

Top Text

Tom Jacobs and James Boogaard

December 15th 2022

Module: SEAR

Venlo, Limburg, Netherlands

Abstract

Following the explosion of artificial intelligence during the year of 2022, there now are many text-to-image generation applications out there. This is great for the field of artificial intelligence as a whole, however it has made knowing which application is best quite hard for artists who are new to text-to-image generation. This paper looks at four of the most popular generators, with the aim to find out which one is the best for an inexperienced artist.

Contents

1	Introduction	1
1.1	Context and background	1
1.2	Problem and hypothesis	1
2	Methods	3
2.1	Data collection	3
2.1.1	Generating images	3
2.1.2	Scoring generated images	4
2.2	Visualizing data	4
3	Results	5
3.1	Figure 3 and 4	6
3.2	Figure 5 and 6	7
3.3	Figure 7	8
3.4	Figure 8	9
4	Discussion	10
4.1	Interpretation	10
4.2	Secondary research question justification	10
4.3	Answering primary research question	11
4.4	Further research	11

List of Figures

1	Results from Google Trends showing the popularity of different text-to-image applications	2
2	3
3	5
4	5
5	6
6	7
7	8
8	8

1 Introduction

1.1 Context and background

During the year of 2022 there has been a massive increase in the popularity and use of text-to-image generators, an artificial intelligence technique that translates human text into a computer-generated image. You will likely have heard of at least one of these if you frequent the internet. This explosion in popularity has led to many different text-to-image applications being made, which on the one hand has given many more options to artists(people who use a text-to-image generator to create and refine an image), but it has also made the landscape more difficult to navigate because of its abundance of choice. This paper makes an effort to try and alleviate this problem by picking four of the most used text-to-image generators and comparing them based on their accuracy and process. This allows us to help people gain some perspective on which text-to-image generator is best for their needs.

A research paper was published on the nineteenth of February 2020 called: "A survey and taxonomy of adversarial neural networks for text-to-image synthesis" that makes use of surveys to compare models for text-to-image generations. This paper aimed to show which text-to-image technique yielded the most realistic results in a wide variety of categories. One issue with this paper was that it was written before the explosion of consumer text-to-image generators happened, which is why we will instead focus on modern consumer-grade generators which are available today.

1.2 Problem and hypothesis

Our goal during the course of this research paper is to find out what the best text-to-image generator is out of the four we chose, in the case of an artist who is not experienced in the world of text-to-image generation. Our hope is that artists who are new to text-to-image generation will be able to make an educated decision on which generator is best for them on the basis of this paper.

Another question we take a look at is what makes an application good for an inexperienced artist. In answering this question we will consider: the tools available to an artist, the difficulty in using these tools, the accuracy in generating images, the learning curve of an application and the speed with which the application can generate images. With these points of consideration it should be possible to score each application on their ease of use, and on their future potential if the artist were to stick with that application.

The text-to-image generators we will be testing during the course of this paper are: Stable diffusion, Midjourney, DallE and Dream by Wombo. We came to these models after looking at several applications in google trends. This gave us a good idea of what the landscape looked like, and which applications artists would be likely to choose. You may notice below in Figure 1 that Dream by Wombo is not that popular anymore, however it was one of the first openly accessible text-to-image applications out there.

