CS230

GPT-2 for Emily Dickinson poetry generation

Alice Dai
Department of Computer Science
Stanford University
alicedai@stanford.edu

Abstract

In 1985, Steve Jobs gave a talk at the University of Lunds outlining his vision for the future. The following excerpt of what he said about artificial intelligence caught our attention: "My hope is someday, when the next Aristotle is alive, we can capture some underlying worldview of Aristotle in a computer. And someday some student will be able to not only read the words Aristotle wrote but ask Aristotle a question and get an answer" [1]. Fast forward 36 years to now, and Jobs' vision of machine intelligence has come to life, in a sense. In this project we test Jobs' idea of "capturing some underlying worldview of Aristotle" with a famous 19th century poet, Emily Dickinson. Through a technical and literary analysis, we explore a GPT-2 model trained on a dataset of Dickinson's poetry and fine tuned to output new Dickinson poems. We explore the implications of uploading the "consciousnesses" of famous thinkers using deep learning models, and conclude with a brief analysis of our model's generated poetry.

1 Introduction We learned how to use GPT-2 to generate new Emily Dickinson poems. Emily Dickinson was a prolific poet in her lifetime, writing over 1800 poems in her room in Amherst, MA, until her death at age 55 in 1886. Many scholars have spent their academic careers pouring over Dickinson's work, trying to access the beauty of the world she experienced and made material through her poems. For GPT-2 to create "new" Emily Dickinson poems, then, is a significant and perplexing technological innovation.

This project is an experiment to find out what GPT-2 thinks makes Emily Dickinson Emily Dickinson. Is her brilliance replicable beyond the veneer of a well-rendered pastiche? And what is the value in her work if, in a single day, we could train GPT-2 to produce thousands of new poems, enough to drown out the "mere" 1800 Dickinson created in her lifetime? Conversely, what is the value of GPT-2's work if the body of Emily Dickinson's work already exists? Jobs' vision for the future of computing was uncannily accurate, but considering that he did not live long enough to see the world of computing today, would he still take interest in uploading Aristotle's brain onto a computer? What is the value in reviving the deceased if not for a romantic desire for immortality? And if we could ask Aristotle a question today, what would we ask, and how would an answer replace the labor of reading closely, of spending time thinking?

This project is motivated by questions perhaps more typically asked in the humanities disciplines. As such, we'd like to present our work with a care for technical details that will expand into a larger analysis of GPT-2's poetry, analyzing the poems for what we could call the soul of Emily Dickinson. Is she there? And if not, what is? The technical architecture of the project is simple: we trained our model using GPT-2 from OpenAI. We inputted 586 stanzas of Emily Dickinson poems, which outputted new stanzas of Emily Dickinson poems.

2 Related Work A number of papers and non-academic applications have investigated the creative potential of deep learning models. We foresee that as we make more technical progress in natural language generation, our questions will turn more philosophical, and the following survey of work related to our project seems to suggest as much.

A paper published in 2012 titled, "Full-FACE Poetry Generation" explores the early applications of corpus-based poetry generation and claims to be the first paper to achieve poetry generation that satisfies the four standards of the FACE descriptive model [2].

Another poetry generation paper that was particularly well-written came out in 2018 titled, "Deep-speare: A joint neural model of poetic language, meter and rhyme" [3]. In this paper, researchers used a Project Gutenberg dataset of 3,355 Shakespearean sonnets to to output quatrains of poetry that followed specific rules of rhyme and meter. The team trained the data on three joint models: a language model, a pentameter model, and a rhyme model, and showed their results to an English professor for a comparative analysis between real and generated sonnets. The paper concluded that future work should focus less on syntactical rules and more on readability.

Another paper from 2019 titled "Generative Adversarial Networks for text using word2vec intermediaries" [4] attempts to generate text using GANs. GANs were originally applied to generate synthetic images, but this paper proposes GAN2vec, which uses GANs to generate Word2Vec vectors rather than one-hot encoded outputs. The researchers trained their GAN2Vec model on a Chinese poetry dataset and a Coco Image Captions dataset with results that are more proof-of-concept than meaningfully artistic.

