

Let Me Collect Myself

IMA Capstone Project by Max Chu

Abstract

This project investigates how the development of machine learning models changes the labor relations of artists with their art. There's a new type of labor that exists within image generation models that is distinctly unique and distinctly invisible compared to art practices of the past, and due to the newness, the language and modes to critique image generation is still in the process of being developed. Therefore, this project aims to participate in this development of language through interrogation of the implicit, "invisible" labor conducted in order to create data sets with which machine learning models are dependent on. Many of the foundation models that are used as the basis of ML image generation are trained on data sets with unattributed sources, such as ImageNet, or data sets compiled by Common Crawl. Further, much of the essential tagging of the data in the sets is performed through crowd-work, using such services as Amazon Mechanical Turk. This disambiguation of labor results in a system that is difficult to audit, and therefore, difficult to critique. In response, the goal of this project is to present a data set created entirely by one individual, in order to create a framework with which critique may more easily occur. The paper aims to provide documentation of the work, as well as supplementation of language and history.

Glossary

- CNN - convolutional neural network
- COCO - Common objects in context - dataset developed by Microsoft

- Common Crawl - dataset hosted by Amazon
- Foundation Models - “broad data (generally using self-supervision at scale) that can be adapted to a wide range of downstream tasks,” (Bommasani et al, 2022)
- LLM - language learning models
- ML - machine learning
- NLP - natural language processing

Introduction

Let Me Collect Myself is an investigation into the human labor required to construct a machine learning model. The project aims to deconstruct such labor by integrating as much of the work as possible onto one individual.

Why care about ML?

The process of ML is inherently reflective of values the populist zeitgeist is upholding. While seemingly an overly broad statement, the goal of much of ML is to catch as much of a variety of individual experience as possible into a centralized, deployable tool. The easiest way to understand this to be true is to look at the datasets that support public image and language generation models. The primary example is the use of a dataset called Common Crawl. Common Crawl “contains petabytes of data, regularly collected [from the internet] since 2008.” Since ChatGPT-3 uses this dataset in training, ChatGPT’s results can be understood to be a direct interpretation of the dominant, crowdsourced internet culture. The issue is, while internet culture may contain elements reflective of a global zeitgeist, such reflection is not equally represented. “In the case of US and UK English, this means that white supremacist and misogynistic, ageist, etc. views are overrepresented in the training data, not only exceeding their

prevalence in the general population but also setting up models trained on these datasets to further amplify biases and harms,” (Bender et al. 2021).

Unique qualities of Image Generation models

Like all ML models, image generation models are generally responsible for “compound(ing) existing inequities by producing unfair outcomes, entrenching systems of power, and disproportionately distributing negative consequences of technology to those already marginalized,” (Bommasani et al, 2022). Further, to compare image generation models to Language Learning Models, image generation models are uniquely subjected to the process of tagging, which adds an element of abstraction to the training datasets. A seemingly innocuous task, the process of tagging images consists of an individual identifying any explicit features in an image and inputting that data. When the fact that many foundation models require, at minimum, 10,000 classified images per variable, the task becomes distinctly more tedious. Much of this type of classification labor is outsourced to locations with looser labor laws than the US, using 5iverr-esque, mass gig based services such as Amazon’s Mechanical Turk. Further, services such as Amazon’s Mechanical Turk takes a noticeable lack of effort in identifying the laborers, thereby disenfranchising the workers to an such a great extent, unique to such labor.

Accreditation of data

Another aspect of the invisible labor involved in producing foundation and ML models is that of the production of the image to begin with. Nearly all of the images are the result of Common Crawl-esque internet scraping. Most of the photographers who developed the images aren’t aware that their images are being used for training, nor have they been notified that that might be possible, or given the option to opt out. In another step of abstraction, many of the image hosting sites, such as Flickr, have only recently

become aware that their images were scraped to begin with. A project revolving around this lack of accreditation is detailed below.

