Synthetic Data

Samyak Thapa

University of Virginia

School of Engineering and Applied Science

STS 4500: STS and Engineering Practice

Dr. Richard Jacques

November 21st, 2023

Introduction

In a world governed by algorithms, implicit biases are often the root of many societal issues. Artificial intelligence and machine learning have become ubiquitous in our modern world. From unlocking your phone, to the advent of self-driving vehicles, to companies advertising which product you are most likely to buy, our world is governed by these algorithms. But what happens when these algorithms, meant to offer convenience and automation, actively propagate, and reinforce social biases?

This can, in part, be alleviated by synthetic data. Synthetic data is defined as "information that's been generated on a computer to augment or replace real data" ("What"). Since the data is fake, you can make a virtually unlimited supply of exactly the type of data you need. Researchers propose that using synthetic data can fine-tuned data to de-bias your models and algorithms ("Five"). For example, in medicine, you would be able to account for edge cases and extremes, meaning a more equipped model, and in social media fake data could mitigate privacy concerns, since training examples do not have to involve real people (Toews).

Synthetic data has promise towards creating more equitable algorithms, but what about its drawbacks? This paper aims to not only dive into the promise of this emerging field from a technical standpoint, but also the ethical considerations necessary for any new technology.

Technical Topic

This field of research is new, relevant, and increasing exponentially. Gartner, a technological research, and consulting firm, released a study that "predicts 60% of data for AI will be synthetic [in 2024]… up from 1% in 2021" ("Gartner"). It is evident that this issue will have broad implications for the future of AI.

As mentioned before, machine learning algorithms suffer from bias in training data. This bias can perpetuate existing societal norms. A prime case study on this topic is Amazon's now obsolete AI tool to recruit top talent. In 2018, the story of Amazon's biased hiring tool made headlines when internal research showed that it strongly preferred men's resumes. This failure was a result of inherently biased data, as men still majorly dominate the technology sector. This is dangerous, as it shows that AI can reinforce established imbalances in opportunity ("Carnegie"). As James Zhou, a member of the Stanford Institute for Human-Centered Artificial intelligence puts it, "one of the best ways to improve algorithms' trustworthiness is to improve the data that goes into training and evaluating the algorithm" ("Data"). Synthetic data has the capability to do this. IBM research has developed a tool to "[create] synthetic text to reduce bias in language classification models," specifically in the context of de-biasing sexist sentiment classifiers ("Five"). It works by taking a sentence, making a new fake sentence with the gender of the subject flipped, and then retraining the data to teach the model that both sentences should have equivalent sentiment. The concrete example they provide is "my boss is a man." The model gives this sentence a positive sentiment-rating. Then, the newly generated sentence flips "man" to "woman." The model produces an output that has a negative sentiment-rating, even though it obviously should not. The model is then retrained to correct this behavior, using similar, fake sentences. It is not hard to imagine how this advancement by IBM could have been used to

improve Amazon's sexist recruitment tool. This clearly shows that synthetic data has the potential to reduce unfair stereotyping in the hiring process.

While the mechanism in how synthetic data generation works is too complex for the scope of this paper, it is worthwhile to examine a high-level overview. Synthetic data is primarily created using a machine learning model called a Generative Adversarial Network, or GAN for short. This is the backbone behind DALL-E, OpenAI's text-to-image model, the same company responsible for creating ChatGPT. GANs work by taking two neural networks and "[pitting] them against one another" (Towes). Rob Toews, a Harvard, and Stanford educated venture capitalist, outlines this process in detail in his Forbes article. One neural network, called the "generator" is responsible for generating new images that resemble its training data. Another neural network, called the "discriminator," tries to discriminate between pictures in the training data and pictures output by the generator. These two neural networks go in a loop, one repeatedly creating more indiscernible images, and the other getting better at figuring out which images are not real. Towes goes onto write that "Eventually the discriminator's classification success rate falls to 50%, no better than random guessing, meaning that the synthetically generated photos have become indistinguishable from the originals" (Towes).

With an understanding of one of the processes to create fake data, and looking at one of the many ways synthetic data can be used to improve AI for society, a good foundation has been provided to examine the ethics of this technology.

STS Topic

Perhaps the most recognizable example of synthetic data by the public are deepfakes. The Oxford English Dictionary defines a deepfake as "various media… that has been digitally manipulated… often maliciously." Deepfakes have been rampant, from being used as a tool for actors to look younger, to spurring misinformation using presidential candidates' faces, to being used to sexualize celebrities and models in a skin-crawling, dystopian manner.

