

CS505 Project Report: AI Poetica

Junyi Li, Tianchi Huang

link to repo:

<https://github.com/JaydenLi0206/AI-Poetica.git>

Abstract

We present AI Poetica, a GPT-2 based model that generates poetry autonomously based on tokenized vocabulary and curated corpora of 20th-century poetry. The novelty of the model lies in its dataset, which consists of a collection of profound poetic works that enable the generation of poetry that is stylistically diverse and coherent. The model generates decent poems with sophisticated writing style. To assess the model's efficacy, perplexity scores and Flesch-Kincaid Readability Tests were employed, revealing high variability indicative of creative depth. Future work includes experimenting with more advanced models and expanding the dataset to refine quality and creativity in poem generation.

Introduction

Poetry is regarded as a sacred form of human expression because it combines the richness of language with the profundity of feeling and creativity. The relationship between artificial intelligence and literature in this era of digital revolution has opened up fascinating opportunities for the creative writing community. Our project, AI Poetica, combines poetry, a classic art form, with cutting-edge technology by utilizing Natural Language Processing (NLP). In order to produce poems that are compelling in terms of both style and content, we have set out on an ambitious project to create a model based on GPT-2, one of the most sophisticated language processing algorithms.

Our project is distinguished by the compilation of a distinct body of work that includes some of the most renowned poetry pieces of the 20th century. This collection of works is more than just a dataset; it's a meticulously crafted tapestry of literary masterpieces that capture a wide range of topics and styles that characterize a rich period in poetry. Our goal is to extract the essence of the thematic resonance and structural elegance of poetry by fine-tuning GPT-2 on this carefully selected corpus. Our model has been customized to identify and imitate the subtleties that contribute to the rhythmic and expressive potency of poetry. This report outlines the thorough procedure, which begins with the preparation of the dataset and ends with the model's fine tuning, laying the groundwork for AI Poetica to produce coherent poetry..