Discussion regarding project data collection

The project began with a sense that, in order to ethically collect data, there must be a meaningful and consensual exchange between parties. This philosophy morphed throughout the project, if only in that I never figured out how I was going to carry around some sort of material reward, and instead became content in an educational discussion, and entertainment. I considered custom, wooden coins, but alas, I am but human and am on an undergraduates' time frame/budget. However, there are explicit rules that I have not intentionally deviated from:

1. it has to be in person, face to face. the goal of the project is to connect with people and learn. further, this provides some sort of potential education-angle. further, this provides some sort of ethical, consensual angle
2. shoulders and up
3. Write in one color (decided on red)
 - a. This is in order to help a future model train on more consistent data.
4. on the back, offer the artist to sign their name, but don't push that idea, it's not necessary
 1. additionally
 1. the portrait number
 2. the date
 3. the time limit imposed

Over time, the only thing in the parameters that shifted was the requirement of the portrait number, in order to keep track of how many had been produced. Oftentimes, though, this number was added retroactively and relatively arbitrarily.

Related works

Myriad: Tulips, Anna Ridler

This is the original inspiration of this project. The expression of detailed, democratic labor was just masochistic enough to warrant investigation. Further, the result of the project seemed frustratingly dispersonal, thereby existing narrowly within a parameter of awe and subtle recognition, while never leaving that site for the choppy waters of education and change. The project was excessively elegant. The work of a machine was merely replicated. I envisioned a further step, into a realm that only a conscious, individual person could complete.

Humans of AI, Phillipp Schmitt

This is the project that proposed context to the idea generated through *Myriad: Tulips*. *Humans of AI* consists of three projects in discussion with the COCO dataset. The title piece, *Humans of AI*, had that same elegance that *Myriad: Tulips* had, while also including a finger pointed at an injustice. The project attempted to confront large amounts of data, and translate those impossible numbers into something comprehensible, something individualists might care about.

Methodology

Discussion of the practical project

The tangible, explicit goal of the project is to collect 10,000 images of myself. These images would all have to have the unique quality of being produced by someone other than me, and other than anyone else who had participated in the project prior. With that tangible goal in mind, I began developing strategies that would allow myself to collect the portraits in a reasonable amount of time, as well as work the skill out as to how to walk up to strangers. What follows are

the excerpts of the documentation of the project, within the first days and weeks when the project was in the most flux, as well as an entry from a much later stage in the project's lifespan.

may 20th

[s] said to me something along the lines of, in order to make this project work, you're going to need to get rid of some of the...[parameters, limiting factors, restraints, preexisting biases]? something along those lines. i had already been going down a deep spiral of panic, surrounded by people trying to get me to ask the drinkers outside of happyfun if they would engage. aristotle even offered to go over with me, which I can only understand as generous in hindsight.

the scale of the thing is terrifying, and the terror is finally seeping in. i feel violently awful, like i'm letting down every single person who i've roped into drawing me. i don't see this feeling going away, but instead growing. slowly, i will take on the weight of all the demands i've made of the people around me and i will snap like a branch in the wind. i've done this to myself. i don't mind. i feel like i knew these feelings would come, and that i've started this project in order to give myself a reason to feel deprecating. otherwise, i would do nothing. that's how it always is...

may 30th

its the two week anniversary of the beginning of the project and i think i have done maybe 35 portraits. moral is low, i think. my sister had three of her friends over today, and instead of interrupting their hangout in order to secure three more portraits, i didn't and let them be. to be clear, it would have been easy and unobtrusive to do so, but i didn't. something about the breadth of the project is allowing me to compartmentalize that there will always be more opportunities, and therefore it's not worth risking anything of my own, nor bothering people when they don't have to be bothered. with this mentality, i'll never get anything done.

