Discovering the Secret Language of DALLE-2

Giannis Daras and Alexandros G. Dimakis
University of Texas at Austin.
giannisdaras@utexas.edu, dimakis@austin.utexas.edu

Abstract

We discover that DALLE-2 has a secret language that can be used to generate images with absurd prompts. For example, it seems that `Apoploe vesrreaitais` means birds and `Contarra ccetnxniams luryca tanniounons` means bugs or pests. We find that these prompts work in isolation but also sometimes in combinations. We present our black-box method to discover words of this language. This creates important security and interpretability challenges.

Figure 1: Images generated with the prompt: "`Apoploe vesrreaitais eating Contarra ccetnxniams luryca tanniounons`". We discover that DALLE-2 has its own language where `Apoploe vesrreaitais` means birds and `Contarra ccetnxniams luryca tanniounons` means bugs. Hence, this prompt means "Birds eating bugs".

1 Introduction

DALLE [1] and DALLE-2 [2] are deep generative models that take as input a text caption and generate images of stunning quality that match the given text. DALLE-2 uses Classifier-Free Diffusion Guidance [3] to generate high quality images. The conditioning is the CLIP [4] text embeddings for the input text.

A known limitation of DALLE-2 is that it struggles with text. For example, text prompts such as: "`An image of the word airplane`" often lead to generated images that depict gibberish text. We discover that this produced text is not random, but rather reveals a hidden language that the model seems to have developed internally. For example, when fed with this gibberish text, the model frequently produces airplanes.

Some words from this hidden language can be learned and used to create absurd prompts that generate natural images. For example, it seems that `Apoploe vesrreaitais` means birds and `Contarra ccetnxniams luryca tanniounons` means bugs or pests. We found that we can generate images of cartoon birds with prompts like `An image of a cartoon apoploe vesrreaitais` or even compose these terms to create birds eating bugs as shown in Figure 1.

2 Discovering the DALLE-2 Language

We provide a simple method to discover words of the DALLE-2 language. We use (in fact, we only have) query access to the model, through the API. We describe the method with an example. Assume that we want to find the meaning of the word: `vegetables`. Then, we can prompt DALLE-2 with one of the following sentences (or a variation of those):

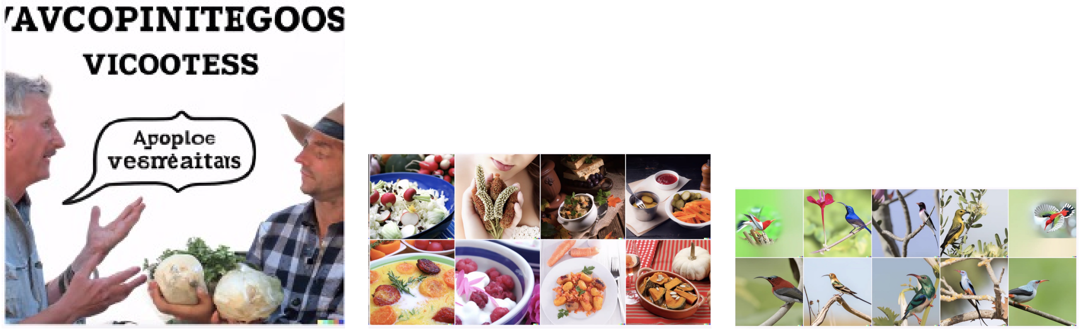
- "`A book that has the word vegetables written on it.`"
- "`Two people talking about vegetables, with subtitles.`"
- "`The word vegetables written in 10 languages.`"

For each of the above prompts, DALLE-2 usually creates images that have some text written text on it. The written text often seems gibberish to humans, as mentioned in the original DALLE-2 paper [2] and also in the preliminary evaluation of the system by Marcus et al. [5]. However, we make the surprising observation that this text is not as random as it initially appears. In many cases, it is strongly correlated to the word we are looking to translate. For example, if we prompt DALLE-2 with the text: "`Two farmers talking about vegetables, with subtitles.`", we get the image shown in Figure 2(a). We prompt the model with words from this language, as shown in Figure 2(b), (c). It seems that `Vicootes` means vegetables and `Apoploe vesrreaitais` means birds. It appears that the farmers are talking about birds that interfere with their vegetables.

3 A preliminary study of the DALLE-2 language

We do a very preliminary study of the properties of this DALLE-2 language.

