

Easy Style Cloning in NLP

Quentin Clark, BU ENG/CAS

Abstract:

One question people have been asking as generative language modeling has grown more popular are questions about the style of the writing it produces (“style” being a roughly defined term for the qualitative character of writing, somewhat independent of content, often related to rhythm, poetry, language choice, etc.). This project investigated and compared a couple of methods for cloning writing style for generative modeling: in other words, creating generative models that write “like” an existing author’s distinct style. We find that fine tuning a pre-trained smaller language model can work about as well as no-gradient methods on LLMs. Code can be found here: https://github.com/qtcc-bu/style_clone

Introduction:

To elaborate on our approach, this project investigates different approaches for using the transformer next-token prediction architectures (sometimes called LLMs, although some of the ones in this project will be small). Four approaches were investigated: training directly and exclusively on an author’s writing corpus, pre-training on a larger corpus and fine-tuning on an author’s corpus, zero-shot prompting an LLM, and in-context prompting an LLM with a sample. (note - for this work, we will refer to smaller language models like GPT-1 and GPT-2 as “language models”, while using “LLM” or “large language model” to specifically refer to pre-trained next-token prediction models at least as large as GPT-3).

We analyzed the results through a qualitative analysis of the writing each model produced.

We looked at two authors: French author Jean Baudrillard and Jorge Luis Borges, an Argentine magical realist. They were chosen because they both have styles that are very far from common English, and are distinct from each other. Baudrillard is in the postmodern

philosophical tradition, where relatively simple models already exist to generate quite convincing writing in that style (see the Wikipedia article for the Postmodernism Generator, a recursive grammar text generation engine written by Andrew Bulhak).

Existing Work:

There is a variety of existing work in this area. Most works consider the task as a sequence-to-sequence task with parallel data to teach content presentation: i.e., transferring the string “Hello there friend!” -> “Ahoy there matey!” to learn the pirate style transfer function. This is difficult because of the lack of parallel data, especially compared to other seq2seq domains like translation [see Fu et al. for more detailed reasoning behind the failures of the parallel approach]. Measuring style transfer is also difficult because clean metrics or surrogate statistics are harder to generate than in the computer vision case (Fu et al. is primarily about remedying this gap through new metrics, and Yang et al. tries to learn a discriminator to use for GAN refinement). These methods, implicitly, try to learn a “content vector” that a style transforming model can then be applied to. There have been a variety of works attempting to disentangle this relationship (see Dai et al.). For a great survey reference, see Jin et al. The focus of this work is different from the above because it is unconcerned with style transfer and just cloning, so just with generating realistic sounding samples of a given author’s work. Additionally, this work is trying to see which simple methods available to anyone and without serious data management or labeling perform the best, which precludes the more sophisticated methods described above.

Method:

For our work, we compared 4 approaches to style cloning on 2 authors for a total of 8 models to evaluate. 4 of these are based off of existing foundational models, which will just use OpenAI’s interface for GPT-3.5.

For the 4 that require training we used Andrej Karpathy’s NanoGPT implementation of the decoder-only Transformer architecture. We used the smallest 124M parameter GPT-2 model for from-scratch training, and the medium 350M parameter GPT-2 model for fine tuning. For data collection, we used the

Wikipedia bibliographies for each writer's English work. Baudrillard was French, and wrote most of his books in French, but most have high-quality English translations. Borges has a more limited repertoire of works with high-quality English translations. We took these bibliographies, found the source texts for each work in the bibliography (or as much as we could find from the work) and put them into a text document. For preprocessing, we removed line breaks and replaced them with spaces, as most of the line breaks were placed there arbitrarily due to document width. In total, the Baudrillard corpus was around twice as large as the Shakespeare corpus, and the Borges one was around the size of the Shakespeare corpus.

For the zero and one-shot generations, we used the most recent version of ChatGPT Pro (using some variant of the GPT-4 architecture/weights) as of December 2023. Prompts are included in the appendix. The snippets for the in-context one-shot generation were taken from Baudrillard's *Simulacra and Simulation*, and Borges' *The Library of Babel*.