I have more ideas about the project, collected through conversations with others. firstly, i decided at the omakase restaurant with my uncle that i will not ask service workers who are currently working. i think that it feels manipulative, as all things are, within the context of capitalistic corrosion and the need for one to earn tips from ones job.

secondly, i think that i want to incorporate an educational element into the project, that i want to use the project as a platform to tell more people about more realistic stakes of ai and machine learning models. i need to be more educated on the facts, to be quite honest. maybe i can make a pamphlet to pass out, and that can be some sort of exchange for participating in the project.

i want to make a website, or maybe an instagram to post all the photos, so that people can look back and review what other people have done. i'm doing this because [e] had a really good time looking through the history so far and comparing the results. this sort of acts as vindication to one's own drawing skills.

i want to keep more diary entries.

there's also a sort of scary element of other people feeling the need to keep me accountable. maybe i should let them...

- [] make website
 - [] with a diary section, with date and piece #, but without name
 - [] with education and further readings
- [] make pamphlet
 - [] doesn't have to be physically printed until back in nyc, but designed so that

july 6th

its the earliest part of the sixth, and there's gotta be something to be said about the gruesomeness of the data entry portion of this project. people who work in such an industry professionally must have such sharp minds. it's like taking my thoughts and sharpening them on a diamond block. the more i work, the more quick and witty i feel, and so therefore the harder the work becomes. breaks and/or drifting thoughts only worsens the final product. it's a cycle of irony.

so far, in order to get through it, im working in batches of 10. 10 formatted on photoshop. 10 uploaded to ig.

july 9

there's a guy named [J] in wash who sits at a table with a sign that reads "what do you believe?" [q] convinced me to talk with him, after her talking with him, and i am grateful she did. [J], after hearing about my project, told me that the most important part is to stay curious and keep trying different methods. he derives this advice from home telling him what i value lost about the project is the conversations that are happening between people. i care about the variance in people's reactions and how they approach my stagnant parameters. today also i made a sign that reads PORTRAIT TRADE: YOU DRAW ME, AND I'LL DRAW YOU (free...obvi). through this sign i got one new drawing, from a straight couple that was visiting new york. the woman was working for a high end makeup brand, interning, and she wasn't satisfied with the work, as she was just making powerpoints.

july 15

cataloged the back logged drawings sitting with remi (*a cat*). we're up to 111 portraits so far. last night, i went to bootleg with [j] and [h], and in between rounds of pool i managed to talk to a pretty substantial amount of people. due to pool's inherent social aspect, it facilitates the level of

trust im looking for in order to ask someone to do a drawing. the fellowship and camaraderie allows people to exist on the same page that a dry walk-up doesn't necessary guarantee. further, [j] and [h] were playing an excruciatingly slow game of pool, which was to my advantage, as new people came to put their names down on the board, the sharks had nothing to do except for look at the pool table. then i struck. then i got portraits. i had to rip paper in half. i drank beers and smoked cigarettes and got portraits. the motivation is becoming more smooth, streamlined, and therefore natural. i'm developing an apathy towards those around me, with the only care the ends. or perhaps a better way to put it is that the portrait project is my social armor. it makes me strong, reflects rejection and allows connections. [J]'s prediction is coming true. the project is abstracting itself based on its needs.

february 19

the time to start taking the project seriously has come. sam heckle emphasized the importance of the discussion of methodology as in reference to the paper, and so i think it would be important to start recording the thoughts that i have when collecting the datapoints. after all, the project focus is one of labor and effort, and so i need to be able to convey what kind of labor and effort goes into the collection of the stuff, because it is not insignificant.

much of the tiring aspect of the project comes actually in the time after collection. there's a certain obligation that i feel in which i have to tell the person about the purpose to the project, and then field questions and suggestions that they might have. while in a good mood, this part can be as enjoyable as any other social interaction. however, when not in the mood, this is arduous and slow, and makes me confront parts of myself that i don't like confronting that much. for example, many of the suggestions about how to perform the task, such as post on social media, or make a poll, or like poor [] suggested, get an ai model to produce images for me, i have to take such an academic lens in order to justify my reasoning not to that it feels

exhausting. further, i have to do so in such a way that would ideally not be perceived as me acting too demeaning to the individual and their (perfectly valid) ideas, as the annoyance comes not from the quality of the idea itself, but from the repetition from hearing the same idea over and over. maybe that makes sense

i would also like to discuss the scenarios in which i meet people and ask them to do the project. while the project has been on hiatus for some months now, i do intend to restart the project in the coming week. this brings into the idea that many of the portraits i have received were received at parties and at bars. this is due to the fact that i really don't interact with new people in my life outside of these two environments these days. previously, this might have been a problem, but i have decided to not worry about it and just let it be with whatever number i have going forwards. the technology will catch up to whatever number of portraits i have by the time the project is due.