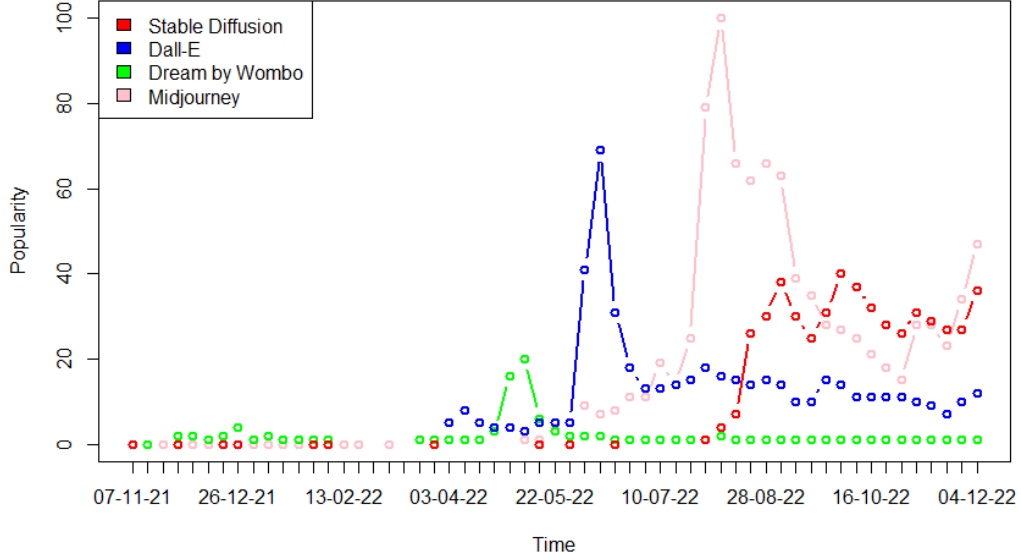


Figure 1: Results from Google Trends showing the popularity of different text-to-image applications

We expect the best text-to-image generator out of the four we chose to be stable diffusion. This is due to the fact that stable diffusion offers its users more tools and techniques like inpainting (allows you to regenerate a specific area of an image without regenerating the whole image), this allows the user more freedom to fine-tune their result. Given enough time this will yield a more accurate result.

2 Methods

2.1 Data collection

Our goal during the course of this research paper is to find out what the best text to image generator is, out of the four we chose, for artists who are inexperienced in AI art generation. In order to find out which is the best one, we will first look at the accuracy with which the generator can create images. We will score the accuracy with an API developed by DeepAI which can calculate image similarity.

2.1.1 Generating images

We have chosen an image that we find sufficiently covers all the challenging parts of recreating an image using ai art generation such as reflections and complex composition(see figure 2 below). Subject 1 has experience with Midjourney and Subject 2 has experience with Stable diffusion and neither of them have any experience in the other applications. Both Subject 1 and Subject 2 then use each of the four text-to-image generation applications to try and recreate that image within a given time frame (15 minutes per application). After each iteration we take the generated image and add it to a list of the iterations for that application.



Figure 2

2.1.2 Scoring generated images

With those sets of data, we then plug the images (each iteration) into the image similarity checking API, developed by DeepAI, that can give you a numerical value for how close one image is to another. This way we can see how close each iteration is to the image we are trying to recreate. The amount of iterations each application is able to create then gives us insight into how fast the application is and how many iterations you have to do to achieve a certain result. We also gain an idea of how easy to use each application is through this process.

In order to compare the image generation applications effectively we observe a multitude of factors that will indicate the strengths and weaknesses of that application. These factors are: The amount of iterations each participant can generate while trying to create the image within the 15 minute time limit, how close each iteration is to the image we are trying to replicate using the ai image comparing API's unit of measurement (the unit is called distance. The closer to zero the distance is, the closer the images are together. Zero means that the two images that are being compared are exactly the same), how easily we were able to access the application and its functionality, how much functionality each application offers and their effects on the final results of our experiment.

2.2 Visualizing data

Once the data is gathered, we can create graphs to compare the applications more easily.

In these graphs the x value is the array of different applications that we use and the y value is the result of that (the distance) for each iteration on the road to trying to recreate an image using these applications. The intervening variable in this case would be the amount of experience using these applications.

We also want to check each applications average result across both Subject 1 and Subject 2 in order to compare how close each iteration gets to the original image on average. This will show us if there are any large differences between the applications as a whole. If there is this will play a role in choosing which is the best application.

The last set of important data we want to visually represent is the min and max of each application and how far spread the results of each application are in order to see how reliable each application is.