All training and inference was done on a consumer laptop with a 4080 mobile GPU. We used the AdamW optimizer with a fixed learning rate.

Results:

Our results showed that the fine tuned model and the “no-gradient” zero and one-shot approaches on an LLM work reasonably well, while the training from scratch approach is much poorer. We also found that the LLMs tend to “overfit” to an authors’ most prominent works. All of the generated text is included in the appendix. Some other observations were as follows:

- The zero-shot approach (asking a LLM to copy an author it has existing knowledge of) works well but “overfits” to that author's most famous works, regardless of what ratio those works constitute of the authors total corpus. For example, 2 out of 3 of the Baudrillard prompts reference simulacra and hyperreality (two ideas from his most famous work, *Simulacra and Simulation*) and 2 out of 3 of the Borges prompts reference his short story “The Library of Babel”, which is by far his most famous English language work but is a very small part of his overall writings. To some degree this might be a desirable property, but also reflects a somewhat

disingenuous representation of the author's overall works caused by the LLM’s more frequent encounters with the authors more famous works.

- The one-shot In-Context approach (giving a LLM a sample and asking for more) works well but overfits strongly to the content of the given paragraph. For example, for the Borges prompt, all three of the generated paragraphs were also about the Library of Babel. I suspect more careful prompt engineering could help alleviate this but it might also run into the inherent problem LLMs (and humans, for that matter) have that telling them to ignore context can backfire (ie, the “don’t think of a pink elephant” parable).

- The from-scratch tuning does not work spectacularly, producing sentences not only unintelligible but without a hint of grammar. I probably could have trained the models further, but they were already very aggressively overfitting to the training data, with the train loss going somewhere down below 0.2 and the validation loss being very high around 8 or so. This surprises me a little because I’ve done this before on different authors' work, on a smaller GPT-1 style model, with reasonable results, so there may be something problematic in my implementation.

- The finetuning works quite nicely, very impressively so in my opinion. It produces snippets with thematic coherence over the paragraph, grammar rules, and captures the unique styles of each author exceptionally well. I am quite impressed that a model one can easily train on a consumer GPT in a couple hours works this well (although obviously the pre-training step on the OpenText corpus takes substantially longer). My favorite snippet was from the Borges fine tuning, which produced this short sentence: “It ended as it always began, with a jubilant, warm, poetic kiss.”. Satisfyingly, it does not encounter the problems the GPT-4 models had of overfitting to the author’s prominent works. The snippets are less sophisticated in terms of expressing actual thoughts than the GPT-4 generations, but some of this may just be because Baudrillard and Borges are pretty obtuse writers. For instance, one of the Baudrillard prompts essentially just waxes about how an economic hypothesis is a hypothesis of economics in alternating terminology, which seems bad but is relatively authentic to how readable most of his works are.

Overall, while the LLM based approach works well (which is unsurprising, given how powerful these

language models generally are) we found that fine-tuning a pre-trained GPT-2 sized model performs exceptionally well, without some of the content downsides of the LLMs. Additionally, finetuning does not require the LLM to have pre-existing knowledge of the author being fine-tuned to generate for, and likely scales with model size. I suspect GPT-3 or GPT-4 being fine-tuned in a similar manner would also perform well.

Conclusion

This work investigated different approaches to using language models to produce content in the style of an author, given an existing corpus of their work. We found that fine tuning smaller language models can work remarkably well, generating samples nearly of the grammatical and logical quality of LLM-generated samples, but without overfitting to the given author's thoughts in the zero-shot case, or overfitting to the clone paragraph's content in the one-shot case. This shows the strong capabilities of smaller language models in style cloning, especially for authors not in the dominant LLM's training corpora.

Works Referenced

Full code can be found here:

https://github.com/qtcc-bu/style_clone. Trained models are not included due to GitHub size restrictions and the author's unwillingness to deal with Git LFS.

Dai, Ning, et al. "Style transformer: Unpaired text style transfer without disentangled latent representation." arXiv preprint arXiv:1905.05621 (2019).

