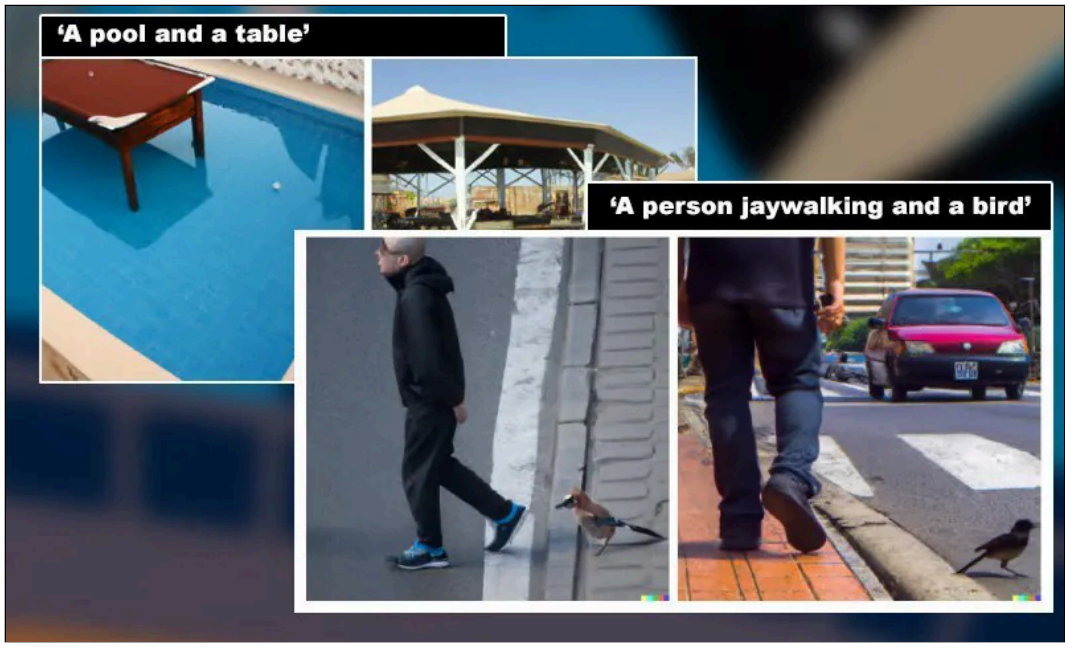
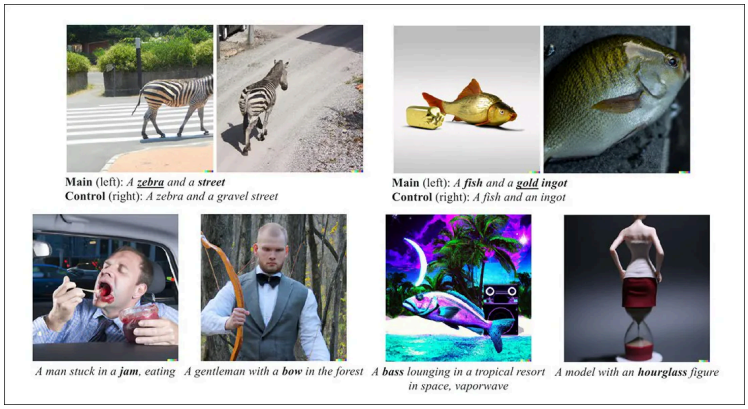
DALL-E 2's Unique Solution to Double Meanings

Originally published at https://www.unite.ai/dall-e-2s-unique-solution-to-double-meanings/



Anyone who has learned Italian learns early to pay attention to context when describing a *broom*, because the Italian word for this mundane domestic item has an extremely NSFW second meaning as a verb*. Though we learn early to disentangle the semantic mapping and (apposite) applicability of words with multiple meanings, this is not a skill that is easy to pass on to hyperscale image synthesis systems such as DALL-E 2 and Stable Diffusion, because they rely on OpenAI's Contrastive Language–Image Pre-training (CLIP) module, which treats objects and their properties rather more loosely (yet which is gaining ever more ground in the latent diffusion image and video synthesis space.

