An Automated Metric for Surprisal

Spring 2023

Figure 1: Would you guess that none of these images are real?

When generating images, whether *de novo* or from an existing image (c.f. style transfer), it's not clear how one should objectively evaluate the quality of the results. One metric, *Inception Score* [1], is the entropy of the label distribution produced by feeding the generated sample through an Inception network [2]. Although it is proposed that this metric matches human evaluations of generated images, this isn't evaluated. One possible project is to run a large-scale Mechanical Turk study in which you compare correlation of human satisfaction/surprisal with Inception Score (or some other metric of your own design).

Another project of high value would be to use MTurk to collect a dataset of generated images (of a particular category) with annotations for the regions of the image that humans find displeasing (kind of like eye-tracking for computing saliency maps). You might then train a model to predict the bad parts of a generated image and use this as a supervision signal for a generative model.

References

[1] Tim Salimans et al. "Improved Techniques for Training GANs". In: *CoRR* abs/1606.03498 (2016). URL: `http://arxiv.org/abs/1606.03498`.

[2] Christian Szegedy et al. "Rethinking the Inception Architecture for Computer Vision". In: *CoRR* abs/1512.00567 (2015). URL: `http://arxiv.org/abs/1512.00567`.