Methods Process

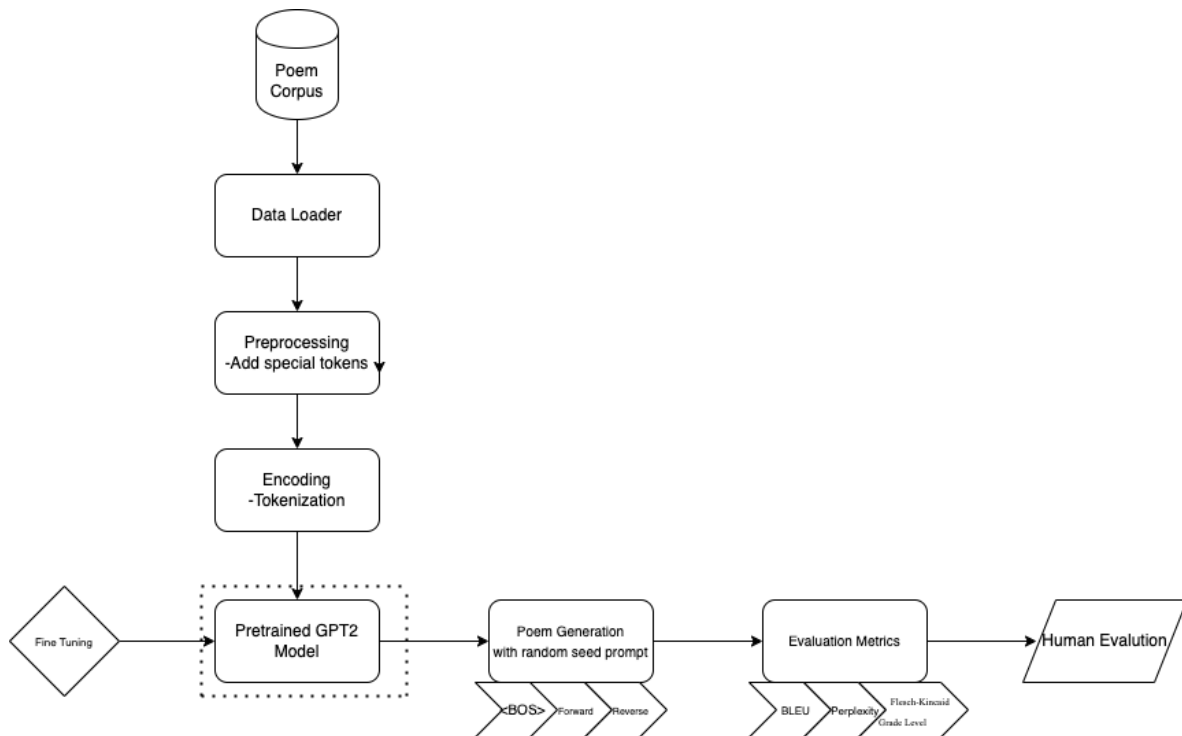


Figure 1: Poem Generation Process for AI Poetica

Dataset Preparation

This project's dataset is a hand-picked compilation of poems that were scraped from Project Gutenberg. Poets from the 20th century include T.S. Eliot, Cyril Scott, Ezra Pound, and others. Two primary fields comprise each record in the dataset: 'title' denotes the name of the poem or segment, and 'stanza_text' holds the poem's or excerpt's actual text. Poetry from the early to mid-20th century is characterized by a variety of styles and profound expressions, which are reflected in the texts' rich imagery and intricate themes. This dataset is an invaluable tool for anyone interested in the development and complexity of modern poetry, literary analysis, or tasks involving natural language processing.

We created a class for a custom dataset designed for a PyTorch model, which takes poem stanzas as input and preprocesses them for training or inference with a GPT-2 model. It tokenizes the input text with special beginning and end tokens, truncates or pads them to a maximum length, and creates attention masks for the model, providing a standardized way to load and handle poem stanza data for training.

Data Loader

Two `DataLoader` objects are initialized for the training and validation datasets of poem stanzas, respectively, in the data loading section of the code. The validation data loader uses a `SequentialSampler` to iterate through the data in order, while the training data loader uses a `RandomSampler` to batch and shuffle the data to improve the model's generalizability. Both loaders use a specified batch size of 3 to determine the number of samples processed in each training step.

Encoding (Tokenization)

To maintain the original model's structure and vocabulary, a pre-trained "gpt2" model is used to initialize a GPT-2 tokenizer during the tokenization process. Then, three special tokens are added to the tokenizer: `PAD` to pad shorter sequences, `EOS` to mark the end of a sentence, and `BOS` to mark the start of a sentence. These tokens must identify the stanzas of a poem in order to preserve sequence length consistency throughout training. The quantity of new tokens added to the tokenizer during training and generation enables the model to comprehend and interpret these newly introduced special symbols.

Model Selection

For this project, we chose to use a powerful language model GPT-2 which is a transformer-based language model developed by OpenAI in 2019. At its core, GPT-2 is an unsupervised transformer language model. Language models are designed to understand and predict text sequences, leveraging probability distributions to forecast the subsequent word in a sequence. They discern and utilize the contextual relationships between words to enhance text predictions. By fine-tuning it on our poem dataset, it can achieve the goal of generating poems for both conditional and unconditional generation.

Model Training

Fine tuning GPT-2 involves iteratively updating the model parameters to minimize the loss function. The process is iterative and typically relies on gradient descent. Here is a condensed version of what a training cycle looks like:

Initialization with pre-trained parameters; for each iteration(epoch), involving forward pass, calculate loss, backward pass, and then update parameters; then repeat until convergence.

We use Adam optimizer to update model weights during training which is a prevailing adaptive gradient descent algorithm. Also, we define key hyperparameters: `learning_rate`, which controls the step size in optimization; `eps`, a small number to maintain numerical stability; and `warmup_steps`, which gradually increases the learning rate at the start of training to stabilize

early iterations. Next, it specifies that computations should be performed on a GPU for efficiency. Finally, it prepares the initial input for text generation by encoding a seed prompt ("**<BOS>**" for beginning of sentence) into a tensor and moving it to the GPU, setting the stage for the model to generate text or continue with fine-tuning.

After 80 epochs, the trained model is saved for generating new poem stanzas.

Poem Generation

Setting the model in evaluation mode signifies that it is now being used to create new content rather than being trained. The model will then use the generate method to produce a new poem based on the original seed text that was supplied. Using a sampling strategy, this method ensures a creative and varied output by controlling the generated text's diversity and randomness through the top_k and top_p parameters. The generated sequences are then decoded from token IDs back into readable text, and the model's ability to mimic the structure and style of the training data it was optimized on is demonstrated by printing out the resulting poem stanzas.