3 Results

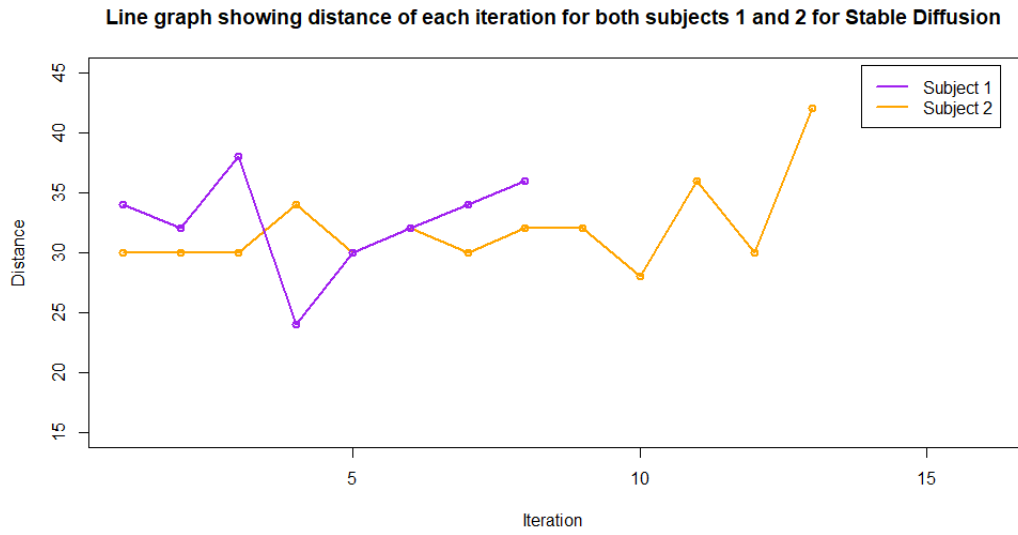


Figure 3

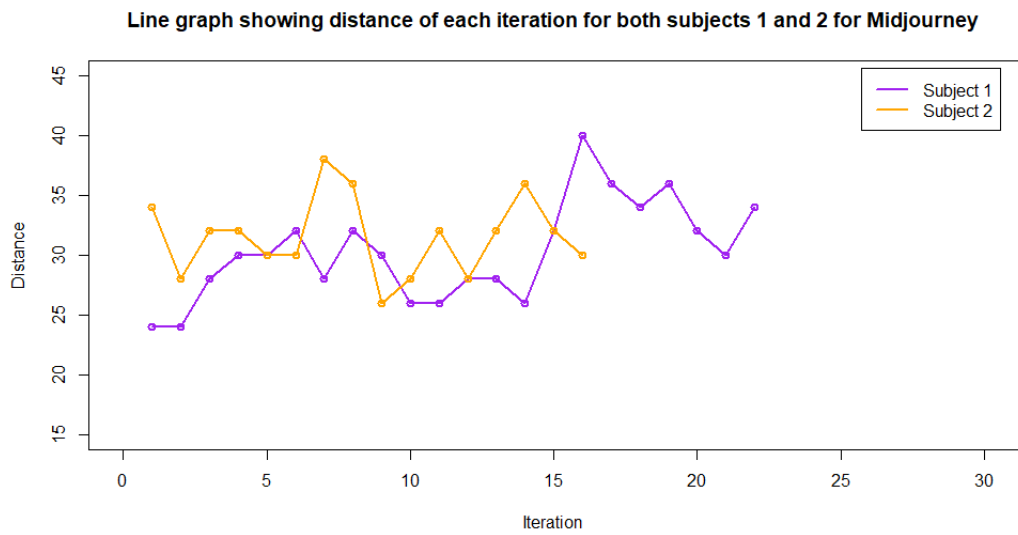


Figure 4

3.1 Figure 3 and 4

What is interesting in both figure one and two is that the discrepancy between the two subjects in both applications is larger than that of any of the other applications we tested. This is seen with Subject two having nearly double the amount of iterations when compared to Subject one in figure 3 and then Subject 1 having more iterations than Subject two in figure 4. Another point of interest is that the application in figure one has as a small amount of iterations when compared to the other applications.