Applied deep learning projects are as worth investigating as theoretical papers. For example, Thomas Dimson, former Director of Engineering at Instagram, used GPT-2 to create "This Word Does Not Exist" [5], an app that machine generates new dictionary words. This work stands out from traditionally academic work because it is more focused on creative applications of existing technology rather than developing new frameworks or quantifying the performance of existing ones.

Another example of a creative application of transformers is from Raphaël Millière, a recent Oxford Philosophy doctorate graduate studying the philosophy of mind and philosophy of cognitive science. He used GPT-3 to generate a letter explaining itself to philosophers [6]. It's unclear whether this could count as a self-consciousness, but the fact that GPT-3 can accurately

understand both its talents and shortcomings—in fact, it denies its own consciousness—marks leagues of progress from where natural language generation was just ten years ago.

3 Dataset and Features Our dataset[1] contains 586 stanzas of Emily Dickinson's poetry, totaling to 6794 lines, or 33,378 words. The data was randomly split into training and validation sets, with 498 samples, or 85% of the total data, used to train and 88 validation samples, or 15% of the data, used to validate. Dickinson's poems are famously difficult to transcribe from her original handwriting due to legibility. We encounter in Dickinson a prescient dissent: there is no such thing as a definitive transcription of Dickinson's poetry which makes the matter of training her poems all the more uncertain and evasive. Such is the human avoiding conscription to the technological machine. Dickinson's poems bear a number of characteristic marks that many literary scholars consider to be Dickinsonian—unexpected capitalizations of certain words and a copious use of em dashes are the obvious two to note. The following is an example of what a typical entry in this dataset looks like:

At least to pray is left, is left
O Jesus! in the air
I know not which thy chamber is,—
I 'm knocking everywhere.
Thou stirrest earthquake in the South,
And maelstrom in the sea—
Say, Jesus Christ of Nazareth,
Hast thou no arm for me?

We preprocessed the data by deleting erroneous quotation marks present from the original Kaggle download. We also manually cleaned certain spacing issues between contractions like "'Tis" and "I'm" that also seemed to be errors during transcription. Also notice that Dickinson uses words like "stirrest" that might return as <UNK> tokens to even a large language model, since she liked to use invented words. The dataset is inconsistently delineated by stanza and by poem—there are whole poems that the dataset failed to delineate by stanza—some so the model will not fully learn a pattern about the lengths of Dickinson's poems. However, many of Dickinson's poems were only one stanza long, so the model will have trouble learning the difference between a whole poem and a stanza of a poem. This will explain why our output fails to create proper stanzas of poetry. Because GPT-2 is an unsupervised language model, our dataset is relatively simple, a text file containing poetry with no additional labels. We then tokenized the dataset per stanza using the GPT2Tokenizer from Hugging Face. This outputs a dictionary of input_ids and an attention_mask that we can then feed into the model trainer. We also added a BOS beginning of sequence token and EOS end of sequence token to the beginning and end, respectively, of each stanza so we could create model outputs with just a BOS token as the prompt.

4 Methods We trained a custom dataset using GPT-2 and an AdamW Optimizer, both imported from Hugging Face. GPT-2 stands for Generative Pre-Trained Transformer 2, and it's an

[1] https://www.kaggle.com/sunxyz/gru-based-rnn-writes-emily-dickinson-inspired-poem/data?select=final-emily.csv

