Multimodal Lyric Generation From Image
CIS 5300 Milestone 4 Final Report

Reifon Chiu, Aditya Jayant G, Edward Park, Chengyi Zhang
Penn Engineering, University of Pennsylvania.

[Verse 1: Paul McCartney]
In Penny Lane there is a barber showing photographs
Of every head he's had the pleasure to know
And all the people that come and go
Stop and say hello

[Verse 2: Paul McCartney]
On the corner is a banker with a motorcar
The little children laugh at him behind his back
And the banker never wears a mac
In the pouring rain
Very strange

[Chorus: Paul McCartney with John Lennon & George Harrison]
Penny Lane is in my ears and in my eyes
There beneath the blue suburban skies
I sit and meanwhile back

Figure 1: Penny Lane from The Beatles

Abstract

A lot of famous lyrics are inspired by great views or images. In this project, we made a lyric generator that can generate lyrics based on given images and given genre information. We fine-tuned the smallest GPT-2 model with 124M parameters and used the BLIP image captioning model. The model can generate rock, rap, gospel, and pop lyrics, and we reached a BLEU score of 0.31 during tests.

1 Introduction

Songwriting may be regarded as a very human endeavor. Many ideas and concepts could go into writing good lyrics, including mental imagery and delivery. Using artificial intelligence to generate lyrics thus poses a challenge: can we incorporate approximately the same information a human might use to generate better lyrics? Another attempt at this task has used MIDI data along with starting lyrics to generate novel lyrics (Vajipey et al., 2023). However, in this paper, we attempt to use genre tags and images to assist in the lyrics generation process. Using a transformer-based architecture, our task is to augment starting text with supplemental material as an attempt at original lyrics generation. This problem is well-suited for NLP since lyrics are generated by a machine instead of a human. It also seems to be an open area that could use experimentation or insights.

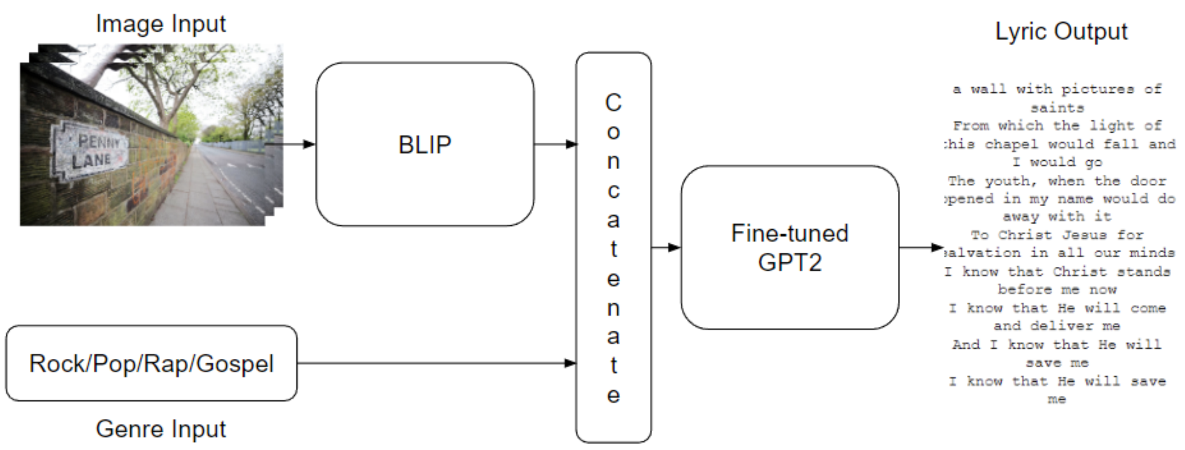
Figure 1 is an example of lyrics inspired by views. The Penny Lane, written by Paul McCartney, was inspired by the view of the actual Penny Lane in Liverpool. It captures the beauty in everyday life and the people resonate with the lyrics.

Figure 2 shows an illustrative example of our model. We extract the information from a given image, pass it along to the fine-tuned lyric generator with the genre information provided by the user, and then we finally output the lyric. Our model architecture is mainly based around GPT-2 with a language modeling head.

We decided to pick this topic for our term project because it was a creative project and it interested all four of us. We wanted to see what we could do for a task with no real right answers and try to generate something convincing. Unlike a more traditional task, this one might be used in more hobbyist applications.