some notes about categorization and labeling:

i recently went through and categorized everything i have so far, and i have 229 right now.

this bleeds into the labeling discussion, which is to say the labeling has fallen off. all that's left is the individual's name and number. however, i'm about to go to a tech conference in march with high schoolers, and i think it might be more ethical to give them the option to put their name or not.

i think the lack of labeling shouldn't be a problem that i worry about going forwards. the number is the most important part, followed by a situational, personal flourish, such as the person's name, or in the case of a few of them, the fact that i wrote "at julia's birthday" on the back.

the tech conference:

i need a working model by then. ideally, one that can be modified on the spot, so that i can display the “back-end” in a live space. i am performing even more than i was before, but that was always going to be the case. i’m used to it now.

Discussion of 10,000

At the beginning of development, I read something somewhere along the lines of “the minimum number of data points necessary to train a CNN was 10,000 images.” Based on that understanding, when development began for a machine learning labor related project, the number of portraits was obvious to me. For 10 months, I conducted the project with the anecdote that 10,000 was the ultimate goal. Then, when it came time to make a presentation, Sam Heckle, my capstone advisor, suggested that I directly quote the source of the CNN-10,000 fact. This is when the reality hit that I couldn’t find where I had read the statistic about 10,000. I looked for as long as I had time for, with no results. Many papers had used 10,000 as a subset for a number of test images when developing CNNs (LeCun, et al.), but when 10,000 showed up, there was never any citation as to why that was the number determined for the minimum.

Finally, I went back and looked at the projects I had been using as inspiration during the devising process, and I found something interesting: Anna Ridler’s *Myriad (Tulips)* was based on 10,000 photographs that she took of tulips. Frustratingly, Ridler doesn’t cite why she decided on 10,000 tulips, although she does mention some other numbers that come from informed places: “The iris flower dataset, created by British statistician Ronald Fisher, contains 50 samples of 3 different irises and is used as a test case for many statistical classification techniques in machine learning,” and “...this is in direct opposition to a dataset such as ImageNet (a canonical dataset often used in computer vision), where there are only around 1,000 tulips out of its fourteen million images, all which have been anonymously found, categorized and labeled.” However, 50×3 isn’t 10,000, and nor is 1,000. Perhaps Ridler, like

me, heard the information about 10,000, and she herself did too much work before deciding to question her core principle.

Regardless, the project remains to be finished, and so there is plenty of time to figure out the truth. I have emailed some old professors with the question, and should frankly try and track down Anna Ridler's email, to see where she got her information. If you're reading this and you know, please reach out to me at maxchu@gmail.com, and I will send you \$10 American dollars, in the circumstance that I haven't already found the information.

Peer Review

The goal of peer review was to determine whether or not the goals of the project were conveyed through the actuation of the project. More specifically, whether or not those who I engaged with *Let Me Collect Myself* came away with a greater understanding of the potential issues of generative MLMs, as well as a premise of the extent of the invisible labor that creating datasets contains. Overall, I believe that these goals were achieved, although only within its possible context. I determined this through open forum discussion with a group of 4 individuals, in which I had them participate in the project without knowledge they would be involved in a discussion, and then prompted a discussion afterwards. This covert method was to simulate a standard environment of collection in which I as the performer and collector would give information regarding the project freely.

One participant in the discussion mentioned that she "got stressed out when you explained the project," and that, while 10,000 may have been a difficult number to imagine, "just thinking about the time it would take to get 10,000 portraits is a pretty effective way to convey the labor." There

was then a discussion about how to calculate exactly how much time the project might take.

