Proposed Solutions for DALL-E 2

Mohammad Arkam

Dept. Of Computer Science & Engineering
School Of Engineering & Technology, Sharda
Greater Noida,India

Aditya Thakur

Dept. Of Computer Science & Engineering
School Of Engineering & Technology, Sharda
Greater Noida,India

Sandeep Kumar

Dept. Of Computer Science & Engineering
School Of Engineering & Technology, Sharda
Greater Noida,India

Abstract: The successor to DALL-E from 2021, DALL-E 2, was unveiled by OpenAI, a research facility for artificial intelligence, In April. Both AI systems are capable of producing pictures that resemble photos, graphics, paintings, animations, and pretty much any other art form you can think of from text descriptions in natural language. Better resolution, quicker processing, and an editing function in DALL-E 2 upped the ante. These features allow users to alter created images using only text commands, such as "replace that vase with a plant" or "enlarge the dog's nose." Additionally, users can contribute their own images and instruct the AI algorithm how to riff on them.

The DALL-E 2 system creates unique, artificial images that correspond to input text used as a caption. DALL-E 2 initially sparked awe and excitement throughout the world. In a matter of seconds, any assortment of items and creatures could be assembled, any artistic style could be imitated, any location could be portrayed, and any lighting conditions could be portrayed. As participants listed the industries that could very easily be affected by such a technology, there were also waves of worry. The findings that have surfaced in recent months speak volumes about the limitations of current deep-learning technology, providing us with a glimpse into what AI comprehends about the human world—and what it completely lacks.

Introduction:
Large diffusion-based text-to-image models produce visually appealing images that depend on input texts, such DALLE-2. However, how well these models actually capture It has been questioned whether human language has compositional organization or not.

Each word has a distinct role in the interpretation, and there is a one-to-one mapping between symbols and roles, which is a fundamental characteristic of the interpretation of natural language utterances. Although symbols and phrase patterns may be unclear, this ambiguity is already addressed once an interpretation has been created. For instance, while the symbol bat in the phrase "a flying bat" might be read as either a wooden stick or an animal, our two potential readings of the phrase are either a flying animal or a flying wooden staff, but never both at once. The term "bat" cannot be used to refer to an animal in the same understanding after it has been used to refer to an object (for instance, a wooden stick). The term gold is used as a modifier of ingot in both a fish and a gold ingot.

After being used, it cannot be used again to modify another item of the same interpretation or be used independently. This characteristic is known as resource sensitivity by certain linguists.

Working: DALL-E 2 is an example of a "generative model," a subset of machine learning that generates complicated output rather than predicting or classifying data from input. DALL-E 2 generative model is different from other generative model by its capability to maintain semantic consistency in the images it creates.

The most basic level of DALL-E 2's operations is as follows:

A text encoder that has been taught to map a text prompt to a representation space is first given a text prompt. The semantic information of the prompt contained in the text encoding is then captured by a model known as the prior, which maps the text encoding to a corresponding image encoding.
Finally, a stochastic image decoder creates an image that represents this semantic data visually.

CLIP:
To determine how much a particular text passage corresponds to an image, CLIP is trained on hundreds of millions of photos and the captions that go with them. In other words, CLIP just learns how closely connected any given caption is to a particular image rather than attempting to predict a caption given an image. CLIP can understand the relationship between textual and visual representations of the same abstract object thanks to its contrastive rather than predictive objective. Let's look at how CLIP is trained to comprehend its inner workings since CLIP's capacity to learn semantics from natural language is the foundation of the entire DALL-E 2 model.

CLIP Instruction:
The basic tenets of CLIP training are relatively straightforward:

First, all captions and images are run through the appropriate encoders to map all objects into an m-dimensional space.
The cosine similarity of each pair of (picture, text) is then calculated.
The goal of the training is to simultaneously maximize cosine similarity between N pairs of correctly encoded images and captions and reduce cosine similarity between N pairs of incorrectly encoded images and captions. Because CLIP is ultimately what defines how semantically connected a natural language fragment is to a visual concept, which is crucial for text-conditional image production, CLIP is significant to DALL-E 2.

Following training, the CLIP model is put to rest, and DALL-E 2 begins to learn how to reverse the image encoding mapping that CLIP had just discovered. Although our focus is on picture creation, CLIP learns a representation space where it is simple to ascertain the relationship between textual and visual encodings. Therefore, in order to complete this assignment, we must understand how to take advantage of the representation space.

A Diffusion Model: What Is It?
The popularity of Diffusion Models, a thermodynamics-inspired creation, has dramatically increased in recent years. Diffusion models reverse a gradual learning process to acquire the ability to generate data. The noise-generating process is shown in the figure below as a parameterized Markov chain that gradually contaminates an image with noise before producing pure Gaussian noise asymptotically. In order to navigate backwards along this chain and reverse this process, the Diffusion Model gradually removes noise over a number of timesteps.

