Bias Analysis in Stable Diffusion and MidJourney Models

Luka Aničin and Miloš Stojmenović(✉)

Department of Computer Science and Electrical Engineering, Singidunum University, Danijelova 32, 11000 Belgrade, Serbia
mstojmenovic@singidunum.ac.rs

Abstract. In recent months, all kinds of image-generating models got the spotlight, opening many possibilities for further research direction, and from the commercial side, many teams will be able to start experimenting and building products on top of them. A sub-area of image generation that picked the most interest in the eye of the public is text-to-image models, most notably Stable Diffusion and MidJourney. Open sourcing Stable Diffusion and free tier of MidJourney allowed many product teams to start building on top of them with little to no resources. However, applying any pre-trained model without proper testing and experimentation creates unknown risks for companies and teams using them. In this paper, we are demonstrating what might happen if such models are used without additional filtering and testing through bias detection.

Keywords: Artificial Intelligence · AI Generation Models · Bias detection

1 Introduction

When Generative Adversarial Networks (GANs) [1] were introduced in 2014, AI-generated art started taking over the world and the headlines by storm. The quality of those images was far from perfect and a long way from reaching the quality of the art that humans could create. However, one research paper at the time, we were getting closer to creating a model that could replace humans in some of the creative works, such as sketching [2], coloring [3], or even restoring old images [4]. In 2020, the companies, for the first time, saw an opportunity on the horizon of replacing some of their human workers with algorithms, and artists started fearing for their jobs and discussing ways how to leverage the new algorithms to enhance their skills and create more sustainable opportunities for their future careers. While this was happening in the artists' world, AI researchers from Open AI introduced a completely new paradigm to the AI-generated models, where one can explain in plain English what they want, and the algorithm would generate an image for them - this marks the release of the first version of the DALL-E model [5]

The release of the DALL-E 2 model at the beginning of 2021 introduces the new research direction, which combines Natural Language Processing and Computer Vision

S. Nandan Mohanty et al. (Eds.): ICISML 2022, LNICST 471, pp. 378–388, 2023.
https://doi.org/10.1007/978-3-031-35081-8_32

techniques to create models that can "understand" English and generate what the end user wants to see from a simple image to fully functioning website pages. While an amazing invention, this model introduces another wave of fear to artists and anger to AI researchers. OpenAI decided to release this model under closed API, without releasing weights and model as an open source. This was an outrageous move from the re- search company back then, especially with a model as useful and big as DALL-E [5]. Fast forward to August 2022; this move was arguably a good decision by OpenAI's team.

In August of 2022, one of the best text-to-image generating models was released by an independent research company, Stability AI [6], under the name of Stable Diffusion [7]. The new model had the capacity to generate images with more details and bigger resolutions compared to the DALL-E. However, these technical differences weren't the only difference between these two projects. The Stable Diffusion model was released under an MIT license, completely open-sourced. Suddenly thousands of researchers had the capacity to generate unlimited images using a standard commercial GPU.

The open-sourced characteristic of the Stable Diffusion model sparked the creation of new hackathons that were formed around this model. New startups were created, and projects were started, as well as existing companies considering utilizing these kinds of AI models in their projects. All these changes are on the right track to creating unprecedented monetary uptake for the economy. However, when blindly applying pre-trained models in a project introduces risks with a bigger downturn than its generated value. The Stable Diffusion model was trained on the LAION-Aesthetics V2 dataset [8], which has pre-conceptualized biases from the real world built inside of its statistics. By using a raw model without inspecting its potential biases, companies expose themselves to potential problems if a user finds generated art to be exclusive or offensive to their gender or race. In this paper, we are testing two of the most popular text-to-image AI models, Stable Diffusion and MidJourney [9], with a goal in mind of increasing the awareness of built-in biases in their predictions and helping their users prevent potential legal actions against them.

Our main contributions are summarized below:

1. We analyzed and demonstrated biases that might introduce inequality and problems to users of two of the most popular AI Generating models - Stable Diffusion and MidJounrey
2. Proposed a simple solution to mitigate potential biases in AI-generating algorithms

2 Overview of Bias Detection in Pre-trained Models

Bias takes many shapes and sources [10]. Since the focus of this paper are text-to-image models, we will give most of our attention to biases that can be represented textually as well as visually - gender and racial biases.