Qualitative Results

Below we present some selected poems generated by our AI Poetica.

1. The Chair is dumbfounded, seeming to slumber; I hear Beside a Creolean lady talking In quaint and quaint tenements, Loud and proud, What music these lovely souls hear; and she, The dame of situations, plays alone The harp on the cell.
The air is heavy with viols, which throughout the whole of the Show are terrific; The viols, which throughout the whole of the Show are terrific, Lie on the banister-organ, some five feet in'd, Broad-shouldered elephants, which always cower Within the banister.

I say these things to myself, To luncheon guests, in the midst of their Being polite and to children;— At the violet hour, when all Is clench'd within the violet walls, All propitious smoke rises In the air, like an incense-wreathed rose; The lady's fingers are sensitized, The foaming mouth, like a lectern, is full Of memories profound.

I say these things to myself, As the evening torches fill My brain, which will burn up

2. The South-wind brings Life with it,— Life with a crust, A clinging mist on its brow.

While Time warms, And the South-wind makes its voice.

The Music swells, And we are sucked in.

The South-wind calls,— "Bard!"— "Courage! come this way!" Our hearts are at their height.

The South-wind makes his art.

3. Asleep, stung by visions, Till the ebbing sea grants one the terror of a lost night; Till, in the calm of some forgotten nest, One poet laments alone The shaft of sleeping Jove's wings, Far from the fiery noon....

I cherish thee thus!

O starless night!

I cherish thee thus!

O starless night!

Stay where!

I cherish thee so!

As I adore thee, Beloved of the Infinite!

To win, to parley With the wintry rabble; To hold, and win, and sink or swim With the tide of history, And at last be man, Earth-bitten!

And live and let live!

4. I bring all I am, certitude, Of power that bring me forth, And make known all I am to say.
To all men at heart, kindred tongue and eye, I say, I am the Avenger, And I am the dawn.
Of old I have stood athwart the lash of the North Star With inflated frenzy, And I will not slumber, Nor whine and moan, Nor
whine and smile.
When each with his momentous art Outshines mine, I sway With the will of the Sprite as through endless swoons, That I give
them speech and joy.
There were but one way to me,——Speak, and I am done.

5. I ask of thee, O Beauty!
To know thy secret, And dive deep in thy depths!
While a star glows at my back, And deep in thy spirit I desire Thee.
Thy form is vain, And thy beauty is shallow.
Thou cannot hide thy sorrows, Nor weep at thy brow, Because I have a prize in my hand, Nor sit by and brood over thy griefs.
Thou delight'st to garnish With gems of wine my whole, And, while my soul prays for thee, I must reign alone.
Thou delight'st to garnish With gems of gold my whole, And when my eyes are full of tears, I weep alone.
Oh, miserable me!
Stoop! and enfold!
Thou delight'st to garnish With gems of gold my whole!

Evaluation Metrics

After generating poems, we have identified key metrics to evaluate our results effectively. The first metric is the calculation of perplexity, a standard measure in our previous assignments. In this approach, we assign a probability of 0.01 to words not found in our poem corpus, ensuring a more realistic assessment of the language model's performance. This method helps gauge how well the model predicts or understands the structure of the generated poems.

The second set of metrics we've incorporated is the Flesch-Kincaid Readability Tests, a well-regarded tool for assessing the readability of English texts. These tests consist of two components:

Flesch Reading-Ease: This test evaluates the text on a 100-point scale, where higher scores indicate greater ease of reading. Texts with simple, short sentences, and familiar words tend to score higher, reflecting their accessibility to a broader audience.

Flesch-Kincaid Grade Level: This aspect of the tests assigns a U.S. school grade level to the text, suggesting the educational level required for comprehension. For example, a score of 8 implies that the text is suitable for eighth-grade readers. The test takes into account sentence length and word complexity, providing insight into the text's complexity.

Employing both the perplexity measure and the Flesch-Kincaid Readability Tests allows us to evaluate our generated poems from different perspectives. While perplexity focuses on the predictability and linguistic consistency of the poems, the Flesch-Kincaid tests offer an

understanding of their readability and accessibility. These tests are instrumental in educational and publishing contexts, ensuring that written materials align with the intended audience's comprehension abilities. Together, these metrics provide a comprehensive framework for assessing the quality and effectiveness of our poem generation model.