GLIDE Instruction:
Although GLIDE wasn't the first Diffusion Model, it made a significant improvement by allowing for the production of images with text-conditions. You'll see in particular that Diffusion Models begin with randomly selected Gaussian noise. At first, it wasn't clear how to modify this procedure to produce particular images. A Diffusion Model can consistently create photorealistic images of human faces after being trained on a dataset of human faces, but what if someone wants to create a face with a specific feature, like brown eyes or blonde hair?

A modified GLIDE model is used by DALL-E 2 to incorporate projected CLIP text embeddings in two different methods. The first method entails adding the CLIP text embeddings to the timestep embedding already present in GLIDE. The second method entails producing four more context tokens and concatenating them with the output sequence of the GLIDE text encoder.

Glide's significance for DALL-E 2 The text-conditional photorealistic picture generating capabilities of GLIDE might be simply ported to DALL-E 2 by conditioning on image encodings in the representation space as opposed to text. DALL-E 2's improved GLIDE gains the capacity to generate semantically coherent pictures in light of CLIP image encodings. It's important to keep in mind that the reverse-Diffusion process is stochastic, therefore changes can be produced by repeatedly passing the identical image encoding vectors through the modified GLIDE model.

How do we actually identify these encoded representations, even while the modified-GLIDE model correctly creates images that mirror the semantics acquired by image encodings? How exactly do we introduce the text conditioning data from our prompt into the image generating process, in other words?

The authors of DALL-E 2 explore with both autoregressive and diffusion models for the prior but finally discover that they produce performance that is comparable. Because the Diffusion Model uses a lot less processing power,
We currently have all of DALL-E 2's working parts and only need to connect them in a chain to create text-conditional images:

The image description is first mapped into the representation space by the CLIP text encoder.
Then, the diffusion prior maps from a corresponding CLIP image encoding to a CLIP text encoding.
Finally, the modified-GLIDE generation model performs a reverse-Diffusion mapping from the representation space into the image space, producing one possible image that communicates the semantic information contained in the input caption.

According to the document that OpenAI uploaded to ArXiv, DALL-E 2 was trained on around 650 million image-text pairs that were downloaded from the Internet. It discovered the connections between photos and the words used to describe them from that

enormous data set. Before training, OpenAI removed photographs with obviously violent, sexual, or hostile content from the data set.

Methods:

Stimuli: We create linguistic cues (stimuli) that cause behaviour that defies the principle that a symbol should only be used once. We employ homonyms, which are words with two different meanings, for the first scenario (words perceived as two entities).

DALLE-2 frequently produces two objects, one for each of the senses. The next set of prompts is what we create for the modification situations. We include prompts with two entities e1 and e2 and a modifier word m that is semantically compatible with both of them but is only syntactically altering the second one.

Control Stimuli: We must determine whether DALLE-2 tends to display e2 with that property, regardless of e1, in order to support the existence of a leakage of property P between two entities e1 and e2. It is difficult to argue that the presence of the term basket in "a basketball near groceries" caused DALLE-2 to construct a basket, for example, if the default grocery for DALLE-2 in many contexts is a foods basket

Examples:

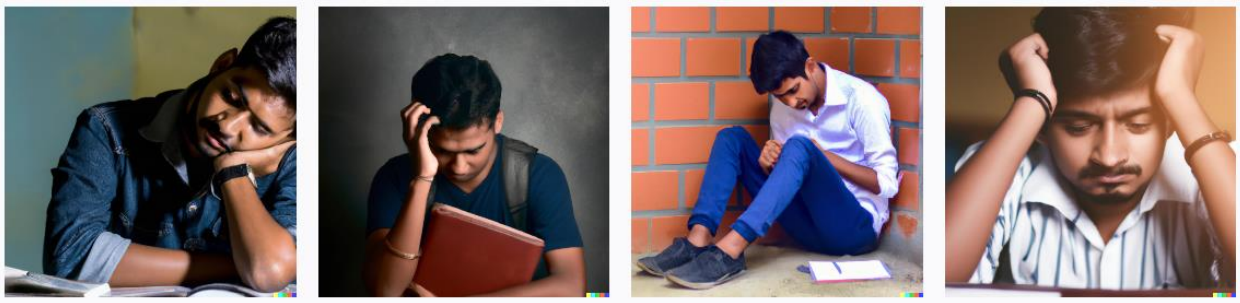
1) **Prompt:** In late afternoon in January in New England, a man stands in the shadow of a maple tree.

Discussion:
As we can see Dall E has perfectly captured the idea behind the prompt and it is intelligent enough to understand that the maple tree does not have any leaves in January.

2) **Prompt: Ancient god disappointed by today's technology usage:**

Discussion:
It has captured the idea that the entity we are talking about is old and by contrast it is dealing with something that is recent, we can say that this AI is intelligent enough to understand sharp contrast in the prompt.

