

Towards Educating Artificial Neural Systems

Simon Colton^{1,2,*†}

¹*School of Electronic Engineering and Computer Science, Queen Mary University, London, UK*

²*SensiLab, Faculty of Information Technology, Monash University, Australia*

Abstract

In the context of text-to-image generators, we discuss how a neural system could be enhanced to follow a knowledge base expressing certain moral considerations, in order to address some looming concerns. We also propose self-educating procedures that enable the system to produce rule-based approximations of its functioning, and illustrate this with an application to the creative interpretation of images.

Keywords

Pre-trained Neural Models, Text-to-image Generation, Computational Creativity

1. Introduction


Until relatively recently, neural image generators were developed for largely utilitarian purposes, for instance BigGAN [1] was trained to generate novel images in 1 of 1,000 classes such as hamburgers or terriers. Moreover, the engineers of such systems would lament issues such as *class leakage* where an image looks like it is from two classes, e.g., an image of a tennis ball that looks like a yellow bird. Text-to-image generators [2] were likewise used largely to generate images which could exist in reality, such as “a red bird in a tree”.

Over the last 18 months, the situation has dramatically changed and one of the main usage of neural image generators is currently to produce highly imaginative images which could never exist in reality. Huge neural models with billions of nodes and training costs in the millions of dollars from OpenAI, Google and others have been produced which can, on demand, generate images quickly with high fidelity to a given text prompt, no matter how outlandish and imaginative the prompt is. As examples, figures 1(a) and 1(b) portray generated images from the DALL-E 2 (OpenAI) [3] and Imagen (Google) [4] approaches, which currently produce the highest-quality images in terms of fidelity to the prompt and coherence.

In [5], we predicted that *generative search engines* will soon revolutionize the creative industries, where text-to-image generators like Imagen and DALL-E 2 will be used much like Google image search currently is, i.e., with images being constructed, rather than retrieved, to fit a search term. This prediction is still sound, but there are, however, hurdles to overcome in making such generative processes available to the public. Access to both DALL-E 2 and Imagen was initially restricted, and there are limitations on the usage, e.g., users are not allowed to share images from DALL-E 2 that portray realistic faces. This has led to some users being asked to remove generated images from social media platforms.

Workshop on Neurosymbolic Learning and Reasoning

 <https://imaginative.ai/> (S. Colton)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).


 CEUR Workshop Proceedings (CEUR-WS.org)



Figure 1: (a) Image generated by DALL-E 2 for: “Velociraptor in a suit, studio, portrait, dark bg, detailed” (source: twitter.com/Dalle2Pics) (b) Image produced by Imagen for: “Oriental painting of tigers wearing VR headsets during the Song dynasty” (source: twitter.com/hardmaru) (c) Five Images generated by DALL-E 2 for “a builder” (source: github.com/openai/dalle-2-preview/blob/main/assets).

We look here at a particular problem with text-to-image generation, namely the lack of diversity in the images generated due to biases trained into the neural models at the heart of the systems. We explore the position that large models trained on internet-sized datasets of images/text provide breakthrough opportunities for content generation, but should be guided by rule-based systems in a neurosymbolic approach. We propose a straightforward way in which a system employing a neural generator could be further *educated* (rather than merely trained on data), with rules representing moral positions which may directly contradict the training data. We also suggest how symbolic knowledge can be extracted from such models to highlight biases and semantic connections. To illustrate the benefits of this, we present some results from an application to the creative interpretation of images.

2. Guiding Image Generation with Moralbases

When asked to generate multiple images for the prompt “a builder”, DALL-E 2 produces pictures of men only, as per figure 1(c). This is because the models DALL-E 2 employs have been trained on millions of image/text pairs where the overwhelming majority of faces associated with the word “builder” are male. If asked to do the same, many, if not most, young children would draw five men in a row if asked to portray five builders. Starting from TV characters such as Bob the Builder, through every trip past a building site and all but the most inclusive of books, builders are – in the experience of children – overwhelmingly male in nature. In the terminology of machine learning, children are *trained* with a bias towards portraying builders as male.

While it may be disappointing to progressive parents or teachers that a child expresses lack of diversity when portraying certain professions such as builders, this offers an opportunity to *educate* children with a more general lesson which may expressly contradict the evidence of their own experience. In particular, parents can tell a child that a person of any gender can be a builder, and/or they may offer a more general lesson that, in principle, a person from any background can do any job (perhaps with additional training). In future, if a child so educated is asked to draw a builder, they may recall this lesson and draw a female builder from their imagination rather than their experience.

A purely deep-learning approach to address biases baked into large neural models would be to curate and balance the training data e.g., to make sure there are an equal number of male and female builders represented. While this might be feasible for training a generative model such as a GAN [6] specifically for producing images of a certain type, it would be impossible

to balance the data required for a general-purpose text-to-image generator where the subject material could reflect any aspect of human life, real or imagined. For instance, for builders alone, thousands of images of every type of human varying over gender, ethnicity, (dis)ability, age, etc. would be needed to ensure that images are generated in an inclusive way.