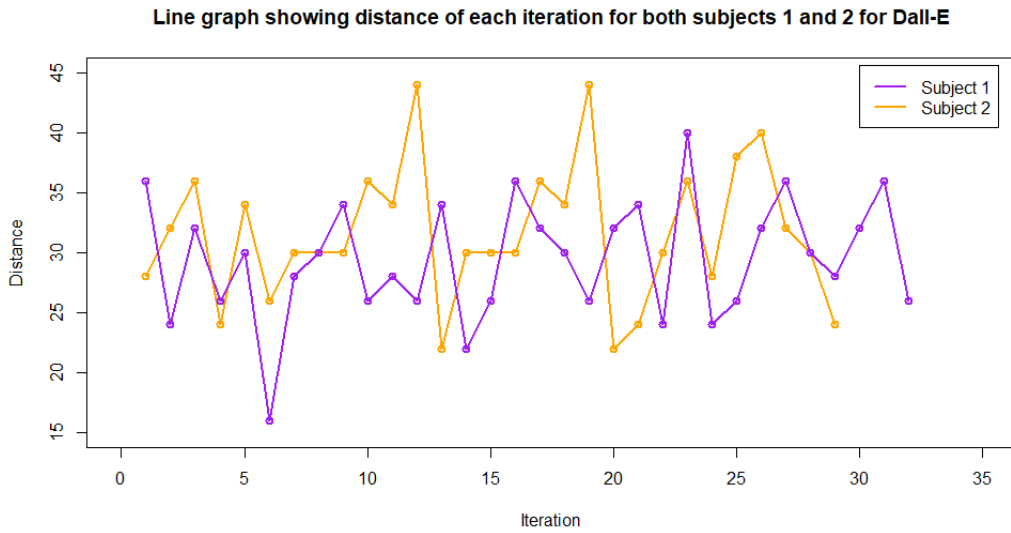


Figure 5

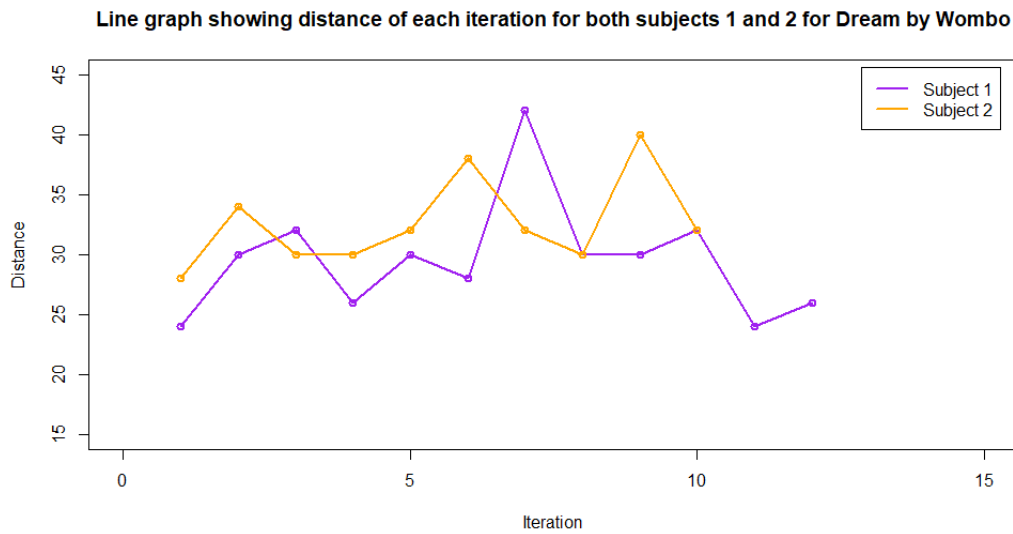


Figure 6

3.2 Figure 5 and 6

In figure 5 we see that Dall-E has the most iterations out of any of the other applications. When comparing the amount of iterations that each subject has in both figure 3 for Stable diffusion and figure 4 for Midjourney, the two applications are very far apart. In contrast the amount of iterations between each subject is very close together in figure 5 for Dall-E and figure 6 for Dream by Wombo.