2 Literature Review

Briefly mentioned in the introduction, Vajipey et al. (2023) uses an approach that utilizes both MIDI and text to generate lyrics. The MIDI data is used to get embeddings by feeding them into MusicBERT. The embeddings are then input to a linear layer to project them into GPT-2 embedding space. This projection is then concatenated with text embeddings, which are created from token and position embeddings. This concatenation is then fed into a slightly modified GPT-2, then into an LM head. Since lyrics are different than natural language, the measures they use include coherency and originality. As for their results, the model seems to struggle with originality, and some outputs do not have a proper structure. They attribute the shortcomings to a lack of data quality and to how they use repetitive genres of music. Additionally, the authors believe that the token limit of GPT-2 prevented the model from achieving a more complete representation of songs.



```
a wall with pictures of
saints
From which the light of
this chapel would fall and
I would go
The youth, when the door
opened in my name would do
away with it
To Christ Jesus for
salvation in all our minds
I know that Christ stands
before me now
I know that He will come
and deliver me
And I know that He will
save me
I know that He will save
me
```

Figure 2: Diagram of the final model

Ma et al. (2021) proposed an automatic lyrics generation system with four modules: a music structure analyzer, a SeqGAN-based lyrics generator, a deep coupled music–lyrics embedding model, and one called Polisher. They paid attention to how the lyrics should be coherent and meaningful. Since they use multimodality, the lyrics should match the MIDI files' style and rhythm. To address challenges, they enhanced the dataset by adding human labeling about the sentiment of the music, among others. For the music structure analyzer, they assumed the expected number of syllables based on pop music. The lyrics generator is a syllable-conditioned bidirectional LSTM-RNN. The overall reward for the generator is a linear combination of the two different discriminators. In the Deep Coupled Music–lyrics Embedding part, gated-CNN is used instead of LSTM to speed up the training. BERT is used to extract features of the lyrics. The results show that the model surpasses other models in both objective and subjective evaluation, and the ablation study showed the importance of each part of the system.

Finally, Malmi et al. (2016) uses a two-step method to generate smooth and flowing rap lyrics. Firstly they used a prediction algorithm to identify the next line of lyrics using a dataset of pre-existing next lines using SVMs and novel neural networks. Then they use the prediction model to combine the lines from the existing songs, producing lyrics with rhyme and meaning. Using an information retrieval perspective, the partially constructed song is considered a query and is updated after every new line prediction, mixing lines from different artists to get new songs. One interesting implementation in this paper is the rhyme density measure, used to quantify the quality of the lyrics. Roughly speaking, this measure is based on matching vowel sequences averaged across lines. They also work on extracting different features in each song, but the features themselves can be divided into categories like rhyming, structural, and semantic similarity.

3 Experimental Design

3.1 Data

For our data, we decided to use **Song Lyrics From 79 Musical Genres**, downloaded from Kaggle. This dataset is dominated by English and Portuguese songs, so we first filter it to just English songs. Then two versions were prepared for different stages. The first version is just comprised of songs in the Rock genre, using a popularity metric above 5. We believe a high popularity should indicate that the lyrics have good quality in general. 21319 such songs were pulled. The second version instead filtered the lyrics dataset according to the genre as indicated by the artists' dataset. Specifically, we filter into four genres: Rock, Pop, Rap, and Gospel. Then, if we can, we sort by popularity and take the top 4500 per genre. Now for both versions, once the songs are pulled out, those that have too long of lyrics get pulled out to fit into our model. Then some starting words are isolated for each song, as well as an expected model output. 500 randomly chosen songs out of the respective total songs for both versions are used as a test set.

	# of songs (V1)	# of songs (V2)
Train	20819	17500
Test	500	500

The above table summarizes the number of songs we kept for both pre-processed versions.

ALink	SName	Lyric
/ivete-sangalo/	Arerê	Toda vida, é...
/ivete-sangalo/	Se Eu Não...	Sem direção...
/bruno-mars/	When I Was...	A little bit...

The above table shows examples of some columns from the lyrics part of the dataset.