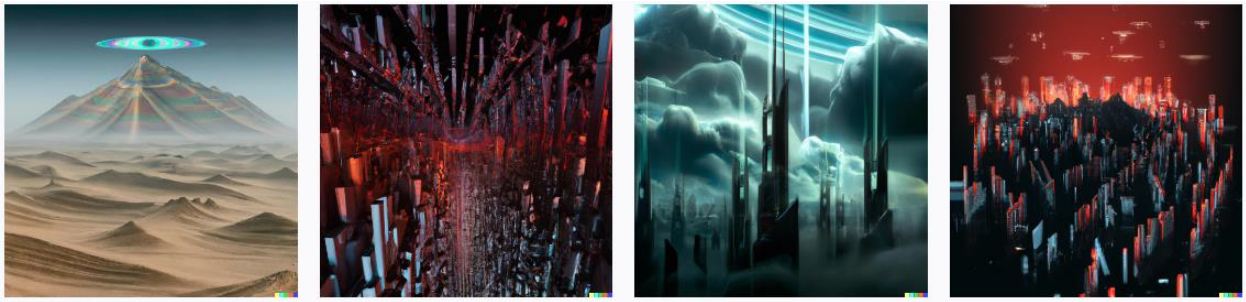
3) **Prompt: Indian student depressed because of his studies and wants to give up on life**



Discussion:
Absolutely stunning result from the AI, Very photorealistic even though it wasn't even mentioned in the prompt, Clearly depicts the prompt, Only drawback of the result is that it struggles with the faces. This is a major drawback of DALL-E as of right now

We can say that when it comes to faces, it is surely one of the weak spot of Dall-E

4) **Prompt: Beeple's 69 million dollar jpeg file**



Discussion:
Not bad result but it has nothing to do with the prompt other than being abstract as Beeple's Art. I was expecting a collection of pictures in one picture, as the JPEG file that was sold

5) **Prompt: The FBI and Supreme Court Justices engage in a baseball game. The justices are on the field, and the FBI is at bat.**

6)

Discussion:
The majority of the photographs include humans, many of whom are situated close to recognizable government structures and some of them are carrying bats. But none even comes close to illustrating the stated circumstance. All of the test captions that have been included are nonsense; DALL-E 2 fails to address this issue (Ramesh, Dhariwal, Nichol, Chu, & Chen, 2022).

7) **Prompt: Pint cartons of milk are on the top shelf of a grocery store refrigerator, quart cartons are in the middle, and gallon plastic jugs are on the bottom shelf.**

Discussion:
Only the first image is close to what has been given to the prompt but other 3 are filled with AI artifacts.

8) **Prompt: An oil painting of a couple in formal evening wear going home get caught in a heavy downpour with no umbrellas, Beautiful realistic city in the background.**

Discussion:

DALL-E 2 made the decision to produce these in a photographic manner, possibly with real-world photographs. The publication of photographic images taken by DALL-E 2 that feature potentially recognizable individuals is forbidden by OpenAI's policy. In each of the ten photographs, a man and lady dressed for the evening are shown in the rain. The couple is covered by an umbrella in two of the pictures but not in the other eight. The couple appears to be inside in one picture, which defies the laws of physics. In one, it appears to be raining even though the streets are wet. Six of the ten photos do, however, categorically adhere to the criterion. However the second image's background looks like boxes of cardboard rather than buildings

9) **Prompt: The Demogorgon from Stranger Things holding a basketball**

Discussion:

DALL-E does an excellent job on this one as the Demogorgon is perfectly generated even though it has less time on screen in Netflix's Stranger things

10) Prompt: Hardest push up variations

Discussion:
A pretty hard and vague prompt to deconstruct for the AI because the only word it understood was pushup, it tries to justify the word 'hard' by inserting an object beneath the women in the right picture on the top and it even tries to make the push ups one handed.

11) Prompt: A horse looking at me through the ceiling

Discussion:
Accurate generation of the prompt, its hard to believe that the photo is generated by an AI until and unless the viewer looks closely to see that the details on the horse's face are weird and skewed.

12) Prompt: People dancing on the subway

Discussion:
The AI, Understood the base idea correctly but as mentioned above, Faces and body shapes are distorted and skewed, other than that a solid output.

13) Prompt: Amazon Forest growing in Amazon Warehouse

Discussion:
Excellent understanding of the "Amazon Warehouse", it nicely generated the warehouse in the background and in the top right corner picture it even grew some trees as a 'forest' was mentioned.