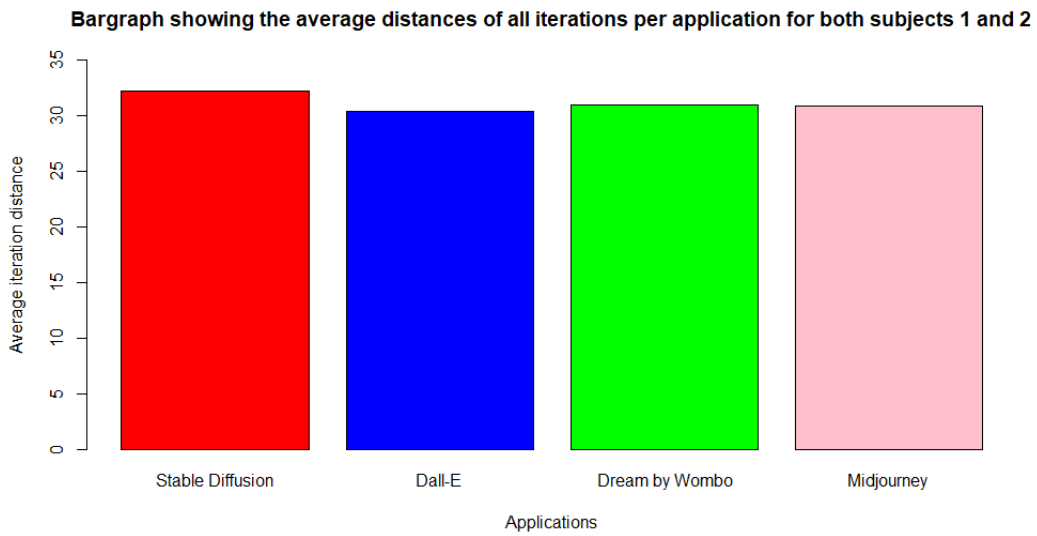


Figure 7

3.3 Figure 7

Figure 5 shows that the average distances of all the applications from both subjects one and two are very close. When comparing the applications Dall-E has the least average distance and stable diffusion has the most average distance but again only by a slight margin.

