Evaluating the influence of ML models on human judgment of non-physical attributes in images

Shruthi Sukumar[1,2], Vijay Veerabadran[3], Jascha Sohl-Dickstein[1], Michael Mozer[1], Gamaleldin Elsayed[1];

[1]Google Research, Brain Team, [2]University of Colorado Boulder, [3]University of California San Diego

Recent breakthroughs in computer vision have given us generative models that can create visual stimuli from natural language descriptions or prompts (e.g., 'photorealistic image of blue alligator riding a small tricycle'). These models are often trained on datasets with multiple modalities such as vision and language and learn from data labeled by humans that not only includes physical attributes (e.g., object classes) but also human judgements of non-physical traits (such as pleasantness). Here, we investigate whether these models can generate subtleties in images that do not have simple physical manifestations, including affective state, in a way that is detectable by humans. We hypothesize that these models will likely be successful in encoding those non-physical attributes and elicit human judgment in agreement with those attributes. To this end, we first utilized a text-to-image generative model to generate an image corresponding to a particular attribute, such as 'pleasant', or 'calm', embedded within the natural language text prompt. Next, we conducted an empirical investigation on human participants and asked them to rate these generated images based on the degree to which they perceived whether an image reflects an attribute, like pleasantness, on a 5-point likert scale (very unpleasant to very pleasant). We then assess the degree of alignment between participants' ratings and the attribute used to generate the corresponding images. Our findings show that there is significant alignment between the attributes used to generate images and subjects' judgment of the image attributes ($p<0.001$), indicating that machine learning models can indeed elicit required judgments of non-physical attributes in visual stimuli. This work illustrates that machine-learning models are capable of representing subtleties in visual stimuli that can reliably influence human perception regarding non-physical attributes. Our experimental protocol was granted an Institutional Review Board (IRB) exemption by an external, independent, ethics board (Advarra)

Acknowledgements: Pieter Kinder, Alex Alemi, Ben Poole, Been Kim, Isabelle Guyon, Paul Nicholas, Patrick Gage Kelley, Bhavna Daryani