This sentiment was shared with the rest of the people who I approached for evaluation.

Evaluation

The most tangible result of *Let Me Collect Myself* that I can outline at this time is the overwhelming human element. While I anticipated the fact that the individual nature of engaging with people would lengthen the time the project took to complete, I couldn't have anticipated the degree to which I would depend on that individuality in order to move the project forwards.

To place the project's ceiling at 10,000 images allows *Let Me Collect Myself* to become a particularly brutal type of effort sink; the more effort I put in, the better the results. There is no top end to that ceiling, and in that way, the act of collecting data further reflects the standard process of digitally compiling datasets. Especially early on in the process, when my own expectations and parameters for the project were still fluctuating, I continuously returned to this feeling of desperation. People blurred into data points, sitting on the train or waiting in the elevator became stressful, due to the mentality that I had wasted the opportunity to collect their portrait.

In this headspace, I found a sort of peaceful solace in the preamble and preface that comes with talking to people. I found that people like to talk while they're drawing. They like to preface with "I'm not an artist" and like to show off their final product to whomever is around them when they're done drawing. They like to ask me questions, and I like to ask them what they liked to draw. Most people have signatures they sign their portraits with, rather than the plain script of handwriting. In these small differences, I drew strength to continue *Let Me Collect Myself*.

The true evaluation of the project looks like this: what work is automated by machine both eliminates the human delight of creation, and any delight that might come from effort. The goal of the project was to create a space where the invisible labor behind machine learning models can be discussed, in an effort to acknowledge the work done. Through the work on *Let Me Collect Myself*, I personally became deeply familiar with the weight of that labor, and through this deeper understanding I believe that I have become more capable of conveying that idea to others.

Conclusion

The project as devised about a year ago has not been completed. However, the project that stands now is different in a handful of priority-driven ways. Firstly, the project had an implicit air of masochism embedded in the work. That masochism, derived from scale over time, had been somewhat intentional, with its purpose as metaphor for the immense amount of labor that data sets take to devise for technical function. This premise of masochism was also prompted within the context of my understanding of performance art and protest: two forms that draw from pressure to express fundamental messages; from John Cage and Marina Abramovic to the protesters of BLM in 2020 and anti-genocide Columbia and UCLA encampments, art and action that made waves and made change were predicated on timing and scale.

However, living in the project for the last year has made me come to the understanding that I am not a masochist. There are models of changemaking that are based on the dissemination of information and developing forms for accessibility. In the past year, I visited third place collections such as the Interference Archive and the Cybernetics Library, and always came away with more tools for change.


The core of the project has always been in the fact that I believe not enough people know the level of labor that goes into the LLMs and image generation models they use. The greatest joy I have had is teaching people about what I know, in both informal and formal contexts. I came to the realization that, while there is a certain, primal understanding that one can glean from watching another push themselves, there is benefit in taking one's time. I believe that, in my case, the metaphor that is represented by scale over time (10,000 images in 1 year), can equally be represented in time over scale (how long does it take someone to collect the minimum amount of data to train a CNN?).

Further, I have begun experimenting with alternate methods of message dissemination. I recently got the opportunity to present my project at a high school STEM fair, and for the students I met there, I designed a zine made up of data that I had collected. This zine represented a new path forwards for the project, as well as a second life outside of the context of development as a BFA capstone project.

Let Me Collect Myself has changed from a data set collection project into a system with which I can both talk and do research into the inequity perpetrated by machine learning. The project represents a desire to collect ideas and information, and is a place to sift through massive amounts of information and misinformation in the world regarding a topic that corporate interests would not want seen from any other angle than straight on. The project encompasses not only the development of the data set, but the discussion surrounding the data set, the bibliography for this paper, and the zine made with the intent of a specific (high school) audience. The project will grow from here, I am sure, until such time that I collect 10,000 portraits. Then everything will be put to bed.