By generating 50 sentences and calculating the perplexity, we get the graph in Figure 2.

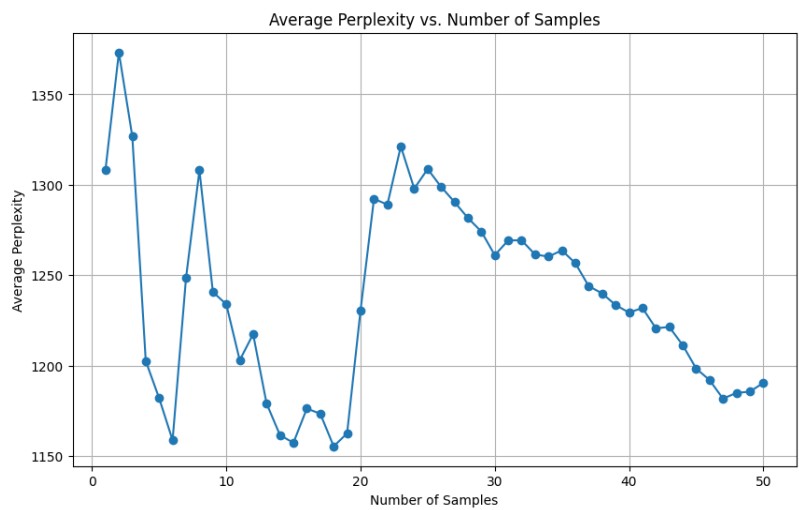


Figure 2: Trend of Average Perplexity Across Generated Poems

A perplexity of around 1200 is relatively high, indicating that the model found the text to be unpredictable or complex, possibly due to the creative and non-standard use of language often found in poetry.

By generating 80 sentences and calculating the scores based on the two tests in the Flesch-Kincaid Readability Test, we get the graph in Figure 3.

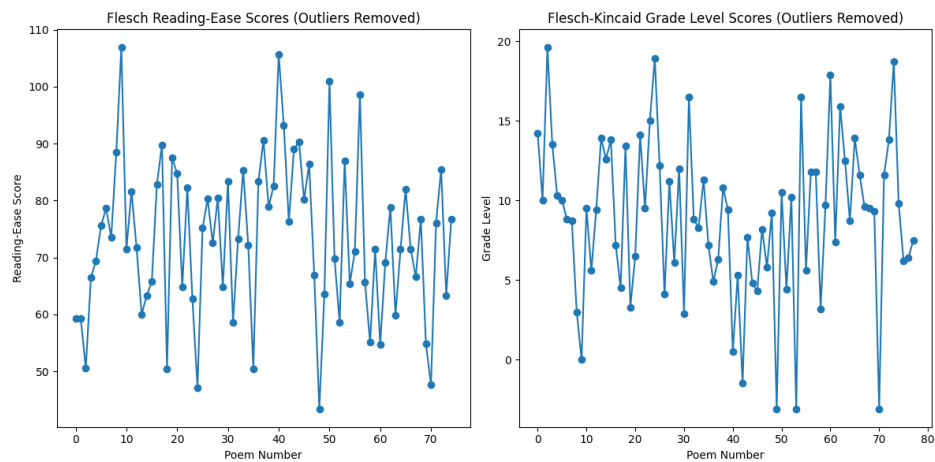


Figure 3: Flesch Reading-Ease and Flesch-Kincaid Grade Level scores graphs

In short, the graphs display a high degree of variability in the readability of the generated poems, which could be indicative of diverse styles and vocabularies being used by the poem generation model. While higher variability in scores could suggest creativity and a wide range of expression, it may also point to inconsistency in the model's output, especially if there is no intentional control over the complexity of the generated poems. If the goal is to maintain a certain readability level, these results could indicate the need to fine-tune the model to produce more consistently targeted outputs.

For human evaluation, we ask for some opinions from our friends since they are the easiest to approach. To conclude their opinions, overall, these poems are rich in language and emotion, each creating a distinct world. Despite their impressive linguistic flair, their effectiveness varies, with some parts being more impactful and coherent than others. Although the poems' vivid imagery and emotional depth are commendable, they might be undermined by occasional obscurity or repetition.

That is to say, we can define our model's performance to be successful.