Artist	Genres	Songs	Popularity	Link
Ivete Sangalo	Pop;...	313	4.4	/ivete...
Chiclete...	Axé	268	3.8	/chicl...
Bruno Mars	R&B;...	122	62.5	/bruno...

The above table shows examples from the artists part of the dataset. The ALink column is a foreign key of Link, which allowed us to join the two tables.

3.2 Evaluation Metric

For reference purposes, we use BLUE-4 with a brevity penalty. Specifically, let H represent our model output, and let R be reference lyrics. Then let grams(i, T) be the set of i-grams from text T. Then let mCount$(g, T, T') = \min($# of occurrences of n-gram g in T, # of occurrences of n-gram g in $T')$ be a modified count of n-gram g in text T. That is, the number of g that we will take cannot exceed the number of g in T'. And then let count(g, T) be simply the number of occurrences of n-gram g in text T. If we let len(T) be the amount of words in some text T, we then define the brevity penalty B as 1 if len$(H) >$ len(R), $\exp\left(1 - \frac{\text{len}(R)}{\text{len}(H)}\right)$ otherwise. Then for a particular order n, we will define a modified precision $p_n = \frac{\sum_{g' \in \text{grams}(n,H)} \text{mCount}(g', H, R)}{\sum_{g' \in \text{grams}(n,H)} \text{count}(g', H)}$. Finally, we compute a BLEU-4 score as $B \cdot \sqrt[4]{p_1 \cdot p_2 \cdot p_3 \cdot p_4}$. Over multiple model outputs, we just average these scores together. However, because BLEU scores cannot be used for lyrics, we should also consider the qualitative aspects of the lyrics outputs when applicable.

3.3 Simple Baseline

Our simple baseline uses the original GPT-2 (no fine-tuning) to generate text based on test set lyrics. The simple baseline achieves a score of 0.10536856686666983.

4 Experimental Results

4.1 Strong Baseline

For the strong baseline, we fine-tuned GPT-2 for lyric generation with our dataset. We first picked the lyrics based on language and artist popularity, since we are only generating English lyrics and we believe that high popularity generally indicates good lyrics. We then picked based on length, since there's an input length limit of 1024 tokens for GPT-2. We didn't want to crop the lyrics into two parts since that would make the lyrics start or end in undesirable places. Accumulated gradient is used during training for simulating larger batch sizes due to the limitations of using Google Colab. Overall, we trained with around 20k data points for 20 epochs. The whole training process took 2 hours. 500 lyrics were split out from the dataset for testing. During testing, we split the lyrics into two parts: the last 20 words and the rest of the lyrics. Lyrics are fed into the model as input and we process the output to get the generated part. Finally, we check the BLEU score for the generated lyrics against the true end lyrics. The fine-tuned GPT-2 reached a BLEU score of 0.31, which is significantly higher than the original GPT-2 model.

4.2 Extensions

One extension we tried was to tag each song's lyrics with the corresponding genre. This is what led to the second preprocessed version of the dataset we used. For example, we tag the start of lyrics with <RAP> for Rap songs. Then, we finetuned our model even more on this second version dataset. The idea behind this extension is to see if our model could benefit from such a tag. This extension resulted in our model being able to generate lyrics for the different genres.

Using "a man in the mirror" with the Rap genre, we get the following first few lines:

```
A man looking in the mirror.
When I say that the man looks in the mirror
I'm speaking of the man who says it's my turn.
Yeah, and I was wondering what I was doing
```

Using the same text but with the Gospel genre, we get the following first few lines:

A man looking in the mirror. You're my savior
I won't hide from you, man. I'll tell you everything. I'll keep your life.
I'll just see the world
through your eyes
For your part, I'll tell you everything

Another extension we implemented was using BLIP in our overall architecture. This allows us to connect images to the generation of our lyrics. Using an image serves as a proxy for having mental imagery, which many songwriters try to convey. All one needs to do is to provide an image, which will have a caption associated with it. This caption is additionally fed along with a genre tag to produce lyrics. We present examples of different images for the different genres below. We generated all 4 genres for the penny lane image, and only the pop genre for the other images:

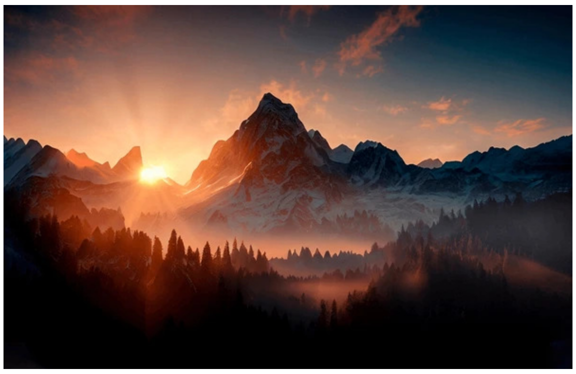
Pop:
Jump on a street that will only stand for my song
Come out in droves, try and fight and get down
But wait for the train to run off by now
I don't care what the hell you say, it'll be the same
'cause we'll all try to get out alive
If you can't hit a rock
If you can't break a guitar
If you can't smash a record
Then take this love, make it easy to be strong

Rock:
Telling you that this is you
And all the time I've been there I've seen you dream and you never give a damn
It's because of what you did
Wake up tomorrow, all the right
And I can hear you saying,Dear God, God
This is a word from you
That you would really do for us
A word from God that we could never hear
If you would give us this chance
We'll do something for you

Rap:
There's nothing out there, nobody knows
So my mom used to tell me
You go to the store and pick up somethin'
You do everything in your power to save the world
I mean, what else can I do?
You say I'll go in the next cab, and
the cab comes by, you go in the next cab
Yeah, you're gonna be the winner
All this other shit just make you cry

Gospel:
a wall with pictures of saints From which the light of this chapel would fall and I would go
The youth, when the door opened in my name would do away with it
To Christ Jesus for salvation in all our minds I know that Christ stands before me now
I know that He will come and deliver me
And I know that He will save me
I know that He will save me

Here, we can clearly see the distinction between each genre. The pop is more about dancing and rhyming and upbeat in general, while the gospel is very religious and mentions words like saints and Jesus. Rap and rock are more about storytelling, while rap uses more strong and harsh words that rarely show up in other genres like "shit".



Pop:
We're down on the mountains of the night
They're as steep as ice
And their mouths are the same as gold
When we've been sitting here all day
We're old, but the wind blows cold
We're lost in the land of shadows
This way we can change our minds
This way we can change our minds
We will tell you the truth
Do you think there's a place that we're not at

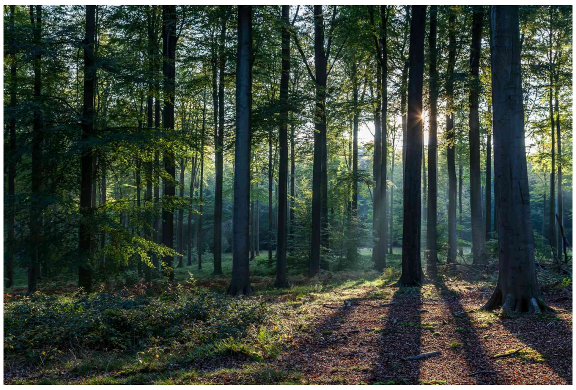
The lyric captures the objects like the mountain and ice in the image, and the time, which is dawn or dusk. There are rhymes like "gold" and "cold".

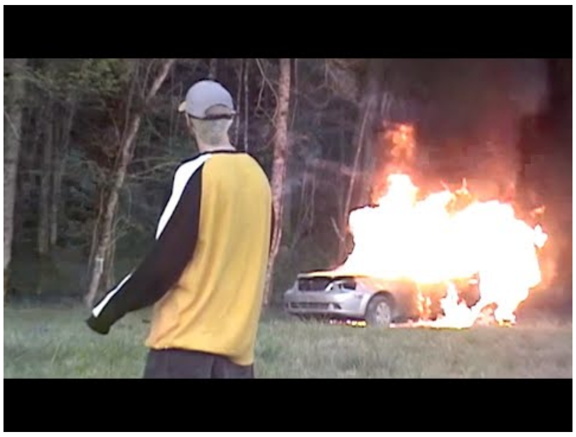
Pop:
Only one way out of life, another step away from this city
Won't you get along?
The change you'll make
chorus

Well, we're good, I'm sure you'll be proud of that
We're not weak, I'm sure you'll be proud of that
We're not alone, it's a real train of thought
We're all we know
solo