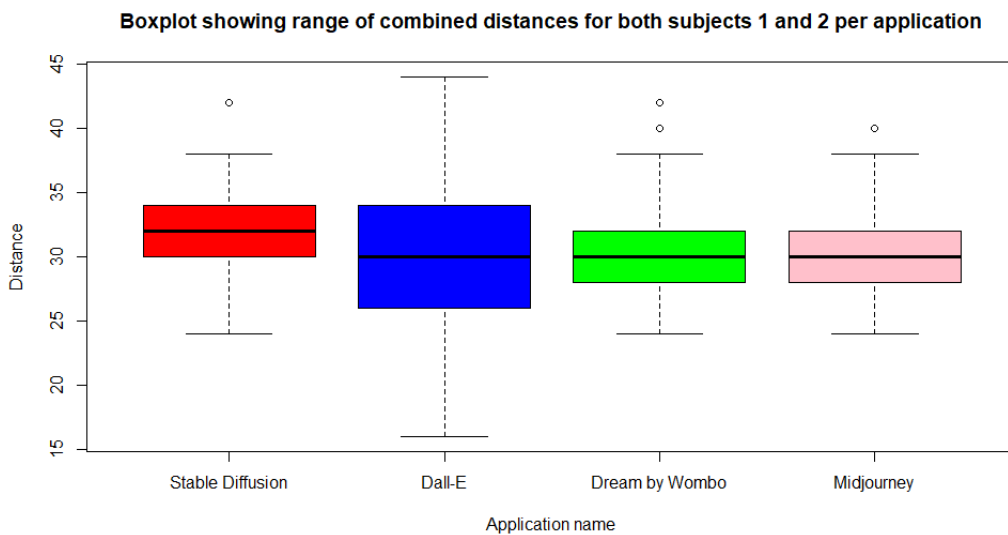


Figure 8

3.4 Figure 8

In Figure 8 there are two important points of interest. Firstly is that Dall-E has a massive range between the minimum and maximum distances when compared to the other applications as well as having the largest space between quartile 2 and quartile 3. This means the spread of the distances is also the largest out of any of the other applications. Second is that Stable Diffusion has the largest median distance.

4 Discussion

4.1 Interpretation

In order to find the best application out of the four we chose we needed to see what actually makes a text-to-image generation application good. This was done by analysing figures 3 through to 7. The applications shown in figure 3 and 4 are applications that both subjects have experience in (Subject 1 has experience with Midjourney and Subject 2 has experience with Stable diffusion). What makes figure 3 and 4 important is that both the graphs show the difference that having experience in an application makes, and that difference is only in the amount of iterations the user is able to generate because the quality of the images is very random which can be seen with how the line graphs from figures 3 to 6 all fluctuate hugely and never set any kind of discernible pattern.

In figure 5 and 6 we then see two applications that neither Subject 1 nor Subject 2 have any prior experience in. using comparative analysis we can then compare both the graphs that the users have experience in (being figure 3 representing Stable diffusion and 4 representing Midjourney) and the graphs that the subjects don't have experience in (being figure 5 representing Dall-E and 6 representing Dream by Wombo) and we can see more clearly that when both subjects have the same amount of experience with an applications the amount of iterations the subjects are able to generate is noticeably closer together compared to when there is a discrepancy in experience level.

When we take a look at figure 7 we can see that even though it is the combinations of the averages of all iterations across both Subject 1 and 2 per application, the differences between the bars is very small. This can be attributed to the nature of text-to-image generation being that of a random series of attempts until the user reaches a point where they are happy with the result. This is further proven when you look at any of the patterns for figures 3 through 6 and how the distance for each application's iterations follow no solid pattern but rather jumps around sometimes even ending up further away from the original image than some of the earlier iterations. An example of this can be seen in figure 4 with Subject 1.

4.2 Answering secondary research question

In order to answer the secondary research question we looked at how different aspects of the applications affected the resulting images that were generated and came to the conclusion that the amount of functionality does not play a role in the quality of generated images with inexperienced users which makes sense as they may struggle to use said functionality. This can be seen when you take into consideration that stable diffusion has the most functionality but it has the worst average iteration quality across both users. We also looked at how easy the application was to gain access to and came to the conclusion that every one of them was very easy to use and gain access to, so that also became an irrelevant comparison factor.

With all this information we can now say that an application's value to an inexperienced user is not about how good the AI generation technique is nor is it about how much functionality it supplies it's user. This is because AI text-to-image generation is very random and relies

a lot on luck with our current technology even for users that are experienced with these applications. What we do know is that currently if all the applications rely on luck to some degree it is then better if the application is able to give many iterations because the more attempts you are able to make, the faster you are more likely to get the result you are looking for.

4.3 Answering primary research question

Our primary research question can then be answered now that we know what makes an AI text-to-image application good for an inexperienced user and that is how fast the application is able to generate iterations. With this in mind we can then say that the best application out of the four we chose is Dall-E as it is able to generate the most iterations. On a side note our experience with it also lead to us believing that it offers a very streamline user experience, is easy to access and intuitive to use.

4.4 Further research

The ways in which our research can be carried out in an improved manner are to firstly use a bigger sample size in order make the results of the research more convincing. This can be done by using more text-to-image applications and trying to recreate more than just one picture.

Secondly you can use many different ways to compare images rather than just one in order to make the results more accurate. This can be done with surveys where you ask people which image is closer to a humans perspective or by using not just one application for image comparison.