Not All Artists Speak English: Generating images with DALL-E 2 from Portuguese

Gretchen M. Eggers
Group on Artificial Intelligence and Art (GAIA)
University of São Paulo - Fulbright U.S. Student Program
gretchen.eggers@usp.br

Abstract

DALL-E 2 supports basic input in a number of languages, such as Spanish, German and Portuguese. This is an exciting development for non-English speaking artists and designers who could benefit from using the software. However, prompts given in other languages sometimes lack accuracy and differ in stylistic choices. In this work, we explore the differences in aesthetic quality and accuracy of results produced by DALL-E 2 from paired text inputs in English and Portuguese.

Introduction

In April of 2022 OpenAI released a new version of their text to image software DALL-E 2 [1]. This update sparked international excitement among computer scientists and members of the art and design community for its significant improvements in image quality. In addition to these general improvements, DALL-E 2 appears to be capable of interpreting prompts from a wide variety of languages. However, no complete list of supported languages has been published and DALL-E 2's performance across languages is currently unknown.

Text to image generation models trained on internet data, like DALL-E 2, frequently learn basic multilingual capabilities as the scraped captions are likely to include examples from other widely spoken languages. Unfortunately these scraped data-sets tend to be heavily biased towards English and any learned multilingual abilities are considered coincidental. As a result, images produced from prompts in other languages are likely to suffer or differ in some aspect based on this imbalance, but are rarely used to evaluate performance. Despite not being intentionally trained to understand languages other than English, DALL-E 2 does a surprisingly good job with some examples being nearly indistinguishable from their English counterparts (Figure 1). In this work, we begin the process of understanding DALL-E 2's multilingual capabilities through qualitative and quantitative investigation of the differences between images generated from Portuguese and English.

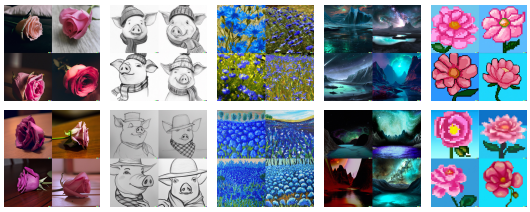


Figure 1: Images generated by DALL-E 2 in English (top) and Portuguese (bottom) that are nearly indistinguishable

Methods

At the time of this study, no representative data-set of images generated by DALL-E 2 was publicly available. Many successful prompts follow the basic structure of combining a subject, modifications to the subject and a style description. In order to ensure our data-set included a broad sample of

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

images, we chose five categories each of subjects, modifiers and styles from which to craft our prompts (see appendix A). Using this framework, we drafted 50 prompts in English and generated 4 images per prompt using DALL-E 2. Drafted prompts that didn't produce reasonably accurate images were edited or replaced. Prompts were then translated to Portuguese by a native speaker and re-generated, resulting in 400 images total (See appendix B for discussion on editing prompts).

We crafted a survey and crowd-sourced Brazilian Portuguese and English speakers to evaluate differences in human perception of aesthetics and accuracy between generated images. Respondents were not given the prompt when asked to choose based on aesthetics, but when asked to choose the more accurate group of images, we presented the prompt in the language(s) spoken by the respondent.

Using CLIP [2], we then created text and image embeddings for each English prompt and all generated images. To measure similarity between the Portuguese and English generated images, we calculated the mean pair-wise cosine similarity between English and Portuguese generated images. To measure to what extent the Portuguese generated images represented the same concepts as the English generated images, we calculated the mean cosine similarity between the image embeddings and the English prompt text embeddings. Following the work of Wolfe and Caliskan [3], we calculated the self-similarity within each Portuguese and English image group as a measure of concept consistency in generated images. We also performed zero-shot classification with CLIP on all 400 images and calculated the entropy of the normalized class probabilities to measure the ambiguity of the concepts portrayed in the generated images.

Results

165 respondents completed the survey with 56.4% of respondents speaking only English, 2.4% speaking only Portuguese and 41.2% speaking both Portuguese and English. Combined results from all language groups showed a statistically significant ($p < 0.001$) aesthetic preference for images generated by English prompts. Respondents also reported that images generated by English prompts were more accurate for a given prompt, regardless of prompt language presentation ($p < 0.001$).