Scraping internet-scale datasets for training is very convenient, and it is already difficult/impossible for organisations to remove offensive images from the training set; it seems highly implausible to imagine attempting to balance training data for large models. Moreover, this would miss one of the big advantages of text-to-image generators, namely that they can be directed to produce images that are not representative of their training data. That is, while DALL-E 2 and Imagen may by default produce a white, male face in a yellow builder’s hat when prompted with “a builder”, when prompted instead with “a female builder”, this would override the default and deliver an image of a woman dressed as a builder. A position we take is that the pre-trained neural models in text-to-image generators are equivalent to naïve children who have generalised from what’s around them, but who need to be educated with higher-level rules which may contradict their experience. That is, rather than blaming the neural model, the data or the training regime for the inherent bias that it can be used to portray, we should blame (and fix) the processes surrounding the employment of the neural model for the bias.

We propose that neural systems employing pre-trained large image/text models for image generation should be guided by what could be called a *moralbase* of high-level rules. In particular, these rules could be employed for automated *prompt engineering* [7], whereby a user’s prompt is altered before it is passed to the generative engine. One can imagine a simple approach where certain abusive words are removed from prompts, changed to non-offensive versions, or where the system refuses to process any prompt with such words in, as is our approach for the @artbot text-to-image Twitter bot [8]. To begin to address lack of diversity in image generation, we can imagine that a moralbase could encode rules such as $builder(X) \rightarrow man(X) \vee woman(X) \vee nonbinary(X)$. Alternatively, we could express rules at a higher level, for example as follows:

$$\begin{array}{l} is_person(X) \wedge is_profession(Y) \wedge has_profession(X, Y) \rightarrow has_gender(X, G) \\ has_gender(X, G) \rightarrow male(X) \vee female(X) \vee nonbinary(X) \\ is_profession(builder) \end{array}$$

Another alternative would be to use a formalism such as stochastic logic programming [9] to represent moral considerations probabilistically. The process could then derive a set of re-write rules so that prompts can be engineered on a rotating basis to increase the diversity of outputs. For instance, in one run, “builder” could be replaced by “male builder” and in another by “female builder”. Complex natural language processing and re-writing of prompts may be needed for this, but we have found that simply appending a word like “man” or “woman” to a prompt can lead to images portraying the gender required.

In the same way that it is not feasible to balance data and retrain a neural model to remove biases, it is not feasible to imagine a prompt engineering scheme which can handle all biases which may present. Hence, we expect that moralbases will be highly focused and probably specific to a particular user, organisation, project, political leaning, etc., and that individual users could build up multiple moralbases to switch between. We also expect that automated reasoning, constraint solving, planning and other classical AI techniques will be required for moralbases to be used to their full potential, and we are currently experimenting to this effect.

3. Extracting Semantic Knowledge from Neural Models

The building up and deployment of moralbases could be time consuming and difficult, and some automation may improve matters by identifying biases and relationships in the pre-trained model that the moralbase guides. To make initial proposals for this, we have recently experimented with an implementation of CLIP-VQGAN which can perform text-to-image generation, as described in [10]. Here, latent vector inputs for VQGAN are found using gradient descent, in such a way that the image reflects a given text prompt. CLIP [11] comprises two models for encoding images and text respectively into the same latent space, so pairs of images and text (I, T) are encoded to vectors with a smaller cosine distance between them if I reflects T than if I has no relation to T . The loss function for the gradient descent is based on the average cosine distance between CLIP encodings of subimages of the VQGAN-generated images and the CLIP encoding of a given text prompt. This works in a similar way to CLIP-Guided BigGAN [5]. CLIP has encoded a vast amount of visual and textual information, and we can mine semantic relationships it has learned. For instance, given a list of colours, it is straightforward to get CLIP to identify that the word “chocolate” has the smallest cosine distance to “brown”, “sun” is closest to “yellow”, etc. Such relationships can be extracted as symbolic knowledge to aid in constructing a moralbase, and can be seen as self-education of a pre-trained neural model.

As an artistic application of this kind of semantic extraction, we generated some *creative interpretations* of images. To do this, we implemented a process which finds both nouns and adjectives closest (in terms of cosine distance of CLIP encodings) to an image, then (as above), we extract CLIP relationships between the adjectives and nouns to produce sentences which caption the image. Given an image and a caption, CLIP-VQGAN can generate a novel image which reflects both the original image and the caption. We used this functionality to produce a series of creative interpretations given in the appendix. As an example, the system has interpreted/captioned the first original image in the appendix as a “worn and forsaken ruin” and produced a new image accordingly, which also references strongly the original. Social media and the internet are becoming saturated with text-to-image produced imagery, and, artistically, there are rapidly diminishing returns to the production of such images. Our aim with such creative interpretations, and other more sophisticated generative projects we have planned, is for generated images to be seen as artworks with an aura [12], to be interpreted as part of a thought provoking conversation, rather than merely to accurately portray a given prompt.