Discussion

At all, we believe our model is doing a great job in generating poems. It obviously learned the pattern and format of how to write a poem. Although the evaluation score may not look good, that could be because poetry often relies on creative structure, metaphor, and rhythm, which are not always captured in the text generation where those metrics evaluate. That is to say, even with an extremely low evaluation score, the poem itself still can be qualitatively good. Also, unlike informative texts, poetry is highly subjective. What one person considers a masterpiece, another might not understand or appreciate. Therefore, the evaluation of poetry can take human evaluation into account where multiple humans could judge the quality of poetry.

To clarify, we only use the smallest version of GPT-2 which is less than a tenth of the parameters in the available XL version of it. However, it still displays a really powerful ability to generate text which shows plenty of room for improvement.

Future Work

For future work, building on the promising results of our GPT-2-based AI Poetica, we propose to experiment with larger-scale models, such as the more parameter-dense GPT-2 XL version, to potentially enhance the quality and depth of the generated poetry. The larger scale version has more powerful capabilities. The increased capacity could better capture the nuances and complexities inherent in poetic language, offering a more refined understanding and generation of poetic structure, metaphor, and rhythm. Additionally, we plan to explore more advanced training techniques, data augmentation strategies, and the integration of multimodal inputs to

enrich the creative process. Also, we can expand the size of the dataset to include more poem samples for the model to better capture the writing skills.

This expansion, coupled with more sophisticated evaluation metrics and continued human evaluation, will drive our AI Poetica toward a new horizon of artificial creativity, pushing the boundaries of what's possible in the automated generation of poetic art.

Conclusion

To sum up, AI Poetica is an excellent illustration of the fascinating combination of poetry and artificial intelligence. Our GPT-2 model demonstrates the potential of AI to capture and recreate the essence of poetic artistry. It has been refined through the use of a wide range of 20th-century poetry. The resulting poems, with their diversity of language and style, validate the model's skill in creating lines that both explore new creative horizons and connect with the essence of traditional poetry.

Evaluation metrics like perplexity and Flesch-Kincaid scores have provided quantitative insights into the model's performance, revealing a propensity for complexity and a breadth of expression reflective of human poetry. Additionally, subjective human assessments have affirmed the model's ability to engage readers, highlighting areas for further refinement.

Looking forward, we aim to enhance AI Poetica by incorporating more sophisticated models and broader datasets to push the envelope of AI-generated poetry. AI Poetica is not just a technological feat but a gateway to the future of creative writing, promising new explorations of verse in the digital age. We eagerly anticipate AI Poetica's continued evolution and its contribution to the cultural tapestry of poetry.

Extension on comparison of different training datasets

Here are some outputs from model training on the modern poem dataset:

0: let us braid
so as to look at the innumerable multitude
the multitude that is gone before us in all its many forms
and its innumerable acts and ceremonies
in all its various forms
we laugh and cry and wave our hands
and say and remember our departed homestead
and remember thee who was in that place

1: he found it quite certain
that it was truth
of old: he had not sense
of old how bitterly he felt

2: a woman who suffers from cardiac and cardiac disease
has no choice
but to take the lead
in her quest
and fulfill her destiny
i am the only child
who can walk with thee
and see thee alone

GPT-2 models trained on modern and 20th-century data sets possess differences in style, language, and thematic content. Modern poetry model-generated verse examples, more personal, in conversational and free-form, reflecting the less structural, more direct and often more personal tone contemporary poetry takes. The poems in this collection range over multitude of topics and emotions from everyday experiences of life to doubly profound as well as personal revelations- often using plain and simple language. Compared to this poetry-trained model of 20th century, the model inspired in 20th century has a richer, more intricate language and often grander, more expansive themes of perhaps evoking the poetic conventions throughout that century. Verse density increases, along with the use of elaborate metaphors, a broader vocabulary, and more traditional forms as an attempt to draw from the depth and diversity within 20th-century poetics. The dataset to be employed in training the GPT-2 model will heavily influence its output. Each dataset leaves its stylistic and thematic features on the generated text, illustrating the capacity of the model to operate through different poetic eras and styles while being open for differences in nuances..

Personal Contribution

Tianchi Huang: Model Evaluation, constructing report, making visualizations, construct dataset, construct model

Junyi Li: Model training, Data preprocessing, Poem generation, constructing report, constructing dataset