Sometimes I try and differentiate the work that I do here and the general premise of having a hobby. The line gets blurry sometimes, but sometimes, it's sharp and clear. Regardless, my passion for engaging with such a topic has not diminished. So I will collect myself for a while longer yet.

Bibliography

- “A People’s Guide to AI — MIMI QNUQHA.” *MIMI QNUQHA*,
<https://mimionuoha.com/a-peoples-guide-to-ai>. Accessed 1 Mar. 2024.
- Aaron Li [@AaronGLi729]. “ChatGPT Calculates How Many Grams of CO2 Your Conversations Generate. <https://t.co/K06cotvyVe>.” *Twitter*, 6 Mar. 2024,
<https://twitter.com/AaronGLi729/status/1765252022012711154>.
- “Anatomy of an AI System.” *Anatomy of an AI System*, <http://www.anatomyof.ai>. Accessed 1 Mar. 2024.
- 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*
- , Association for Computing Machinery, 2021, pp. 610–23.
- ACM Digital Library*
- ,
- <https://doi.org/10.1145/3442188.3445922>
- .
- Bender, Emily M., and Alexander Koller. “Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data.” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky et al., Association for Computational Linguistics, 2020, pp. 5185–98. *ACLWeb*,
<https://doi.org/10.18653/v1/2020.acl-main.463>.
- Betker, James, et al. *Improving Image Generation with Better Captions*.
- Caesar, Holger, et al. *COCO-Stuff: Thing and Stuff Classes in Context*. arXiv:1612.03716, arXiv, 28 Mar. 2018. *arXiv.org*, <http://arxiv.org/abs/1612.03716>.
- Gradient-Based Learning Applied to Document Recognition*.
<http://yann.lecun.com/exdb/publis/pdf/lecun-01a.pdf>. Accessed 29 Apr. 2024.
- Hovy, Dirk, and Shannon L. Spruit. “The Social Impact of Natural Language Processing.” *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, edited by Katrin Erk and Noah A. Smith,

Association for Computational Linguistics, 2016, pp. 591–98. *ACLWeb*,

<https://doi.org/10.18653/v1/P16-2096>.

“How to Comprehend Incomprehensibly Large Numbers.” *Gizmodo*, 26 Feb. 2014,

<https://gizmodo.com/how-to-comprehend-incomprehensibly-large-numbers-1531604757>.

Humans of A.I. <https://humans-of.ai/editorial/>. Accessed 1 Mar. 2024.

Irani, Lilly C., and M. Six Silberman. “Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk.” *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2013, pp. 611–20. *DOI.org (Crossref)*,

<https://doi.org/10.1145/2470654.2470742>.

McGowan, Jamie. “AGI, AI, DL, ML... What’s the Difference?” *The Startup*, 27 Feb. 2021,

<https://medium.com/swlh/agi-ai-dl-ml-whats-the-difference-cfdf749667c9>.

“Myriad (Tulips).” *Anna Ridler*, <http://annaridler.com/myriad-tulips>. Accessed 1 Mar. 2024.

On the Opportunities and Risks of Foundation Models - ProQuest.

<https://www.proquest.com/docview/2562092492?parentSessionId=M2dVT0NL7rgNFJdUsnfUcrgXI9m7WTg6Y9K4oSdC%2Fk8%3D&pq-origsite=primo&accountid=12768&sourcectype=Working%20Papers>. Accessed 1 Mar. 2024.

Stalder, Felix, et al. *Hyperemployment – Post-Work, Online Labour and Automation*. NERO, 2020.

“The Research Origins of Stable Diffusion | Runway Research.” *Runway*,

<https://research.runwayml.com/the-research-origins-of-stable-diffusion>. Accessed 1 Mar. 2024.

Xu, Yaoda, and Marvin M. Chun. “Selecting and Perceiving Multiple Visual Objects.” *Trends in Cognitive Sciences*, vol. 13, no. 4, Apr. 2009, pp. 167–74. *www.cell.com*,

<https://doi.org/10.1016/j.tics.2009.01.008>.