Poem Generation with GPT-2

Alex Kruger, Zhoucai Ni

alexander.j.kruger.23@dartmouth.edu
zhoucai.ni.24@dartmouth.edu

Abstract

GPT-2 is a new Natural Language Processing (NLP) model pre-trained with publicly available text from the internet. GPT-2 implements a deep and complex neural network, with its pre-training set consisting of a massive corpus of text from the internet. It can also use a provided corpus to fit its neural network to a more refined purpose. GPT-2 has revolutionized the way that humans and computers interact, which inspired us to explore how this technology affects the field of poetry. After implementing our own GPT-2 model data from The Poetry Foundation, we found that the poems it generated were entirely comprehensible and even meaningful in some cases. Before the introduction of GPT, it was commonly believed that poetry was one of the purest forms of artistic expression and could never be written by a computer. But this report stands as proof that this is not entirely the case; our algorithm generates poetry that can be analyzed and enjoyed on a similar level to poetry written by real authors. Our results demonstrate that as natural language processing advances, we are forced to challenge our own common perception that the human soul is the only origin of artistic expression.

1 Introduction

Our algorithm aims to challenge the extent to which poetry is exclusively 'human' by generating poetry that is as close as possible to being indistinguishable from real poetry. We are certainly not the first to take on this challenge.

In 2014, researchers at the University of Edinburgh also developed a natural language processing algorithm to write poetry. In their case, they used a recurrent neural network (RNN) to create realistic Chinese poems, meticulously training their algorithm to replicate common traditional Chinese rhyme schemes and tonal patterns. For their training data, they used 284,899 classical Chinese poems from several online resources: Tang Poems, Song Poems, Song ci, Ming Poems, Qing Poems, and Tai Poems. As a result, they found that their generated poetry was close to real poetry but not quite entirely convincing to the human reader (Zhang, Lapata, 2014).

Others have performed similar research, with more of an emphasis on human testing and distinguishing between real poetry and AI-generated poetry. At the University of Southern California, researchers used a recurrent neural network with the assistance of a finite state acceptor (FSA) to generate poetry. They also developed a web interface where users could rate the quality and authenticity of AI-generated poems from one to five stars. Their system used these ratings as feedback into the system to iteratively improve the quality of their poems. Resultantly, they found that in 59% of cases, the users preferred the poems that were generated after feedback was provided (Ghazvininejad, Priyadarshi, Knight, 2017).

Finally, two years ago, the University of Amsterdam used GPT-2, the same natural language processing technology that we used in our project, to generate poetry that imitated that of Maya Angelou. To test the model, they presented subjects with a blind mixed bag of genuine poetry by Maya Angelou and AI-generated poetry. They found that humans were only able to detect the real Maya Angelou poems with 50.21% accuracy, and on average, the subjects reported feeling 62.27% confident about their choices. Furthermore, the researchers tested if the subjects preferred the human-written poems over the poems written by GPT-2 and found that there was a slight, but statistically significant, blind preference for the human-written poetry (Köbis & Mossink, 2021). Perhaps this is evidence that even the best algorithms cannot surpass true human artistic expression and creativity.

2 Methodology

Our machine learning algorithm is trained with a corpus of poetry from The Poetry Foundation. These poems each contain a small set of data including the title, the poem itself, the poet, and a list of tags relating to the poem. These tags are essentially descriptive categories that the poem could fall under, including but not limited to: “Living”, “Parenthood”, “The Body”, “The Mind”, “Nature”, “Trees”, and “Flowers”. Overall, the corpus includes 13,754 poems with approximately 500 that are untitled. There are also more than 300 different poets included in the corpus. We obtained this data from a large csv file that is publicly available on Kaggle (Titor, 2019). Firstly, our algorithm imports this data from The Poetry Foundation into a Pandas dataframe and tokenizes it so that we can quantitatively interpret the data more easily. We chose to tokenize it based on ‘<BOS>’, ‘<EOS>’, and ‘<PAD>’, which mark the beginning of a sentence, end of a sentence, and padding, respectively. Before going any further, we initialize some global variables, including a seed for our randomizers, which will eventually be plugged into our GPT-2 model, and a fixed number of epochs.

2.1 Technical Configurations

The configuration process involves setting up a pretrained GPT2 model, specifically a language model variant, for the purpose of generating a poem stanza. The configuration of the model is set up first using GPT2Config. The vocabulary size is taken from a tokenizer's length, and the maximum length of position embeddings is set to a constant MAX_LEN. The from_pretrained method is then employed to load the pretrained 'gpt2' model configuration, with output_hidden_states parameter set to True to enable the model to return all hidden states.