Deepfakes are just one example of where the power of synthetic data can go wrong. With an infinite amount of any kind of data you want, the possibilities can be catastrophic. Consider China and their facial recognition system for its citizens. With synthetic data, one could easily increase the accuracy of an identification, simply by spawning more images from different angles, lightings, and vantage points. Furthermore, a facial recognition model could be trained solely to pick our individuals of a certain race. The idea of foreign governments having access to near-perfect photo identification or the ability to target a certain race with precision should be deeply concerning. The idea of any government, including our own, fleshing out this sort of idea, is deeply troubling.

Indeed, these concerns are corroborated by Dr. Mauro Giuffrè and Dr. Dennis Shung and Dr. Dennis Shung, researchers from Yale who work at the intersection of machine learning and healthcare. While their paper in Nature takes on a lens more suited for healthcare than potential foreign technology, their section on pitfalls in this technology is relevant to this thesis. In particular, they state "if a synthetic dataset is trained on a dataset of facial images that majorly includes people from a certain ethnicity, the synthetic images generated will naturally reflect this imbalance, thus perpetuating the initial bias" (Giuffrè). This statement signifies that facial recognition in general, can be fine tuned and biased towards a particular ethnicity, if deemed

worthwhile by an adversary training the model. It is not hard to imagine governments around the world honing technologies like this, away from the view of the rest of the world. At least in the private sector, models are subject to scrutiny, and companies can be held accountable for the decisions they make in the models they release. This is not the case in government research facilities that hide behind security clearances and thick walls.

A potential scenario that is relevant to current events is the use of misinformation and synthetic data in global politics. A hypothetical scenario could involve one side of conflict using synthetic images of bloody civilians subject to war, to propagate whatever narrative they so desire. This would be irrespective of its accuracy. Such a scenario is very feasible given the current global climate and given the technology readily available.

Conclusion

There are no deliverables since this prospectus is independent research on a CS topic, I found interesting.

Like other bleeding edge technologies, synthetic data is a weapon that can lead to more equality, or further division. It is a powerful tool in the current growing space of artificial intelligence, with a wide variety of ethical considerations to keep in mind. When used properly, it can be a force for good and can debias the algorithms that perpetuate existing norms. In the hands of an adversary, synthetic data can lead to the spread of misinformation and shatter our beliefs of what is real and what is not.

Works Cited

Data-Centric AI: AI Models Are Only as Good as Their Data Pipeline. Stanford HAI. Accessed October 27, 2023. https://hai.stanford.edu/news/data-centric-ai-ai-models-are-only-good-their-data-pipeline

Five ways IBM is using synthetic data to improve AI models. IBM Research Blog. Published February 9, 2021. Accessed September 28, 2023. https://research.ibm.com/blog/synthetic-data-explained

Gartner Identifies Top Trends Shaping the Future of Data Science and Machine Learning. Gartner. Accessed October 27, 2023. https://www.gartner.com/en/newsroom/press-releases/2023-08-01-gartner-identifies-top-trends-shaping-future-of-data-science-and-machine-learning

Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *npj Digit Med*. 2023;6(1):1-8. doi:10.1038/s41746-023-00927-3

Jacobsen BN. Machine learning and the politics of synthetic data. *Big Data & Society*. 2023;10(1):205395172211453. doi:10.1177/20539517221145372

Toews R. Synthetic Data Is About To Transform Artificial Intelligence. Forbes. Published June 12, 2022. Accessed September 28, 2023. https://www.forbes.com/sites/robtoews/2022/06/12/synthetic-data-is-about-to-transform-artificial-intelligence/

University CM. Amazon Scraps Secret AI Recruiting Engine that Showed Biases Against Women - Machine Learning - CMU - Carnegie Mellon University. Machine Learning | Carnegie Mellon University. Accessed October 27, 2023.

What is synthetic data? IBM Research Blog. Published February 8, 2023. Accessed October 27, 2023. https://research.ibm.com/blog/what-is-synthetic-data

Zewe A. In machine learning, synthetic data can offer real performance improvements. MIT News | Massachusetts Institute of Technology. Published November 3, 2022. Accessed September 28, 2023. https://news.mit.edu/2022/synthetic-data-ai-improvements-1103