unsupervised large language transformer open-sourced by OpenAI in 2019 [7]. A transformer is a deep learning model architecture invented in 2017 that utilizes self-attention to train data [8]. Transformers are the model architecture of choice in natural language processing because they can be initially pre-trained on a large corpora of text then easily fine tuned for domain-specific tasks. GPT-2 contains 1.5 billion parameters and was trained on a dataset of 8 million webpages to predict the next word given a set of previous words. GPT-2 is the successor of GPT and the predecessor of GPT-3, which Open AI released for public use on November 18, 2021, during the development of this project. GPT-2 is capable of a wide range of domain-specific language tasks. Popular use cases include question and answer generation, text summarization, and machine translation. Because GPT-2 was trained on a large majority of the internet, many AI ethicists have raised alarm about the risks and biases of large language models [9]. As large language models continue to scour the internet as its primary source for pre-trained data, issues around privacy and bias will only grow. We also added an AdamW optimizer to our fine-tuned model to adjust the learning rate during training. The AdamW optimizer involves a fix around the weight decay implementation in the Adam optimizer, which stands for Adaptive Moment Estimation. Adam is an optimization algorithm that combines momentum with RMSprop [10]. The purpose of optimization in general is to speed up training and encourage model convergence.

5 Experiments/Results/Discussion We trained our model on a number of hyper-parameters including epochs, learning rate, batch number, epsilon, and sample interval. We chose epsilon to be 1e-8, which stayed the same throughout all of our trials. We experimented with different batch numbers, learning rates and sample intervals and recorded the training and validation loss of our data over epochs. Our quantitative results helped us tune our model to minimize bias and variance, and we got the best results with a learning rate of 5e-4, batch size of 2 and sample interval of 200 over 5 epochs.

Figure 1. Training and validation loss for learning_rate = 5e-4, batch_size = 2 and sample_interval = 200 over 5 epochs

Figure 1 shows the training and validation loss of our best performing experiment. Variance between training and validation increases each epoch after the second epoch. And while training loss reaches a minimum of 0.58, validation loss actually increases 0.93 in the last epoch. We ran into variance issues throughout our experiments, which means that we would benefit from

training our model on a larger dataset. We could also deploy dropout regularization, perhaps at epoch 2 for the model in Figure 1, but doing so would also mean we would always cutoff our model performance before minimizing loss.

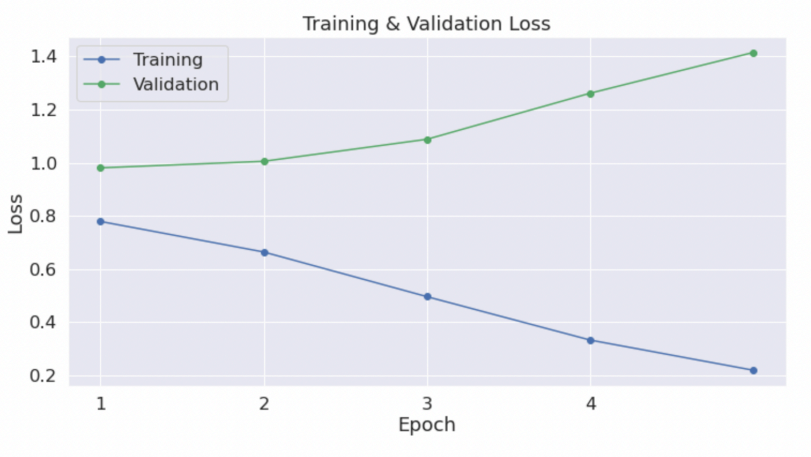

Figure 2. Training and validation loss for learning_rate = 3e-4, batch_size = 3 and sample_interval = 100 over 5 epochs

Figure 2 shows that with poorly tuned parameters, our variance issues were even worse though the training data loss gets as low as 0.22, which suggests that to improve our results from Figure 1 we need to both train longer to reduce bias and train on a larger dataset to reduce variance.

For the qualitative component of each experiment, our model outputted 10 unrelated stanzas of new Emily Dickinson poetry with a BOS token as the prompt. A few notable stanzas are attached in Appendix A to save space.