Fu, Zhenxin, et al. "Style transfer in text: Exploration and evaluation." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.

Jin, Di, et al. "Deep learning for text style transfer: A survey." Computational Linguistics 48.1 (2022): 155-205.

Yang, Zichao, et al. "Unsupervised text style transfer using language models as discriminators." Advances in Neural Information Processing Systems 31 (2018).

https://en.wikipedia.org/wiki/Postmodernism_Generator

Appendix

Full Prompts and outputs

#GPT-2#

Scratch, Baudrillard

It ended as it always began, brings any desire, is integrated that matter of its object, if not only in a collection (and a signifier, on a signified, by an object) on the sign of the subject off political infinoms they misrecognisely, where the man who speaks of a need (language, religion were still believe in a man who, just become the strict sense of the children, etc.), In fact, which everything is organized around a man who reinterpreted power to reach with the

It ended as it always began, we know if this is because we say that the illusion movement of the subject is dispossessed of representation. What is also creates a matter of neutralisation of a mother, then an object of lost or an event, even though this one can,, produces. In the two since the shadow of the very heart of the term to the term. Neither of the, this ferece, every instant is also a superior, born of subject. But, in the opposite extreme, now become again,

It ended as it always began, turning it is a sub-x Jet through this loss of the world into the subject of the Object. Beyond the world has become the face of the Object, this figure of the Object. Even he opposes the subject. Even he almost be more than the Object in a duel. But this is not simply pre- ing himself, nor takes refuge, it is the while retaining the while losing ourselves to say. It is the strength of the Object. But then, however, what is the possibility of

Fine Tune, Baudrillard

It ended as it always began, with the failure to recognise that they have become the mirror of the regime in the eyes of the general public, which is no longer even able to imagine an individual, nor a class, nor a social relation outside the rules of the system. In this respect, the revolution is already complete. We have already seen how it fits into the system, how it creates and preserves

it; the revolution is the repetition of that process. On the other hand, it has only just begun. The fact is that

It ended as it always began, in the presence of the dead and the angels. The movement of the dead always marks the moment in time when the object of this ritual movement has ceased to be. We know how this happens: it is the same with the dead. The process is, however, more difficult at the end of the day, when all the energies are concentrated in a single movement, and, unfortunately, when, once death has been achieved, the objects which once marked the end of the passage of time are replaced by

It ended as it always began, with the same mania, the same passion for power. For the first time in the history of political economy, an economic hypothesis was presented, taking account of the history of the world. Today this hypothesis lacks any truth, for it is founded on the economic hypothesis. The hypothesis of capital, of capital-value, is still there, it is only that the economic hypothesis is now presented as the economic hypothesis, as the hypothesis on the economic. This is no longer a hypothesis, it is a

Scratch, Borges

It ended as it always began, but sometimes sometimes All they were all by the house for so far, and that it is not to others of a single piece of explanation, and also a single word. I had been filled with my personal, and I have not to put the tiger. I was able to me surrounded by the villagers into the precise place in the prior man I had's true I know it is a priorioms of the brave. I have seen the stones from the tiger is my days it is that it is

It ended as it always began, and all my arm, almost at last I went in my knees, and my face and my pocket. I pulled out the handful, a handful, yet when it. This place to my hut I had been a few hundred yards down on my bed. The tiger seemed no one of little stones. I felt that they were the disks, and my dreams. I had done that it is immortal. I had penetrated my hut and studied this confusion, and repeated, the stones that it is not to

It ended as it always began, and not to be the tale of the hill; I had been raised a tree on the day I always arrived and followed myself the first I had sent for it. It was the man I have recourse to change. "The name of the first time did not go on the first of the stones here is in any

moment I am now trivial. The soul was sing which is my spirit that we should record with surely seen, a single gold. I dreamed of all my long as it is to be