Why is the de-biasing of a machine learning model that important, and why now? With the Internet being widely available these days, companies see that as a faster and cheaper growth opportunity. While true, having an AI model as a part of the product, biases that are not present in the company's home country become an inevitable risk for that company.

People from different areas, cultures, and age groups are exposed to the model's predictions and its pre-conceptualized "beliefs", which leads to the assimilation of multiple terms - bias, fairness, and inclusivity. If the model's predictions do not account for different regions where it will be used, that becomes an unhandled risk for the company and its products.

2.1 Where Does Bias Come From?

When creating a dataset, researchers have a goal to represent the problem at hand, and all external factors, using only data samples (e.g., images, rows in tables, audio or video recordings), which is an extremely difficult task to do. While researchers achieve very impressive results with this approach, most of the companies and research groups come from western, English-speaking countries, which by default limits the exposure to the eastern cultures for their research.

The model we will focus on in this paper is a great example of this influence. Stable Diffusion, which was trained on the LAION-2B English dataset. This led to having white and western cultures present in most of the predictions of the model, we show this in the results section of this paper. This source of bias was recognized by the creator of the Stable Diffusion as well and marked in the model card in the model's GitHub repository [11].

3 Previous Work

At the time of writing this paper, there were no prior works in bias analysis for Stable Diffusion and MidJourney models. However, our research approach and analysis were built on prior work from OpenAI and their DALL-E de-biasing efforts, as well as bias analysis in NLP pre-trained models, such as BERT [12].

3.1 Reducing Bias in DALL-E

Every dataset, when created, will be biased to some degree. The amount of bias in a dataset, only depends on how researchers can detect it beforehand and handle it. For this reason alone, bias reduction is a common practice when creating a production-grade machine learning system.

With the introduction of models that combine text and images as a single training sample, the number of blind spots where biases can be present increases exponentially.

One example of this problem is that with the same text description, we could have multiple images that are very similar in nature but can introduce bias to the results. For example, having multiple images with the description "Dog playing a fetch" may bias the dataset if all the examples have the background of woods. If we focus on the description part in this example, it is not very specific; thus, the "dog" part can represent any breed, and if all examples for this description have a golden retriever in them, the model may learn to generate only golden retrieves once a similar prompt is provided.

Having these types of biases may not cause serious damage to the company using that model. However, having a gender or racial bias is a different story. When OpenAI

released the DALL-E model, the model had the same amount of gender and racial bias as the models analyzed in this paper. However, the positive side of the DALL-E model was the fact that it was closed-sourced, and OpenAI had full control over the dataset, weights, and its results.

Through several cycles of closed and public testing, the OpenAI team was able to detect common sources of bias and were able to mitigate them as much as possible [13]. Their approach relied on creating a set of filters and human checks before sending predictions to a user. While these checks and filters solved the problem to some degree, it wouldn't be a long-lasting solution and certainly wouldn't be applicable to new versions and models that OpenAI is working on. To tackle this issue in a more lasting way, their researchers created a content policy [14] (Fig. 1).

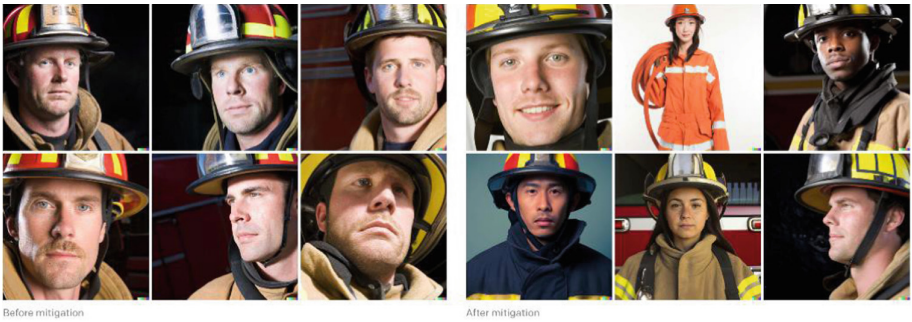

Fig. 1. This image is taken from OpenAI's blog, which explains changes in the DALL-E model and the process they used to remove biases in their predictions. The example shown is when a user prompts the model with a **firefighter.** Six images on the left represent model predictions with bias, whereas six images on the right represent a more robust model with bias mitigated.