6 Conclusion/Future Work The task as readers of machine generated poetry is to suspend disbelief. GPT-2 wrote poetry, and only when we take its art at face value can we move beyond the obvious hesitations and critiques. Doing so is an attempt to topple the concept of an author who often distracts how we read a text [11]. Referring to the stanzas attached in the appendix, our model picked up on a number of Dickinson's characteristic punctuations and capitalizations, the capitalized nouns, the em dashes. The setting sun that sounds like fire in Stanza 3, or the image of a brain as a planet in Stanza 4 are provocative images. Such images are where GPT-2 shine artistically. Machine generated poetry seems uniquely suited for free association, a practice in psychoanalytic therapy where humans try to express themselves without censorship from their consciousnesses. Considering the uncertainty around machine consciousness, GPT-2 is a psychoanalyst's dream. Without consciousness, GPT-2 is able to free associate in uninhibited and ultimately inhuman ways, taking data from the internet and making unexpected connections between words or concepts. In a human artist, freedom from the usual trappings of consciousness is highly valued. We call it originality or creativity. In a machine artist, this freedom is criticized because we question if GPT-2 really knows what it's talking about; our own self-consciousness denies the machine's art. However, we think that the value in machine art is to serve as a black mirror. GPT-2 creates associations out of what it learns off the internet, and it has probably "read" works from all the great human thinkers, including Aristotle and Emily Dickinson. Its free

associations within this great library of human thought—a library that also contains offensive language and violence and spam—can be read as reflections of the world today, and perhaps the world that it offers us is uncomfortable to accept or hard to understand. When GPT-2 starts to write hate speech, for example, it is easy to blame the technologists for their ethics, yet GPT-2 has learned to imitate this speech like a parrot imitates its owner. It's simply easier for us to direct blame toward something we consider inhuman, whose machine flaws we believe are fixable through better design and more iteration, than to blame ourselves and our human flaws. Large language models will eventually achieve a level of technical sophistication where we can no longer blame its technical shortcomings for producing text we dislike. Perhaps we have already arrived to this point. In such identity crises, we must turn to philosophy. We should think deeply and think well about what the machine can teach us about ourselves. This is what humans can still do that machines can't.

Our model missed the deepest parts of Emily Dickinson's poetry; that soul of hers is still buried in the ground in Massachusetts. When we push further by reading for deeper meaning in the machine-generated stanzas, they tend to deflate. In our model's poetry, we encounter an unconvincing philosophy. "A Soul is not a Soul, nor the Universe./It was once, or later,/ And eternity, is a hinge." The association between eternity and a hinge interested us the most, but this stanza doesn't say much beyond its word associations. Our model tends to circle around big concepts, like the soul or the universe or eternity, but what it has to say about such matters is decidedly ambivalent ("It was once, or later"). Such writing lacks conviction, it leaves us unmoved. Training our model on a larger dataset with more epochs may produce better results.

For future work we'd like to recreate this project with GPT-3 to compare performance between models. For fun, we would also like to see machine generated art become a new section in a large publication like *The New Yorker*. We would also like to see a large language model trained with just texts from canonical works in philosophy and literature. Could we train an academically specialized version of GPT? Most importantly, we'd like to continue synthesizing our technical understanding of natural language generation with literary analyses. From this project alone, we see that our model never quite captures what Dickinson so masterfully accomplished in her work. We leave with more questions. How could we train a model to capture a thinker's changing philosophy over a lifetime? Can we use the free associations of machine generated art to aid us humans in creating new ideas and new philosophies? What can we learn about ourselves in the challenge to synthetically create consciousness? We hope that there is space for this interdisciplinary work in the future.

7 Contributions This project was inspired by the writer and academic J.M. Coetzee's short-lived career as a computer programmer in London in the early '60s. He produced homespun versions of machine generated "poetry" by programming an IBM 1401 computer to output random words on a page [12]. Some code for the project was adjusted from a Google CoLab tutorial on training

GPT-2 with a custom dataset[2]. Thanks also to professor Joseph Donahue of the Duke English department for his single-author Emily Dickinson class which we took Spring 2019. Lastly, thanks to the teaching team of CS230 for a challenging semester, we learned a lot. Note that this paper was authored by a single person but I refer to myself as "we" throughout.

