Towards a Principled Evaluation of Likeability for Machine-Generated Art

Lia Coleman[1], **Panos Achlioptas**[2] , **Mohamed Elhoseiny**[12]
[1] KAUST [2]Stanford University
lia.coleman@kaust.edu.sa, optas@stanford.edu, mohamed.elhoseiny@kaust.edu.sa

Abstract

Creativity is a cornerstone of human intelligence and perhaps its most complex aspect. Thus, it is very interesting to understand how AI is already being used by professionals in creative domains like the arts and fashion. Namely, do artists actually *like* AI-generated "paintings"? In this study we collect and analyze responses on these questions from various contemporary artists and compare them to more naive, crowd-sourced ones. We highlight the importance of considering artists' opinion when evaluating AI-based art, and present a promising approach for researchers to do this easily.

Background. Computational creativity attempts to generate original content that is both realistic and aesthetic [5, 7, 1]. Generative Adversarial Networks (GANs) [3, 8, 4] are often the model of choice; however, the classic GAN training objective does not promote the production of novel content beyond the training data. A GAN trained on artwork will generate Da Vinci's "Mona Lisa" again, but it will not produce a new painting. However, recent work has been able to encourage GANs to produce novel images. Inspired by [6] Elgammal [2] adapted GANs to generate new novel paintings by encouraging the model to deviate from existing art styles. In fashion, Sbai [9] developed a model that generated an unseen fashion design. This is done by Elgammal and Sbai by adding an additional head on the GAN's Discriminator D, which predicts the *class* of an image. The Generator is then encouraged to not only generate real-looking examples, but also examples which are hard for D to assign a class to. More concretely:

$$\mathcal{L}_G = \mathcal{L}_{G \text{ real/fake}} + \lambda \mathcal{L}_{G \text{ creativity}} \tag{1}$$

Methods. We train a CAN model using multiclass cross entropy for our loss as in [9], but on the WikiArt dataset[1] for art instead of for fashion. After observing that our network produces novel artwork, we perform a human evaluation study on 120 images. One half of the 120 images are synthetic generations; the remaining half are art sourced from the contemporary art movements of minimalism, abstract expressionism, and art basel. We ask a group of 13 professional artists to: (1) rate on a scale of 1-5 the likeability of the artwork, and (2) guess whether the artwork was created by a human artist or generated by a computer (i.e., *do a Turing test*). We also collect 5 responses per image from Amazon Mechanical Turk raters, with the intention of scaling our evaluation by using artist data to validate the MTurk responses. These responses came from 82 distinct turkers across the 120 images.

Given our dataset of 120 images, we simulate real-world evaluation conditions. We split the 120 images into two sets: the 'seen' set, which we presume we have labels from artists on, and the 'unseen' set, which we presume we cannot get labels from artists on (because the size of the set is too large, etc.). We split our 120 images into these two sets randomly with a 60%-40% split. Our evaluation procedure is then as follows: From the 'seen' set, calculate the Cohen's Kappa (a measure

[1]https://www.wikiart.org/

32nd Conference on Neural Information Processing Systems (NeurIPS 2019), Montréal, Canada.

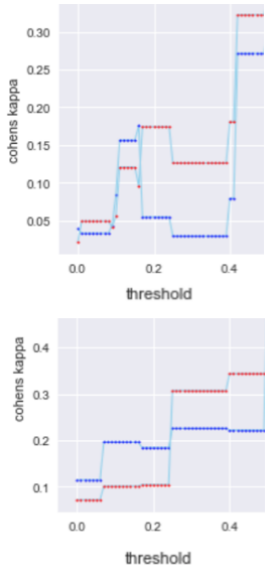

Figure 1: Gain in Cohen's Kappa, before (blue) and after (red) filtering at a particular threshold. 'Seen' (top) vs. 'unseen' (bottom) data.

Figure 2: Extreme samples of likeability.

of similarity) between each turker's likeability responses and the artist majority vote on likeability. Choose a Cohen's Kappa threshold to exclude turkers from the 'unseen' set who deviate from artists the most.

Results. We show that using Cohen's Kappa to filter out turkers who deviate from artists the most on the 'seen' set, results in better turker overall performance with the artist majority vote on unseen images. This implies that when performing evaluation of AI-generated artwork, we can use abundant turker labels to supplement the scarce labels of artists. In more detail: from the graphs in Figure 1, we see that 0.2 is the best threshold from the 'seen' 60% set, in terms of the amount of gain in Cohen's Kappa from the filtering. Then, we also see that 0.2 as a threshold performs well in the 'unseen' 40% set, as well.

For qualitative examination, we present the most-liked and least-liked works generated by our network in Figure 2. We find that our best machine-generated image is on par with the 3rd best human-created artwork, both attaining 75% of artists' votes. The fact that this margin is small is very promising.

Furthermore, we investigate the ability of a simple linear classifier to predict the likeability of a holdout set of test images. To achieve this goal, we exploit the semantically rich feature space provided by a VGG-16 [10] neural network, pretrained on ImageNet [2]. We use this network to embed each of the 120 images in a 4096D space (using the penultimate (fc7) layer), and in this space we train a linear SVM on a binary "likeability" problem. We use MTurk labels as our training data. Namely, an image is treated as a positive example if a strong majority (4 out of 5 Turkers) cast a positive score for likeability (4-5) and negative if they cast a negative score (1-3). These conditions hold for 70.8% of the data, while the SVM achieves $78.2\% \pm 0.4\%$ test accuracy under a 10-fold cross-validation. This is an encouraging (and statistically significant) result which indicates that some aspects of human preference in creative arts are shared and learnable.

Conclusion. Overall, we show that using MTurk to help scale artist responses on likeability for unseen images is a viable approach. We also show promising results on likeability of our network's novel artwork, and present preliminary results that likeability of artwork could be learnable.

[2] www.image-net.org

Ethical Implications. Our research involves obtaining labels from humans. We obtain full consent of our participants beforehand, and use their responses only for our research as we originally stated. Our expert evaluators opt-in as volunteers; Amazon MTurkers opt in and were compensated for their time. Additionally, our dataset, WikiArt, consists of artworks that are open-source.

Acknowledgements. We want to thank the artists who helped us with our research: Brooke Cheng, Dwayne Jones, Joseph Wilk, Luisa Fabrizi, Mark Hernandez, Taís Mauk, Michelle Cheung, Julia Peter, Francisco Rojo, Mathilde Mouw-Rao, Iain Nash, and Achim Koh.

References

[1] S. DiPaola and L. Gabora. Incorporating characteristics of human creativity into an evolutionary art algorithm. *Genetic Programming and Evolvable Machines*, 10(2):97–110, 2009.

[2] A. Elgammal, B. Liu, M. Elhoseiny, and M. Mazzone. Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms. In *International Conference on Computational Creativity*, 2017.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[4] D. Ha and D. Eck. A neural representation of sketch drawings. *ICLR*, 2018.

[5] P. Machado and A. Cardoso. Nevar–the assessment of an evolutionary art tool. In *Proc. of the AISB00 Symposium on Creative & Cultural Aspects and Applications of AI & Cognitive Science*, volume 456, 2000.

[6] C. Martindale. *The clockwork muse: The predictability of artistic change.* Basic Books, 1990.

[7] A. Mordvintsev, C. Olah, and M. Tyka. Inceptionism: Going deeper into neural networks. *Google Research Blog. Retrieved June*, 2015.

[8] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[9] O. Sbai, M. Elhoseiny, A. Bordes, Y. LeCun, and C. Couprie. Design: Design inspiration from generative networks. In *ECCV workshop*, 2018.

[10] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.