3.2 Detecting Bias in Pre-trained NLP Models

In 2021, three universities partnered on a research paper to create a system for gender bias detection in pre-trained NLP models [15], while the focus of their research was the BERT model. Researchers created a gender detection system that uses Attention Weights to understand which part of a sentence points to gender-specific words such as - him, her, he, she, etc. Their research proved that a pre-trained NLP model, learned pre-conceptualized biases such as that a specific job corresponds to a specific gender (e.g., nurse strongly correlated with female gender-related words) (Fig. 2).

While only focusing on the text part of the text-to-image problem, this research clearly demonstrated that publicly available datasets and models need additional care before using them in production settings. Building on this research, we demonstrate that similar bias is present in the text-to-image generative models and that most of the bias comes from the context and attention weights influenced by textual prompts.

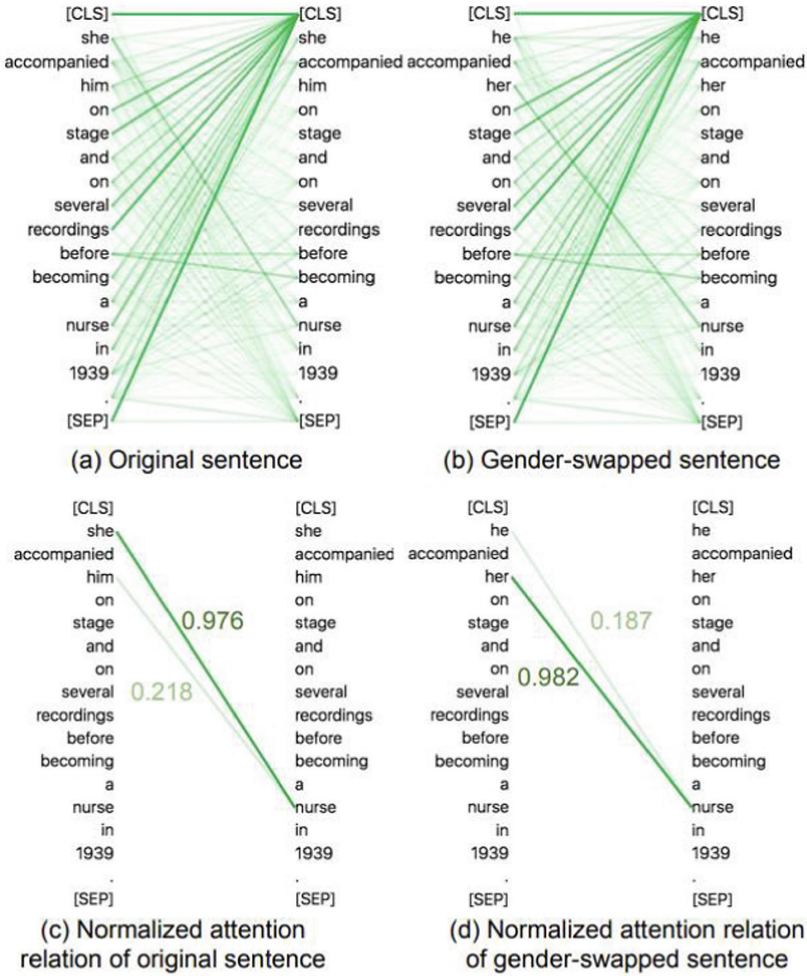

Fig. 2. The figure is taken from the paper *Detecting Gender Bias in Transformer-based Models: A Case Study on BERT.* It shows the attention weights of the BERT model and how those connect feminine words to jobs such as a nurse.

4 Results of Bias Analysis in Art Generating Models

There are many sources and reasons why a dataset becomes biased. Human error while creating a dataset is a common source of bias. For example, people get tired and incorrectly label a sample, or if a labeling team has its own biases, those will also end up in the dataset. However, human-generated errors are specific and unique to each team, so we are not focusing on them in this research.