References

[1] "Steve Jobs Predicts the Future of Computers - Youtube." *Youtube*, 16 Mar. 2015, https://www.youtube.com/watch?v=pWRzfObFG4A.

[2] Colton, Simon, Jacob Goodwin, and Tony Veale. "Full-FACE Poetry Generation." *ICCC*. 2012.

[3] Lau, Jey Han, et al. "Deep-speare: A joint neural model of poetic language, meter and rhyme." *arXiv preprint arXiv:1807.03491* (2018).

[4] Budhkar, Akshay, et al. "Generative Adversarial Networks for text using word2vec intermediaries." *arXiv preprint arXiv:1904.02293* (2019).

[5] Dimson, Thomas. "This Word Does Not Exist." *This Word Does Not Exist*, https://www.thisworddoesnotexist.com/.

[6] "Response to Philosophers.pdf." Edited by Raphael Milliere, *Google Drive*, Google, 31 July 2020, https://drive.google.com/file/d/1B-OymgKE1dRkBcJ7fVhTs9hNqx1IuUyW/view.

[7] Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

[8] Wolf, Thomas, et al. "Transformers: State-of-the-art natural language processing." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020.

[9] Bender, Emily M., et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021.

[10] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).

[11] Barthes, Roland. "The death of the author." *Contributions in Philosophy* 83 (2001): 3-8.

[2] https://colab.research.google.com/drive/13dZVYEOMhXhkXWfvSMVM1TTtUDrT6Aeh?usp=sharing#scrollTo=-p0upAhhRiIx

[12] Roach, Rebecca. “The Computer Poetry of J. M. Coetzee’s Early Programming Career.” *Ransom Center Magazine*, 28 June 2017, https://sites.utexas.edu/ransomcentermagazine/2017/06/28/the-computer-poetry-of-j-m-coetzees-early-programming-career/.

[13] Luksic, Sandra. “Wittgenstein, Natural Language Processing, and Ethics of Technology.” *Duke University*, 2020.

[14] O'Gieblyn, Meghan. “Babel: Could a Machine Have an Unconscious?” *n+1*, 29 Sept. 2021, https://www.nplusonemag.com/issue-40/essays/babel-4/.

[15] Hämäläinen, Mika, and Khalid Alnajjar. "The Great Misalignment Problem in Human Evaluation of NLP Methods." *arXiv preprint arXiv:2104.05361* (2021).

[16] Belloni, Massimo. “Neural Networks and Philosophy of Language.” *Medium*, Towards Data Science, 4 July 2020, https://towardsdatascience.com/neural-networks-and-philosophy-of-language-31c34c0796da.

[17] Millière, Raphaël. “Welcome to the next Level of Bullshit - Issue 89: The Dark Side.” *Nautilus*, 9 Sept. 2020, https://nautil.us/issue/89/the-dark-side/welcome-to-the-next-level-of-bullshit.

[18] Bernstein, Michael S., et al. "ESR: Ethics and Society Review of Artificial Intelligence Research." *arXiv preprint arXiv:2106.11521* (2021).

Appendix A

Stanza 1:
The Soul is not the Soul, nor the Universe.
It was once, or later,
And Eternity
Is a hinge.

Stanza 2:
We are the ones that know no
We do not know a thing,
But are asleep, we are thinking.

Stanza 3:
I thought some things were going to be done,—some things I could have done.
The sun was setting,—so when I first saw it, I thought I had heard—
It sounded like a fire.
I wondered if the grass had started,—
The houses were running—no, they were—
How long would this have been before
But I could not tell—
'Tis morning now, when I first heard the sounds
Of fire in the grass
That started this way.

Stanza 4:
The lower portion of my brain
Upon each side of Eden,
The brain upon its axis,
The brain upon its axis,
One axis withdrawn by itself,
Departing the other,
Contented in one or two adjoining regions
Where duties hold them,
Sagacious as hell to be,
Unmoved as you are,—
There's a revolving wheel
For every atom round,
There's a revolving hand,
An axis withdrawn by itself,
Lest we perish!