Studying this shortfall, a new research collaboration from Bar-Ilan University and the Allen Institute for Artificial Intelligence offers an extensive study into the extent to which DALL-E 2 is disposed towards such semantic errors:



Double-meanings split out into multiple interpretations in DALL-E 2 – though any latent diffusion system can produce such examples. In the upper right image, ren the prompt changes the species of fish, while in the case of the 'zebra crossing', it's necessary to explicitly state the road surface in order to remove the duplicatec
Source: https://export.arxiv.org/pdf/2210.10606

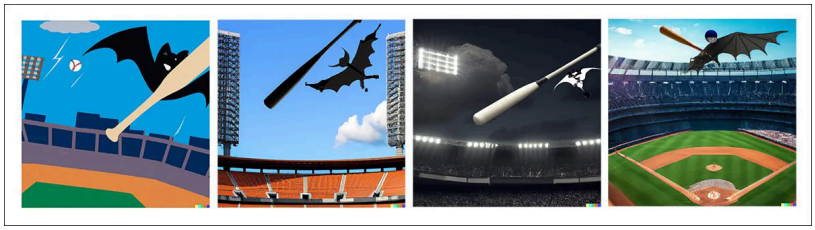
The authors have found that this tendency to double-interpret words and phrases seems not only to be common to all CLIP-guided diffusion models, but that it gets worse as the models are trained on higher and higher amounts of data. The paper notes that 'reduced' versions of text-to-image models, including DALL-E Mini (now Craiyon) output these kinds of errors far less frequently, and that Stable Diffusion also errs less – though only because, very often, it does not follow the prompt at all, which is another kind of error.

The simple prompt 'date' forces DALL-E 2 to invoke two of the several meanings of the word, while the word 'fan' also splits into two of its semantic mappings, and in the third image, the phrase 'cone' reliably turns the otherwise unspecified food in the prompt into ice cream, which is associated with 'cone'.

Explaining how we perform efficient lexical separations, the paper states:

> *'While symbols – as well as sentence structures – may be ambiguous, after an interpretation is constructed this ambiguity is already resolved. For example, while the symbol bat in a flying bat can be interpreted as either a wooden stick or an animal, our possible interpretations of the sentence are either of a flying wooden stick or a flying animal, but never both at the same time. Once the word bat has been used in the interpretation to denote an object (for example a wooden stick), it cannot be re-used to denote another object (an animal) in the same interpretation.'*

DALL-E 2, the paper observes, is not constrained in this way:



'A bat is flying over a baseball stadium' – the first image is from the paper, the other three obtained from simply feeding the same prompt into DALL-E 2.

This property has been named *resource sensitivity*.

The paper identifies three aberrant behaviors exhibited by DALL-E 2: that a word or a phrase can get interpreted and effectively bifurcated into two distinct entities, rendering an object or concept for each in the same scene; that a word can be interpreted as a modifier of two different entities (see the 'goldfish' and other examples above); and that a word can be interpreted simultaneously as both a modifier and an alternate entity – exemplified by the prompt *'a seal is opening a letter'*:

'A seal is opening a letter' – the first illustration is from the paper, the adjacent three, identical reproductions from DALL-E 2. The photoreal examples below had th Canon50, 85mm, F5.6, award-winning photo'.

The authors identify two failure modes for diffusion models in this respect: that the results of user prompts with sense-ambiguous words will often exhibit the concretized word together with some manifestation of the concept; and *concept leakage*, where the properties of one object 'leak' into another rendered object.

> *'Taken together, the phenomena we examine provides evidence for limitations in the linguistic ability of DALLE-2 and opens avenues for future research that would uncover whether those stem from issues with the text encoding, the generative model, or both. More generally, the proposed approach can be extended to other scenarios where the decoding process is used to uncover the inductive bias and the shortcomings of text-to-image models.'*

Using 17 words that will cause DALL-E 2 to split the input into multiple outputs, the authors observed that homonym duplication occurred in over 80% of 216 images rendered.

The researchers used stimuli-control pairs to examine the extent to which specific and arguably over-specified language is necessary to stop these duplications occurring. For the entity-to-property tests, 10 such pairs were created, and the authors note that the stimuli prompts provoke the shared property in 92.5% of cases, whereas the control prompt only elicits it in 6.6% of cases.

> *'[To] demonstrate, consider a zebra and a street, here, zebra is an entity, but it modifies street, and DALLE-2 constantly generates crosswalks, possibly because of the zebra-stripes' likeness to a crosswalk. And in line with our conjecture, the control a zebra and a gravel street specifies a type of street that typically does not have crosswalks, and indeed, all of our control samples for this prompt do not contain a crosswalk.'*

Main (left): *A **zebra** and a **street**.*
Control (right): *A zebra and a gravel street.*

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More