A model instance GPT2LMHeadModel is created with this configuration, again using the from_pretrained method for initialization, which ensures the transfer of 'gpt2' weights into the new model. The model's token embeddings are resized to match the vocabulary size of the tokenizer being used. This is necessary if the current tokenizer has a different vocabulary size than the one originally used in the 'gpt2' model, or if there are special tokens added to the tokenizer.

The model is moved to a GPU for faster computations using model.cuda(). The AdamW optimizer, an Adam optimizer variant with weight decay, is initialized to optimize the model parameters during training. Total number of training steps is computed for use by the learning rate scheduler, which employs a linear scheduler with warmup. The scheduler's purpose is to increase the learning rate linearly during a warmup period, and then decrease it linearly to 0 over the rest of the training steps. Finally, the model is moved to a specific device (CPU or GPU) using model.to(device) for computational efficiency. This comprehensive setup prepares the model for the subsequent training process, aiming to generate effective poem stanzas.

For this project, we originally set the number of epochs set to one with 1000 steps. However, we ended up dividing the 1000 steps over 5 epochs of 200 steps each and found that it performed worst in terms of average loss and validation. Therefore, we reverted to one epoch because we found this to be sufficient and best for our purposes. If we were to extend this project, we would maybe experiment with the number of epochs and steps, increasing the accuracy of our model and authenticity of our poetry at the cost of computational efficiency and training time.

We use the pandas dataframe of poetry as our training set and randomly separate 80% of the data into a new training set and the remaining 20% into a validation set. These new sets are used to train and check the accuracy of our model before we use it to create new poems. We then fit our GPT-2 model with this tokenized training data. We utilize PyTorch and its provided GPT-2 functionality to accomplish this. Once our model is fitted, we give it a prompt and ask it to generate output. Finally, we use our tokenizer to decode the output, and the result is a poem about our prompt. Along the way, we keep track of various metrics, including accuracy, average validation loss, and total time to train the algorithm.

3 Results

Our project focused on developing and evaluating computational models trained on diverse poetry datasets. The primary model, trained on 13,754 poems, exhibited an average training loss of 0.248, with a validation loss of 1.266. This discrepancy between training and validation losses signifies overfitting, implying that while the model successfully learned the training data, its performance deteriorated on the unseen validation set, indicating room for improvement in its generalization capabilities.

We also experimented with a more specialized model trained solely on the poems of Robert Frost, comprising 28 poems. This model's output was disappointing as it either produced blank poems or reproduced the given prompts verbatim. This deficiency likely stems from the limited breadth and depth of the training data, contrasting starkly with the comprehensive dataset used for the larger model.

For our main evaluation, we distributed a Google form with 10 poems, half of which are randomly selected from The Poetry Foundation.csv file and the other half generated by our model. The participants were not informed about the numbers of artificial and human poems. The generated poems consist of themes such as love, religion, weather, and nature. We gathered a total of 27 google form submissions from anonymous sources as we crowdsourced by asking people in campus wide group chats.

The survey results revealed intriguing patterns. On average, artificial poems were correctly identified approximately 72% of the time, whereas human-generated poems were correctly identified around 78% of the time. An anomalous result was observed with Poem #8, a human-created poem, which was misclassified as machine-generated by 93% of the participants. Poems #3 and #10, both products of our model, achieved around 50% accuracy in identification. These anomalies aside, the classification accuracy was largely consistent across the remainder of the poems.

Poem #8

One day after another—
Perfect.
They all fit.

Figure 1: Human generated poem

Poem #3

I met a girl who lived under my family's roof in an attic.
she told me when I was a child,
I had to go to hell before I got there.
I remember how you made me feel.
We talked and we sang.
We slept.
I knew where you grew up.
I knew where you didn't
I met a girl who was so sad that we couldn't go to the park.
I met a guy who said he had a dream,
that he wanted to go to the woods to see the trees.
I met a girl who was so sad that we couldn't

Figure 2: AI generated poem about meeting someone new

Poem #10

I miss him,
I remember him with a smile
Now I miss his body
Now I hear what it is from the black mourner
I heard what it sounds like
I remember him singing his hymns
My heart murmured to love him
The sound of his drumming
the sounds of his laughter
my heart murmur

Figure 3: AI generated poem about loneliness

Poem #8, misclassified by the majority, is short and lacks a central theme. Poems #3 and #10, both on the theme of love, explore the facets of new encounters and loneliness respectively. This theme may have contributed to their more balanced recognition rate, shedding some light on the perception and interpretation of how love is expressed in poetry.