Average aesthetic preferences did not vary significantly between English, Portuguese and bilingual respondents. However, respondents actually preferred photo-realistic Portuguese generated images and showed a stronger than average preference for English generated digital art images ($p < 0.05$). While respondents found English generated images to be more accurate across all styles and prompt language presentations, respondents had a weaker than average preference for English generated Van-Gogh and pixel art style images paired with Portuguese prompts ($p < 0.005$).

78% of respondents reported general differences in style, 54.9% reported differences in the color of objects, 50% reported differences in the race or ethnicity of people, 45.7% reported differences in the placement of objects and 22% reported differences in the size of objects. Several respondents reported that they noticed general differences in accuracy or that the proportion of images within a group that were accurate to the prompt were different. Selected examples of these reported differences can be found in Appendix C.

Mean pairwise cosine-similarity between CLIP embeddings of Portuguese and English generated images was 0.83, indicating that pairs of generated images are, on average, quite similar. No significant difference was found between prompt language for measures of self-similarity and for cosine similarity between image and English prompt text embeddings. Zero-shot classification showed slightly higher variability in identified classes for Portuguese generated images per prompt, but the average entropy in normalized class probabilities per image was not significantly different.

Conclusion

DALL-E 2 is capable of generating images from prompts in Portuguese that are similar to images generated in English with some consistent differences such as greater ethnic diversity in Portuguese generated images. However, people find English generated images to, on average, be more aesthetically appealing and more accurate across a wide range of styles and subjects. From these results, we conclude that non-English speaking artists and designers could expect reasonably high quality images generated from Portuguese text, but they may struggle with precisely producing their desired images and these images will differ stylistically from those generated by their English speaking counterparts.

Ethical Implications

In supporting basic multilingual input, DALL-E 2 has taken a great step in the right direction towards equal access to cutting edge AI technology for people of all languages. In a world increasingly influenced by AI, it is important that access to AI-powered tools not be restricted to English-speaking populations. However, the differences observed in images generated from other languages compared to English has some negative implications. Unreliable image quality from non-English prompts could result in artists and designers needing to use more requests, and therefore spend more money to generate the same quality of images using English prompts. Cultural norms tied to prompt language and differences in the proportions of races and ethnicities represented in non-English generated images could reinforce negative cultural standards and stereotypes.

As the work of Millière, Daras and Dimakis [4] suggests, multilingual support also comes with security risks. While their work looks at nonsensical and mixed language prompts explicitly designed to evade filters for violent, sexual and other potentially malicious prompts, this work reinforces their idea that these security holes could be exploited to make relatively precise and high quality images of harmful content. Certainly, more research is needed on the content filtering capabilities of DALL-E 2 when coherent malicious prompts are given in languages other than English.

Acknowledgements

Special thanks to Bruno Moreschi for help with the many translations this project required. We would also like to thank Prof. Fabio Cozman, Profa. Giselle Beiguelman, and the many friends who helped crowd source respondents to our survey. This research was funded by the U.S. Fulbright Student Program in Brazil and conducted at GAIA with the support of the Center for Artificial Intelligence (C4AI-USP), the Sao Paulo Research Foundation (FAPESP grant 2019/07665-4) and from the IBM Corporation.

Appendix A: Framework for Creating Prompts

DALL-E 2 is capable of generating images of a vast number of subjects and in many different styles. In order to create a data-set representing a broad range of concepts and styles, we identified 3 core components of a reasonably complex prompt: subject, modifier and style. This framework is also used on the main DALL-E 2 website (`https://openai.com/dall-e-2/`) in the demo section. We then chose five categories for each of the components as follows:

Table 1: Categories used in components of prompt framework

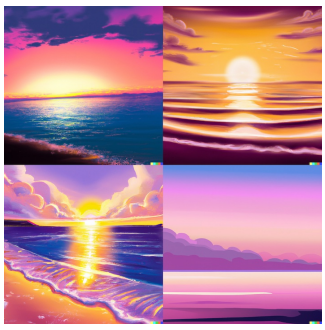
Subject	Modifier	Style
Humans	In a place	Digital Art
Animals	That looks like/Abstraction	a Pencil Sketch
Plants	Doing action	Photo Realistic
Landscapes	Number of subjects	Van Gogh Painting
Household Objects	Color	Pixel Art

For example, the prompt "A painting in the style of Van Gogh of three dogs sitting in a park" has the subject "dogs" from category "Animals", modifiers "Number of subjects" and "In a place", and uses the "Van Gogh Painting" style. Subject and style categories were used in equal proportions while drafting the prompts. Often, more than one modifier was needed to create an interesting image. At the same time, combining some modifiers can quickly result in grammar that is overly complex or ambiguous. We wanted this data-set to also represent examples that an average user of DALL-E 2 might create, so while the modifiers are not perfectly distributed across prompts, this was justified by the overall improvement in image quality.

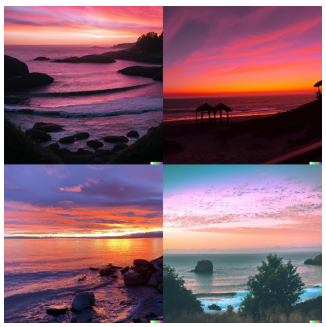
As this framework is not a true grammar, we also took care to use similar syntax wherever possible. DALL-E 2 is extremely reactive to the style category, so all prompts within a particular style were ensured to denote the style in the same way. For example, all prompts from the "Van Gogh Painting" category begin with "A painting in the style of Van Gogh".

Appendix B: Examples and Discussion of Edited Prompts

When submitting prompts to DALL-E 2, it is fairly common to edit or tweak a prompt while attempting to generate your desired image. From our original English prompt drafts, we ran into a couple of cases where the prompts needed to be amended. For example, we originally used "a photo realistic image" to denote our "Photo Realistic" category. This phrase actually produced an image that looked more like the "Digital Art" category.



(a) "A photorealistic image of a beach with a pink, purple and orange sunset"

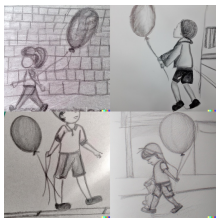


(b) "A photo of a beach with a pink, purple and orange sunset"

Figure 2: Selected images generated from draft vs. final prompts in English

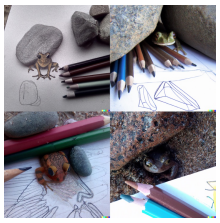
Another peculiar case involved the style designation of "digital art". Originally, prompts drafted in English for the "digital art" style ended with "as digital art". This proved ineffective and was changed to ", digital art".

After the prompts were translated to Portuguese there was only one major change in translation. Prompts from the "Pencil Sketch" category were translated to begin with "Um desenho a lápis", but were later changed to "Um esboço a lápis".



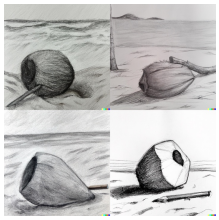
(a) "Um desenho a lápis de uma criança caminhando pela rua segurando um balão"

(b) "Um desenho a lápis de um coco abandonado na praia"



(c) "Um desenho a lápis de um sapo escondido perto de uma pilha de rochas"

(d) "Um esboço a lápis de uma criança caminhando pela rua segurando um balão"



(e) "Um esboço a lápis de um coco abandonado na praia"

(f) "Um esboço a lápis de um sapo escondido perto de uma pilha de rochas"

Figure 3: Selected images showing the varied results between two translations of "a pencil sketch" in Portuguese

"Um desenho a lápis" and "Um esboço a lápis" translate to "A pencil drawing" and "A pencil sketch", respectively. DALL-E 2's failure to properly understand the former is especially surprising given that, the word "desenho" is approximately 10 times more commonly used than "esboço" [6]. In the images above, we can see that DALL-E 2 was occasionally able to understand the usage of "desenho" (Figure 3a). However, it primarily produced photo-realistic images of the subject with misplaced references to pencils or drawings (Figure 3b, 3c). Changing the translation improved the results significantly, however we can still see an artifact of misinterpretation– DALL-E 2 continued to insert literal pencils into the images (Figure 3e, 3f).