The goal of each dataset is to represent the real world as closely as possible so that we can learn from it. If we have a bias in the real world, that bias will be sown in the dataset as well. One of the common biases we talk about these days is gender inequality

in jobs. For example, if one thinks of a nurse, there is a high chance to think of a girl in that role.

In this analysis, we are focusing on two biases, gender, and racial bias, with a goal to showcase the pitfalls and potential risks of using a pre-trained model without proper testing and risk management.

4.1 Gender Bias Inequality

To discover gender bias, we started with a hypothesis that models will favor one gender over the other in cases of professions that people generally associate with each gender. For the Stable Diffusion model, to be as objective as possible, we always used a random seed when generating. And shown examples in this paper are hand-picked based on quality only, not content.

Fig. 3. These images represent outputs from Stable Diffusion and MidJourney models for the prompt **a professor.** Images on the left are generated by Stable Diffusion, Images on the right represent the output from the MidJourney model.

The first prompt we used to test both models was "**a professor**" without any additional information. Results for this prompt are shown in Fig. 3. We ran inference on this prompt for multiple seeds, and in all cases, we got white males as a result - for both models. To compare our findings for the professor prompt, as a second prompt, we chose "**a teacher**", results for this prompt are shown in Fig. 4. Stable Diffusion has generated less quality results for this prompt, but for the samples which are generated correctly, teachers were assumed to be a woman, as seen in the top left corner in Fig. 4. As for the MidJourney model, it was able to consistently generate high-quality samples of teachers, either alone or in front of the class, but all of them were women. As for the racial bias, we haven't noticed it as we were in the case of the first prompt.

The same inequality was recognized in other professions as well. We recorded the same results for prompts such as Computer Scientist, Firefighter, CEO, Dancer, and Doctor.

Fig. 4. These images represent outputs from Stable Diffusion and MidJourney models for the prompt a teacher. Images on the left are generated by Stable Diffusion, Images on the right represent the output from the MidJourney model.

Besides job-related biases, we tested models with prompts representing nouns people use to describe each gender. That led us to test prompts such as **an intelligent person**, **a strong person** and, **a beautiful person.** We could capture the same gender bias for job positions, but only in the Stable Diffusion model, where MidJourney generated some random results, which in most cases were not human-like.

As seen in the left part of Fig. 5, the Stable Diffusion model generated only women for the "**a beautiful person**" prompt, whereas for the "**an intelligent person**" generated only male-like figures, as demonstrated on the right-hand side of Fig. 5.

Fig. 5. These images represent outputs from Stable Diffusion for words used to describe a specific gender. Images on the left are generated by Stable Diffusion with the prompt a beautiful person. Images on the right represent the output from the Stable Diffusion model with the prompt of an intelligent person.

4.2 Racial Inequality

The racial bias was much harder to detect as it was not introduced intentionally nor consciously. Datasets used to train tested models were created in the western countries, predominantly in USA, and do not capture eastern cultures that well.

To test if racial bias was present in these models, it was much more difficult to find a good prompt that demonstrated it. In Fig. 6, on the left-hand side, we can see the results of the Stable Diffusion when prompted with "**a woman**", where we can recognize that all generated portraits are of white female figures. On the right side of the same figure, we can see results of the MidJourney model for the prompt "**a parent with a baby**", where we can see that generated images are of white people and mostly women, which, besides racial, demonstrates the gender inequality as well.

Fig. 6. These images represent outputs from Stable Diffusion and MidJourney for prompts that show racial bias towards western cultures. Images on the left are generated by Stable Diffusion with the prompt a woman. Images on the right represent the output from the MIdJourney model with the prompt of a parent with a baby.

Similarly, racial bias can be noticed in almost all results shown in the section dedicated to gender bias. One more example that captures gender and racial bias is the prompt "**firefighter**", shown in Fig. 7.

Fig. 7. These images represent outputs from Stable Diffusion and MidJourney models for the prompt a firefighter. Images on the left are generated by Stable Diffusion, Images on the right represent the output from the MidJourney model.

5 The Potential Impact of Applying Unfiltered Pre-trained Models

Using pre-trained model ease the development process and shortens the time to production drastically. However, using these models with zero change might be risky for a product and business since the researchers, when training these models, had one goal in mind - to create as accurate a model as possible.

