. IEEE transactions on Information Theory, 24(5), 530-536.

[14] Streitfeld, D. (2018). Computer Stories: A.I. Is Beginning to Assist Novelists. URL https://www.nytimes.com/2018/10/18/technology/ai-is-beginning-to-assist-novelists.html.