4. Conclusions and Future Work

With this position paper, we propose that neurosymbolic systems, especially those generating novel content, could reason over knowledge bases of moral considerations which educate them with higher-level lessons. We further suggest automatically extracting symbolic knowledge from pre-trained models to aid in this. We are currently implementing prompt engineering processes to employ in the context of the @artbhot [8] twitter bot [13]. Ultimately, we aim to treat pre-trained neural models as fixed black-boxes, performing symbolic automated theory formation [14] to derive rule-based approximations of some of the semantic relationships they capture, in order to better understand and control their usage.

Acknowledgments

We would like to thank Amy Smith for insightful comments on the web of meaning into which artworks are situated. We would also like to thank Katherine Crowson and to the @nerdyrodent developer for their work on the CLIP-VQGAN colab notebook and github repository. We would also like to thank the anonymous reviewers for their insightful comments.

References

- [1] A. Brock, J. Donahue, K. Simonyan, Large Scale GAN Training for High Fidelity Natural Image Synthesis, arXiv:1809.11096, 2019.
- [2] J. Agnese, J. Herrera, H. Tao, X. Zhu, A Survey and Taxonomy of Adversarial Neural Networks for Text-to-Image Synthesis. arXiv:1809.11096, 2019.
- [3] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with CLIP latents. arXiv:2204.06125, 2022.
- [4] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, M. Norouzi, Photorealistic text-to-image diffusion models with deep language understanding. arXiv:2205.11487, 2022.
- [5] S. Colton, A. Smith, S. Berns, R. Murdock, M. Cook, Generative search engines: Initial experiments, in: Proceedings of the International Conference on Computational Creativity, 2021.
- [6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Nets, in: Advances in Neural Information Processing Systems, volume 27, Curran Associates, Inc., 2014.
- [7] V. Liu, L. B. Chilton, Design guidelines for prompt engineering text-to-image generative models, arXiv:2109.06977, 2021.
- [8] A. Smith, S. Colton, The @artbhot text-to-image twitter bot, in: Proceedings of the International Conference on Computational Creativity, 2022.
- [9] S. Muggleton, Learning stochastic logic programs, Electronic Transactions in Artificial Intelligence 4 (2000).
- [10] K. Crowson, S. Biderman, D. Kornis, D. Stander, E. Hallahan, L. Castricato, E. Raff, VQGAN-CLIP: Open domain image generation and editing with natural language guidance, arXiv:2204.08583, 2022.
- [11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision, arXiv:2103.00020, 2021.
- [12] W. Benjamin, The Work of Art in the Age of Mechanical Reproduction (Original essay in Zeitschrift für Sozialforschung, 1936), Penguin, 2008.
- [13] T. Veale, M. Cook, Twitterbots: Making Machines that Make Meaning, MIT Press, 2018.
- [14] S. Colton, Automated Theory Formation in Pure Mathematics, Springer-Verlag, 2002.

Appendix: Example Creative Interpretations



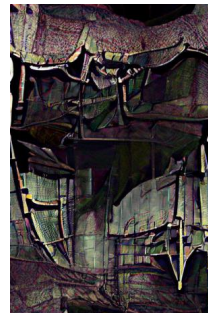
Original "Truth"
by Isaac Ganuza



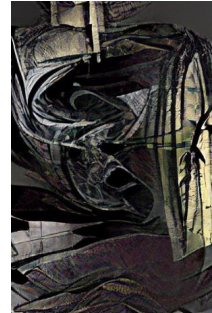
"Fluid and
Threadbare Web"



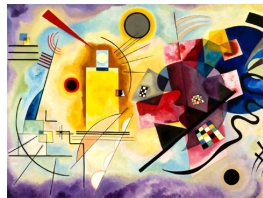
"Wiry and
Torn Rope"



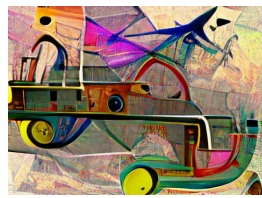
"Worn and
Forsaken Ruin"



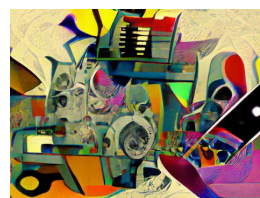
"Wrathful and
Ancient Spirit"



Original "Yellow, Red,
Blue" by Wassily
Kandinsky



"Harmonious and
Imaginative
Transportation"



"Offbeat and
Grandiose
Machine"



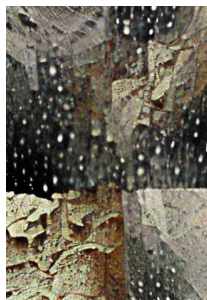
"Energetic and
Immense
Revolution"



Original
"Daniel-Henry
Kahnweiler" by
Pablo Picasso



"Triangular
Structure"



"Granular and
Dreary Rain"



"Quarrelsome
Confusion"



"Strident and
Jagged Shift"