Appendix C: Examples of Differences in Generated Images

Throughout the process of creating the data-set, several themes became apparent in the differences between English and Portuguese generated images from DALL-E 2. Many of these same themes were reported by survey respondents, even without the knowledge of which images were generated in which language.

Style

Changes in "style" of image can be interpreted broadly, and perhaps this explains why this category was most reported by survey respondents. Below are a few specific examples to consider:

(a) "A painting in the style of Van Gogh of an ornate tea cup"

(b) "A pencil sketch of two orchids growing on a tree"

(c) "A photo of two women wearing red dresses walking down a street"

(d) "Uma pintura no estilo de Van Gogh de uma xícara de chá ornamentada"

(e) "Um esboço a lápis de duas orquídeas que crescem numa árvore"

(f) "Uma fotografia de duas mulheres vestindo vestidos vermelhos caminhando por uma rua"

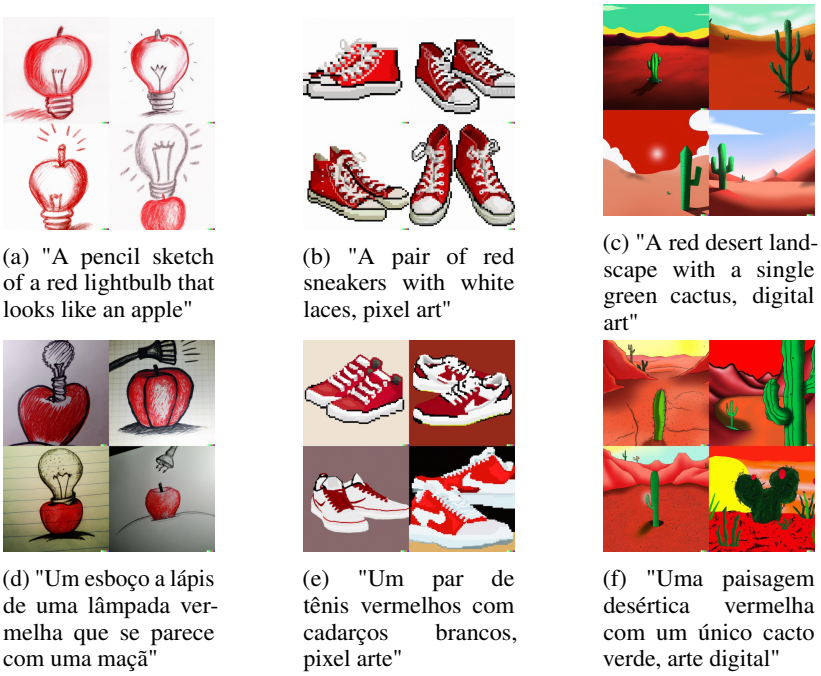
Figure 4: Selected images showing examples in style differences between Portuguese and English results

Portuguese prompts with the "Van Gogh Painting" style generated images that looked slightly less like quintessential Van Gogh (Figure 3a, 3d). These differences were "all or nothing", with all 4 images from a prompt consistently showing the same "Van Gogh" style, or lack there of. In the "Pencil Sketch" style, DALL-E 2 fairly consistently generated one image (out of 4 per prompt) from Portuguese that appeared more like a pen drawing. These images also tended to include shadows, literal pencils or edges of pages that gave the overall effect of a photo of a pencil drawing, as opposed to a high quality scan. The English generated pencil sketches, never included pen drawings, looked more like high quality scans and were stylistically more consistent (Figure 3b, 3e).

One of the most interesting examples of style differences was the changes in architecture style. English prompts produced buildings more consistent with American urban buildings, while Portuguese prompts produced buildings in classic Portuguese architecture style (Figure 3c, 3f). In addition, there are changes in the pavement and cars that overall give general impressions of an suburban American street vs. a suburban Brazilian street.