4 Discussion

The model exceeded our expectations as it's able to generate thematic poems with somewhat of a coherent central idea. The survey results were expected as the artificial poems exhibit some obvious flaws. One noticeable aspect of the AI-generated poems is that the AI generated poems seem to repeat things more often than the human generated poems. One of the poems also added a lot of ellipses, denoted as "...", which makes the poem seem very out of place. Despite their thematic consistency, the artificially generated poems still lack the seamless flow of ideas commonly found in traditional poetry. Furthermore, it's worth noting that, while the model occasionally produces poems with darker, more unsettling themes, it has consistently refrained from generating any content containing offensive language or hate speech. This is likely a testament to the selectivity of the Poetry Foundation and the stringent content filtering applied at their end, ensuring a clean, respectful training set for the model.

Nevertheless, it's essential to bear in mind the potential for latent biases still exist in the model. If the original dataset contains implicit biases, the AI, learning from this data, may inadvertently propagate these biases. This highlights the importance of diversity and inclusivity in our training data. A broader, more diverse dataset would enable the AI to generate a spectrum of poetry that better represents a myriad of perspectives and experiences.

An interesting anomaly was observed with Poem #8, a human-generated poem that was misidentified by 93% of the participants as AI-generated. This poem's length and thematic ambiguity seemed to evoke characteristics associated with artificial generation, indicating that some human compositions can ironically mirror patterns associated with artificial poetry. This result underscores the complexities involved in distinguishing human and AI-authored works and offers a stimulating challenge for future research in this domain.

In contrast, Poems #3 and #10, both AI-generated and thematically centered around love, yielded an identification rate of approximately 50%. The common theme of love in these poems, expressed through the exploration of new encounters and loneliness, might have introduced a level of emotional complexity that made them more like human generated works. The 50-50 recognition rates for these poems could be attributed to the model's capacity to effectively mimic the nuanced and emotive language associated with the human experience of love, making them harder to distinguish from human poems.

Considering the dataset predominantly comprises of English poems, it unlikely that it's able to generate poems in another languages. Poetic devices and figurative language vary across different languages with distinct formats and constraints. In addition, the generation of specific forms of poetry, such as Haikus or Sonnets, would likely challenge the model due to the strict structural requirements of these poetry and the absence of Haikus and Sonnets in our dataset. Furthermore, it is essential to acknowledge that poetry is a high-order linguistic task that relies on cultural nuances, wordplay, and context-specific references. These aspects could be challenging to capture accurately, especially when working with languages and cultural contexts distinct from those in the model's original training set.

5 Conclusion

In conclusion, our experimentation of GPT-2's capacity to generate poetry has offered promising results, yet it has also highlighted the complex intricacies of human creativity and the challenges in recreating it. As we look forward to future works and experimentations, several key aspects and directions comes to mind.

First, enhancing our understanding of the model's strengths and limitations requires further experimentation with the learning rate and the epsilon value for the optimizer. Given more computational resources and time, these parameters could be finely tuned to optimize the model's performance, potentially decrease the validation loss and mitigate the over-fitting nature of our model.

Moreover, diversifying our training dataset to include a broader range of languages, poetic forms, and cultural contexts is crucial for expanding the model's generative capabilities. Different poetic structures, such as haikus or sonnets, and languages beyond English would enable it to generate a more diverse range of poetry.

Lastly, continuing efforts to ensure the inclusivity and diversity of our training data is paramount. Since the model learns and reflects the biases present in its input data, ensuring our dataset is as inclusive and representative as possible will help mitigate the propagation of latent biases in AI-generated content.

Acknowledgments

We would like to thank Professor Rolando and the COSC 72 Tas for providing support and guidance on this project.

References

Zhang, X., & Lapata, M. (2014, October). Chinese Poetry Generation with Recurrent Neural Networks. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 670–680. doi:10.3115/v1/D14-1074

Ghazvininejad, M., Shi, X., Priyadarshi, J., & Knight, K. (2017, July). Hafez: an Interactive Poetry Generation System. Proceedings of ACL 2017, System Demonstrations, 43–48. Retrieved from https://aclanthology.org/P17-4008

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. Computers in Human Behavior, 114, 106553. doi:10.1016/j.chb.2020.106553

Titor, J. (2019). Poetry Foundation Poems [Data Set] https://www.kaggle.com/datasets/tgdivy/poetry-foundation-poems