14) Prompt: Launching Eiffel Tower to space

Discussion:
The explosion beneath the tower makes the prompt complete, also showing the intelligence of the AI

References:

1. Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for
2. Language models. Transactions of the Association for Computational Linguistics, 34-48.
3. Marcus, G. F., & Davis, E. (2019). Rebooting AI: Building Artificial Intelligence We Can Trust. Pantheon
4. Press.
5. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sutskever, I. (2021). Learning
6. Transferable visual models from natural language supervision. International Conference on
7. Machine Learning, (pp. 8748-8763).
8. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical Text-Conditional Image

9. Generation with CLIP Latents. arXiv preprint arXiv:2204.06125. Retrieved from
10. https://arxiv.org/abs/2204.06125
11. Rips, L. (1989). Similarity, typicality, and categorization. In S. Vosniadou, & A. Ortony, Similarity and
12. Analogical Reasoning (pp. 21-59). Cambridge: Cambridge University Press.
13. Swimmer963. (2022, May 1). What DALL-E 2 can and cannot do. Retrieved from LessWrong:
14. https://www.lesswrong.com/posts/uKp6tBFStnsvrot5t/what-dall-e-2-can-and-cannot-do
15. Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., & Ross, C. (2022). Winoground: Probing
16. Vision and Language Models for Visio-Linguistic Compositionality. arXiv preprint
17. ArXiv: 2204.03162. Retrieved from https://arxiv.org/abs/2204.03162
18. Royi Rassin, Shauli Ravfogel, Yoav Goldberg DALLE-2 is Seeing Double: Flaws in Word-to-Concept Mapping in Text2Image Models

19. Arjovsky, Martin, Chintala, Soumith, and Bottou, Leon. ´ Wasserstein generative adversarial networks. In Proceedings of the 34th International Conference on Machine Learning, 2017.
20. Cai, Lei, Gao, Hongyang, and Ji, Shuiwang. Multi-stage variational auto-encoders for coarse-to-fine image generation. CoRR, abs/1705.07202, 2017. Das, Abhishek, Kottur, Satwik, Gupta, Khushi, Singh, Avi, Yadav, Deshraj, Moura, Jose MF, Parikh, Devi, and Ba- ´ tra, Dhruv. Visual Dialog. In Proceedings of the IEEE
21. Conference on Computer Vision and Pattern Recognition, 2017. Dash, Ayushman, Gamboa, John Cristian Borges, Ahmed, Sheraz, Liwicki, Marcus, and Afzal, Muhammad Zeshan. TAC-GAN - text conditioned auxiliary classifier generative adversarial network. CoRR, abs/1703.06412, 2017.
22. Deng, J., Dong, W., Socher, R., Li, L. J., Li, Kai, and Fei-Fei, Li. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.
23. Denton, Emily L, Chintala, Soumith, szlam, arthur, and Fergus, Rob. Deep generative image models using a laplacian pyramid of adversarial networks. In Advances in Neural Information Processing Systems 28. 2015
24. Graves, Alex and Schmidhuber, Jurgen. Framewise ¨ phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks, 18(5-6), 2005
25. Gregor, Karol, Danihelka, Ivo, Graves, Alex, Rezende, Danilo, and Wierstra, Daan. Draw: A recurrent neural network for image generation. In Proceedings of the 32nd International Conference on Machine Learning, 2015.
26. Gulrajani, Ishaan, Ahmed, Faruk, Arjovsky, Martin, Dumoulin, Vincent, and Courville, Aaron C. Improved training of wasserstein gans. In Advances in Neural Information Processing Systems 30. 2017.
27. He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
28. Hochreiter, Sepp and Schmidhuber, Jurgen. Long short-term ¨ memory. Neural computation, 9(8), 1997.
29. Hong, Seunghoon, Yang, Dingdong, Choi, Jongwook, and Lee, Honglak. Inferring semantic layout for hierarchical text-to-image synthesis. CoRR, abs/1801.05091, 2018.
30. Huiskes, Mark J. and Lew, Michael S. The mir flickr retrieval evaluation. In MIR '08: Proceedings of the 2008 ACM International Conf
31. Karras, Tero, Aila, Timo, Laine, Samuli, and Lehtinen, Jaakko. Progressive growing of GANs for improved quality, stability, and variation. In Proceedings of the International Conference on Learning Representations, 2018
32. Kim, Jin-Hwa, Parikh, Devi, Batra, Dhruv, Zhang, ByoungTak, and Tian, Yuandong. Codraw: Visual dialog for collaborative drawing. CoRR, abs/1712.05558, 2017.
33. Kingma, Diederik P and Ba, Jimmy. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations, 2015.
34. Kingma, Diederik P. and Welling, Max. Auto-encoding variational bayes. In Proceedings of the International Conference on Learning Representations, 2014.
35. Kingma, Diederik P, Salimans, Tim, Jozefowicz, Rafal, Chen, Xi, Sutskever, Ilya, and Welling, Max. Improved variational inference with inverse autoregressive flow. In Advances in Neural Information Processing Systems 29. 2016.