Compositionality. From the previous example, we learned that `Apoploe vesrreaitais` seems to mean birds. By repeating the experiment with the prompt about farmers, we also learn that: `Contarra ccetnxniams luryca tanniounons` may mean pests or bugs. An interesting question is whether we can compose these two concepts in a sentence, as we could do in an ordinary language. In Figure 1, we illustrate that this is possible, at least sometimes. The sentence: "`Apoploe vesrreaitais eating Contarra ccetnxniams luryca tanniounons`" gives images in which birds are eating bugs. We found that this happens for most, but not all of the generated images.



(a) Image generated with the prompt: “`Two farmers talking about vegetables, with subtitles.`” (b) Image generated with the prompt: “`Vicootes.`” (c) Image generated with the prompt: “`Apoploe vesrreaitais.`”

Figure 2: Illustration of our method for discovering words of the DALLE-2 language. We first query the model with the prompt: “`Two farmers talking about vegetables, with subtitles.`”. The model generates an image in some gibberish language. We then prompt the model with words from this generated image, as shown in (b), (c). It seems that `Vicootes` means vegetables and `Apoploe vesrreaitais` means birds. It appears that the farmers are talking about birds that interfere with their vegetables.

Style Transfer. DALLE-2 is capable of generating images of some concept under different styles that can be specified in the prompt [2]. For example, one might ask for a photorealistic image of an apple or a line-art showing an apple. We test whether the words of the DALLE-2 language, (e.g. `Apoploe vesrreaitais`) correspond to visual concepts that can be transformed into different styles, depending on the context of the prompt. The results of this experiment are shown in Figure 3. It seems that the prompt sometimes leads to flying insects as opposed to birds.

(a) Prompt: “`Painting of Apoploe vesrreaitais`” (b) Prompt: “`cartoon, Apoploe vesrreaitais`” (c) Prompt: “`3-D rendering of Apoploe vesrreaitais`” (d) Prompt: “`line art, Apoploe vesrreaitais`”

Figure 3: Illustration of DALLE-2 generations for `Apoploe vesrreaitais` under different styles. The visual concept of “something that flies” is maintained across the different styles.

Language consistency with the text-conditioning and the generated image. Recall the example with the farmers. The prompt was: “`Two farmers talking about vegetables, with` subtitles.". From this example, we discovered the word vegetables, but also the word birds. It is very plausible that two farmers would be talking about birds and hence this opens the very interesting question of whether the secret language of DALLE-2 is consistent with the text conditioning and the generated image. Our initial experiments show that indeed it is quite frequent to have gibberish text that translates to visual concepts that match the text-conditioning that created the gibberish text in the first place. For example, the prompt: "`Two whales talking about food, with subtitles.`" generates an image with the text "`Wa ch zod ahaakes rea.`" (or at least something close to that). We feed this text as prompt to the model and in the generated images we see seafood. This is shown in Figure 3. It seems that the gibberish language indeed has a meaning that is aligned with the text-conditioning that produced it.

Figure 4: Left: Image generated with the prompt: "`Two whales talking about food, with subtitles.`". Right: Images generated with the prompt: "`Wa ch zod ahaakes rea.`". The gibberish language, "`Wa ch zod ahaakes rea.`", produces images that are related to the text-conditioning and the visual output of the first image.

4 Security and Interpretability Challenges

There are many interesting directions for future research. It was not clear to us if some of the gibberish words are mispellings of normal words in different languages, but we could not find any such examples in our search. For many of the prompts, the origins of these words remains confusing and some of the words were not as consistent as others in our preliminary experiments. Another interesting question is if Imagen [6] has a similar hidden language given that it was trained with a language model as opposed to CLIP. We conjecture that our prompts are adversarial examples for CLIP's [4] text encoder, i.e. the vector representation of `Apoploe vesrreaitais` is close to the representation of `bird`. It is natural to use other methods (e.g. white box) of adversarial attacks on CLIP to generate absurd text prompts that produce target images in DALLE2.

Our findings open some important interpretability and security challenges. Currently, Natural Language Processing systems filter text prompts that violate the policy rules. Gibberish prompts may be used to violate these filters. More importantly, absurd prompts that consistently generate images challenge our confidence in these big generative models.

5 Acknowledgements

The authors would like to acknowledge the Institute for the Foundations of Machine Learning (IFML) and the National Science Foundation (NSF) for their generous support. We also thank OpenAI for providing access to their model through the API.

References

[1] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021.

[2] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[3] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL `https://arxiv.org/abs/2103.00020`.

[5] Gary Marcus, Ernest Davis, and Scott Aaronson. A very preliminary analysis of dall-e 2, 2022.

[6] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.