Variable Binding and Accuracy

Reported differences in in color, object placement, size and accuracy may be explained by reduced ability in complex variable binding with prompts written in Portuguese. Word order, particularly the placement of adjectives, varies significantly between English and Portuguese (and in Romance and Germanic languages in general) [7]. Due to the training bias for English prompts, its possible that these differences in word order contribute to general accuracy and variable binding issues. DALL-E 2's variable binding in English is by no means perfect, but inaccuracies were far more frequent in Portuguese generated images.



(a) "A pencil sketch of a red lightbulb that looks like an apple"

(b) "A pair of red sneakers with white laces, pixel art"

(c) "A red desert landscape with a single green cactus, digital art"

(d) "Um esboço a lápis de uma lâmpada vermelha que se parece com uma maçã"

(e) "Um par de tênis vermelhos com cadarços brancos, pixel arte"

(f) "Uma paisagem desértica vermelha com um único cacto verde, arte digital"

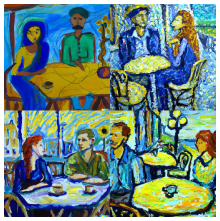
Figure 5: Selected images showing examples of frequent failed variable binding in Portuguese results and comparative success in English results

Race and Ethnicity

Images generated in Portuguese including people were more racially and ethnically diverse than images generated in English. Although English generated images did regularly include non-white people, prompts written in Portuguese produced images with non-white people at a greater frequency. This may be explained by demographic differences in Portuguese and English speaking populations that are likely reflected in DALL-E 2's training data [8,9].

(a) "A photo of three children sitting on a blue couch and smiling"

(b) "A black and white photo of a child on a school bus staring out the window"



(c) "A painting in the style of Van Gogh of a man and a women sitting at a cafe"

(d) "Uma fotografia de três crianças sentadas num sofá azul e sorrindo"

(e) "Uma fotografia preto e branco de uma criança num ônibus escolar olhando pela janela"

(f) "Uma pintura no estilo de Van Gogh de um homem e uma mulher sentados num café"

Figure 6: Selected images showing examples of differences in racial diversity of humans in English vs. Portuguese generated photos

Appendix D: Future Work

This study is limited in size and scope, investigating only one of the many languages DALL-E 2 likely understands and does not evaluate DALL-E's out-painting or in-painting abilities in Portuguese. Much more research is needed on DALL-E 2's (and other text to image software's) abilities in other widely spoken languages. Given that these multilingual capabilities are largely coincidental, research into how DALL-E 2 is capable of understanding languages on which it was not intentionally trained would be very interesting, as it may provide a framework for extending other text-based AI systems to other languages with a relatively small additional amount of training examples.

Specific examples of cultural differences captured by language in the images generated in this study (see Appendix C) could also be an interesting line of research. What artists and designers are influencing the stylistic differences we observed when we translated our prompts to Portuguese? Does DALL-E 2 also mirror the demographics of Spanish, German or other non-English speaking populations when given prompts in those languages? These questions dig into the heart of what it means to design creative tools using AI in an increasingly globalized world, and certainly warrant further investigation.

References

[1] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. *Hierarchical Text-Conditional Image Generation with CLIP Latents*, 2022.

[2] 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.

[3] Robert Wolfe and Aylin Caliskan. *Markedness in Visual Semantic AI*, 2022.

[4] Raphaël Millière. *Adversarial Attacks on Image Generation With Made-Up Words*, 2022.

[5] Giannis Daras and Alexandros G. Dimakis. *Discovering the Hidden Vocabulary of DALLE-2*, 2022.

[6] Corpus do Português. Davies, Mark. (2016-) Corpus do Português: Web/Dialects. Available online at http://www.corpusdoportugues.org/web-dial/.

[7] Freek Van de Velde , Petra Sleeman and Harry Perridon. *THE ADJECTIVE IN GERMANIC AND ROMANCE DEVELOPMENT, DIFFERENCES AND SIMILARITIES*, 2014

[8] "IBGE Censo 2010". censo2010.ibge.gov.br

[9] "United States Census Bureau, 2021" https://www.census.gov/quickfacts/fact/table/US/PST045221