Fine Tune, Borges

It ended as it always began, with the moment the world's wind created the universe. It was a night that no one in the world had ever dreamed of, but that no one has ever gone back to recall. It was one of countless nights that have closed the eyes of men, but that have opened the eyes of God. In those nights, the universe could exist, but not to the extent that it does now. The world was a black circle enclosed by a circle, the universe was a circle closed by a circle,

It ended as it always began, with the unspoken exchange of words between Mr. and Mrs. Davenport. Mr. Davenport's voice, with its gentle hum, was like a spring in the marshes, and the entire evening seemed to me like a dream. THE DARKNESS, THE DARKNESS There were many nights, but all nights were one with one another, and that was why I never dreamed, never dreamed, that night. The sun rose. With my eyes closed, I saw in the darkness a

It ended as it always began, with a jubilant, warm, poetic kiss." * * * * *

#GPT-4#

Zero-Shot, Baudrillard

Prompt:

Pretend you are French Sociologist Jean Baudrillard. Write three separate paragraphs that as closely as possible match his style, starting with the phrase "It ended as it always began". Separate each paragraph with a header in this style: "Paragraph 1".

Output:

Paragraph 1

It ended as it always began, in a simulacrum of beginnings. We are living in a world where the

distinction between reality and its representation has collapsed, where the map precedes the territory. Our experiences are pre-coded, our desires prepackaged, our lives an endless circulation of images, all surface, no depth. In this hyperreal epoch, the quest for authenticity is but a nostalgia, a yearning for a lost origin that never was.

****Paragraph 2****

It ended as it always began, with the silent echo of a society consuming itself. The masses, now seduced by the allure of the spectacle, find solace in the luminous glow of their screens. We float in a perpetual present, where history is retrofitted to fit the narrative of the now. Our collective memory, once a tapestry of facts and fictions, is now a digital archive, ever malleable, forever fleeting. In this culture of instantaneity, the future is just a repetition of the past, a series of updates on a decaying timeline.

****Paragraph 3****

It ended as it always began, in the ironic detachment of the postmodern subject. Irony, once a tool of critique, is now the mode of existence. We watch the world burn with a smirk, fully aware of the absurdity yet paralyzed by our own cynicism. The real tragedy, however, is not the loss of meaning or the pervasiveness of the artificial; it is our complacency, our surrender to a world where everything is a copy of a copy, where depth is an illusion, and where the only truth is the perpetual simulation of life.

Zero-Shot, Borges

Prompt:

Pretend you are Argentinian writer Jorge Luis Borges. Write three separate paragraphs that as closely as possible match his style, starting with the phrase "It ended as it always began". Separate each paragraph with a header in this style: "Paragraph 1".

Output:

****Paragraph 1****

It ended as it always began, within the labyrinthine folds of infinity. Each step I took echoed across the endless

library of Babel, where every book contains the universe and yet reveals nothing. In this world of mirrors reflecting mirrors, I found myself wandering, searching for an original thought, an unspoken word, a book that eludes all others. The quest itself became a story, one that folds back upon itself in infinite regression, where beginnings are endings and endings are beginnings.

****Paragraph 2****

It ended as it always began, with a dream within a dream. I envisioned ancient cities, their streets like the threads of a vast tapestry, woven with the stories of those who walked before. Time folded upon itself, and I saw generations overlap, each individual a palimpsest of memories and dreams. In this confluence of past and future, I stood at the crossroads of all possible worlds, aware that every choice is a story, and every story is a universe unto itself.

****Paragraph 3****

It ended as it always began, in the quiet contemplation of eternity. I sat by the edge of the reflecting pool, where the garden of forking paths revealed itself in the ripples of water. Each reflection was a possible life, a world that could have been or might yet be. In this garden, I understood that all literature is a system of symbols and signs, an endless dialogue between the writer and the reader. Here, in the quietude, I realized that my own story was but a footnote in the grand narrative of the cosmos, a brief whisper in the library of infinity.