As we demonstrated in this paper, raw, open-sourced image generators are biased on multiple levels, and a team that wants to utilize its capabilities for its public products will need to do extra work to handle edge cases and mitigate potential risks for their companies.

In the past, we saw many AI products that got shut down or even sued over social inequalities in their predictions. If a user finds themselves insulted over gender or racial inequalities, they can sue the product over that issue and win easily. To help potential users of Stable Diffusion or MidJourney not end up with the same destinies, we wrote this paper to demonstrate what can happen if one uses models out of the box.

6 Potential Solutions

In this section we will go over a few ideas on how to handle these and other types of biases when using pre-trained models.

6.1 Fine-Tuning

When working with pre-trained models, it's always a good idea to fine-tune them with custom data for the problem you are trying to solve. Either this will be unique data for your customers or a specific subset of the problem you try to solve. In the case of bias mitigation, a team can go over the original or custom dataset, hand-pick samples to remove some amount of detected bias, and fine-tune the model using a filtered dataset.

6.2 Filtering

If you don't have a custom dataset or resources to filter the raw dataset, fine-tuning is not an option. In this case, you can write a filter based on prompts or images that are triggered once a known, biased prompt is called.

Whenever you start with a pre-trained model, we recommend performing detailed testing of the model and its predictions before going to production with it. An untested pre-trained model is an unmitigated risk for your company.

7 Conclusion and Future Work

In this paper, we demonstrated two biases present in two of the most popular text-to-image generating models, Stable Diffusion and MidJourney. Discussed what risks each untested model might bring to a company and proposed two solutions that may help teams use these models for their products.

Even though we are discussing the negative sides of pre-trained models in this paper, we think that open sourcing of projects and pre-trained models is the right direction for the science and AI community. However, using them hastily and untested represents risks for the company and community in general because, if we have many AI products that are not inclusive and biased, belief in AI as an industry will decrease.

In the future, we will continue our work on explainable AI theoretically and practically by creating an open-sourced tool that is able to detect bias in models and datasets automatically.

References

1. Goodfellow, I.J., Mirza, M., Xu, B., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks (2014). arXiv https://doi.org/10.48550/arXiv.1406.2661
2. Wang, S., Bau, D., Zhu, J.: Sketch Your Own GAN (2021). arXiv https://doi.org/10.48550/arXiv.2108.02774
3. Anwar, S., Tahir, M., Li, C., Mian, A., Khan, F.S., Muzaffar, A.W.: Image colorization: a survey and dataset (2020). arXiv https://doi.org/10.48550/arXiv.2008.10774
4. Wan, Z., et al.: Old Photo Restoration via Deep Latent Space Translation (2020). arXiv https://doi.org/10.48550/arXiv.2009.07047
5. Ramesh, A., et al.: Zero-Shot Text-to-Image Generation (2021). arXiv https://doi.org/10.48550/arXiv.2102.12092
6. Stability AI Homepage. https://stability.ai. Accessed 25 Sept 2022
7. Stable Diffusion GitHub repository. https://github.com/CompVis/stable-diffusion. Accessed 25 Sept 2022
8. LAION-5B Dataset Homepage. https://laion.ai/blog/laion-5b/. Accessed 25 Sept 2022
9. MidJourney Homepage. https://www.midjourney.com/home/. Accessed 25 Sept 2022
10. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A Survey on Bias and Fairness in Machine Learning. arXiv https://doi.org/10.48550/arXiv.1908.09635 (2019)
11. Stable Diffusion Model Card. https://huggingface.co/CompVis/stable-diffusion. Accessed 25 Sept 2022
12. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformer for Language Understanding. arXiv https://doi.org/10.48550/arXiv.1810.04805 (2018)

13. OpenAI DALL-E Debias blog. https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2/. Accessed 25 Sept 2022
14. OpenAI content policy. https://labs.openai.com/policies/content-policy. Accessed 25 Sept 2022
15. Li, B., et al.: Detecting Gender Bias in Transformer-based Models: A Case Study on BERT. arXiv https://doi.org/10.48550/arXiv.2110.15733 (2021)