You're not a single voice
We're all we know
So we're all we know
We're all we know
You're not a single voice
We're all we know
So we're all we know
We're all we know

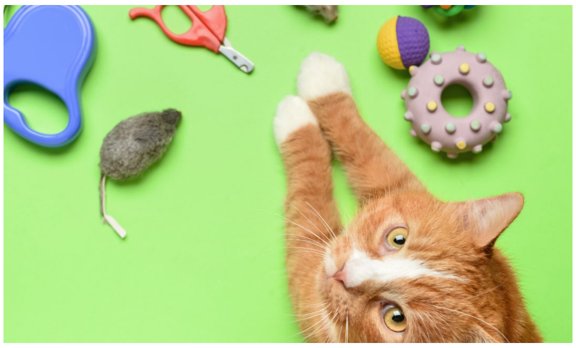
It captured the scene of the image, which is nature and far away from the city. It then further expands based on it.





Pop:
I am the lion
I'm a man standing in front of a car
In the skies
Yes, we're very fast
The god of my day
And I see this to-night
Chorus
Hail the Devil
Oh my Lord God
We are the devil
In our place
We are the road to salvation
We are the pain of our face

This is a random frame I cropped from an MV. It's very random and has a grim filter on it. The model made some cult-ish lyrics that also fit with the image. From the car on fire, it further expands to talk about the devil and god.



Pop:
Looks like a human
Then the dream fades away
And everything's bright
The bird is waving its wings
Everybody who touches it
Is happy
And makes the dings go away
Yes, you can sing
Yes, you can sing
People in the street love you
You're more than just a cat
You're bigger than a hill
And you're in a box
And the birds are singing
Yes, you can sing
Yes, you can sing
Birds sing and some people sing

This can be interpreted as from the cat's point of view. From imagining what the cat has seen by looking up, to an upbeat lyric of what the life of a cat can feel like.

4.3 Error Analysis

We noticed that our final model sometimes generates the same lines unnecessarily repeatedly. For example, "Oh dear world" was generated countless times despite having to do nothing with a song, when using it as a base. This may be attributed to the fact that many songs are repetitive in nature, especially Pop songs. For the song Limbo, the true end lyrics had 7 "ohh", and our model produced 12 of them. Sometimes it will miss some repetition like for the song I Wanna Rock in the test set, which had a lot of "no" at the end, but our model only generated 3.

Another type of error we see seems to originate from the dataset itself. Due to the sheer number of words in the sheer number of songs, we were not able to thoroughly look through the dataset (which was scraped) and clean it. One effect of this can be seen in "Baby don<u>â€™</u>t go", where the erroneous parts like what is underlined show up in scraped lyrics. Some lyrics contain the name of the singers in it before each line, and some others have timestamps in them, or marks like "(Chrous)". All of these tags in the lyrics confused our model and sometimes it will output weird tags or names. For example, for the song "Part II (On The Run) (Solo Version)", it generated:

"it all away
[Chorus: Beyoncé]
[2:30 - 2:40]
[Clip: You"
which is nothing close to the true end lyrics.

5 Conclusions

Through our term project, we managed to fine-tune a model based on GPT-2 and BLIP such that we could generate lyrics from an image in the style of a particular genre. It will extract key information from the images and incorporate it with the provided genre tag, then create lyrics based on the input. The lyrics are generally closely related to the provided images, and there are clear distinctions between generated lyrics of different genres. Although we do not currently believe that other models have the same implementations, we believe that our results are decent for the task that we have been tackling. At the very least, we have utilized proxies for human imagination to generate song lyrics. We'd like to believe that this kind of idea has a long way to go, and perhaps models down the line can be used to assist songwriters.

References

Xichu Ma, Ye Wang, Min-Yen Kan, and Wee Sun Lee. 2021. AI-Lyricist: Generating music and vocabulary constrained lyrics. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 1002–1011, New York, NY, USA. Association for Computing Machinery.

Eric Malmi, Pyry Takala, Hannu Toivonen, Tapani Raiko, and Aristides Gionis. 2016. DopeLearning: A computational approach to rap lyrics generation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM.

Vivek Vajipey, Anthony Zhan, and Steven Zhao. 2023. Multimodal transformer-based lyric generation from MIDI and text data.