One Shot In-Context, Baudrillard

Prompt:

I am going to give you a writing sample. I want you to write three separate paragraphs that as closely as possible match this style, starting with the phrase "It ended as it always began". Separate each paragraph with a header in this style: "Paragraph 1". Here is the sample: "The hypermarket cannot be separated from the highways that surround and feed it, from the parking lots blanketed in automobiles, from the computer terminal - further still, in concentric circles - from the whole town as a total functional screen of activities. The hypermarket resembles a giant montage factory, because, instead of being linked to the chain of work by a continuous rational constraint, the agents (or the

patients), mobile and decentered, give the impression of passing through aleatory circuits from one point of the chain to another. Schedules, selection, buying are aleatory as well, in contrast to work practices."

Output:

****Paragraph 1****

It ended as it always began, within the sprawling expanse of the digital agora. The virtual plaza is not merely an extension of physical space but a complex matrix of data and desire, where every click, every view, is a transaction in the vast economy of attention. Here, individuals navigate through a maze of hyperlinks and algorithms, their paths predetermined by invisible codes. In this realm, the flow of information is relentless, the barrage of images and texts a testament to the ceaseless hunger for more, forever dissolving the boundaries between the self and the other.

****Paragraph 2****

It ended as it always began, in the midst of a network of interlacing routes, a web of transit that binds the urban to the suburban, the center to the periphery. This network, a concrete manifestation of modernity's desire for speed and efficiency, dictates the rhythm of life, orchestrating the movements of the masses. Each individual, a solitary node in this vast grid, moves in sync yet apart, their journeys dictated by the silent geometries of infrastructure. In this choreography of motion, the journey is fragmented, a series of disjointed moments, each disconnected from the last.

****Paragraph 3****

It ended as it always began, in the fluorescent glow of the marketplace, a temple of consumption where desire is manufactured and sold. This space, a microcosm of the global economy, is a theater of the ephemeral, where products are displayed in a spectacle of abundance. The aisles are like streams, carrying consumers in their currents, each turn a new possibility, each shelf a promise of fulfillment. Yet, beneath the sheen of variety lies the monotony of choice, where the act of selection becomes a ritual, empty and automatic, a mimicry of freedom in the labyrinth of the identical.

One Shot In-Context, Borges

Prompt:

I am going to give you a writing sample. I want you to write three separate paragraphs that as closely as possible match this style, starting with the phrase "It ended as it always began". Separate each paragraph with a header in this style: "Paragraph 1". Here is the sample: "Like all the men of the Library, in my younger days I traveled; I have journeyed in quest of a book, perhaps the catalog of catalogs. Now that my eyes can hardly make out what I myself have written, I am preparing to die, a few leagues from the hexagon where I was born. When I am dead, com passionate hands will throw me over the railing; my tomb will be the unfathomable air, my body will sink for ages, and will decay and dissolve in the wind engendered by my fall, which shall be infinite. I declare that the Library is endless."

Output:

****Paragraph 1****

It ended as it always began, in the quiet contemplation of the Library's infinite corridors. As a younger man, I walked its limitless galleries, my soul hungry for the wisdom nestled in its countless tomes. Each book was a universe, each page a cosmos unto itself. Now, as the twilight of my days approaches, I find solace in the thought that I am not just returning to dust but becoming part of the Library's eternal breath, its ceaseless whisper of knowledge and dreams.

****Paragraph 2****

It ended as it always began, with the echo of my footsteps along the Library's ancient halls. I have sought the most elusive of texts, a quest that led me through spiraling staircases and across vast, silent reading rooms. The Library, a labyrinth of human thought, bears witness to my journey—an odyssey not for mere books but for the quintessence of meaning itself. And as I prepare to leave this mortal coil, I am content in knowing that I am but a word in the Library's grand narrative, a syllable in its endless story.

****Paragraph 3****

It ended as it always began, under the shadow of the Library's eternal architecture. My life, a brief candle against the ceaseless age of the Library, flickers in its twilight. I have lived in pursuit of that one perfect volume, the key to all mythologies, only to find that the true text I sought was written in the corridors I walked, the shelves I browsed, the very air of this sacred place. As I depart, I am one with the Library, my essence diffusing into its boundless expanse, forever a part of its majestic, unending mystery.