Evaluating the Category Representations in Conditional Generative Adversarial Networks via Human Category Learning

Victor Navarro (navarrov@cardiff.ac.uk)
School of Psychology, Cardiff University, 70 Park Place
CF10 3AT, Cardiff, UK

Christoph Teufel (teufelc@cardiff.ac.uk)
Cardiff University Brain Research Imaging Centre (CUBRIC) and
School of Psychology, Cardiff University, 70 Park Place
CF10 3AT, Cardiff, UK

Abstract

Generative adversarial networks (GANs) hold great potential as a tool for cognitive scientists, equipping researchers with the ability to generate a theoretically infinite number of complex stimuli. For this purpose, a good understanding of the correspondence between human and GAN representations is critical. In the present work, we evaluated the category representations developed by conditional GANs using human category learning. Specifically, we asked whether humans can learn to categorize class-specific GAN-generated samples, and if so, whether they can generalize that knowledge to real samples. Two groups of participants first learned to categorize either real or GAN-generated histology samples depicting benign or malignant breast cancer. Then all participants were probed for generalization to novel samples from both image sources. Categorization performance, as characterized by sensitivity and bias, showed no reliable differences between groups during training. During generalization, categorization performance with samples matching the image source seen during training was maintained. Most critically, categorization performance generalized across image sources with no loss: participants trained with GAN-generated samples were as sensitive and unbiased in categorizing real samples as those trained with real samples, and vice versa. Our results thus support a close correspondence between how humans and deep networks represent natural categories.

Keywords: generative adversarial networks; GANs; deep neural networks; breast cancer; categorization

Background

Generative adversarial networks (GANs) learn a generative model through the adversarial cooperation between two networks (Goodfellow et al., 2014). The generator network transforms inputs from an arbitrary distribution (e.g., a multivariate Gaussian) into samples from a target data distribution (e.g., a specific image domain). The discriminator network learns to distinguish real from generated samples, and its discrimination performance provides a teaching signal to the generator network. At the end of training, the generator network creates artificial samples that the discriminator network cannot distinguish from the real samples.

GANs have enormous potential as a research tool for cognitive science, offering novel approaches to answer complex methodological and theoretical questions (Goetschalckx, Andonian, & Wagemans, 2021). However, the correspondence between human and GAN representations is rarely directly assessed. Here, we used human category learning to evaluate the category representations developed by conditional GANs (Mirza & Osindero, 2014), a type of GAN that can be conditionalized to generate class-specific samples. We first trained online participants to categorize real or GAN-generated histology samples of breast cancer as benign or malignant, and we later tested them for generalization to novel images from both image sources. In doing so, we sought to answer: 1) whether humans can learn the category representations in conditional GANs, and 2) whether there is enough overlap between GAN-generated and real category representations to enable generalization across them.

Methods

Stimuli

We trained a style-based GAN (StyleGAN-v2; Karras, Laine, et al., 2020) to generate histology samples at a 100x magnification of human benign and malignant breast cancer, sourced from the BreakHis database (Spanhol, Oliveira, Petitjean, & Heutte, 2016), with adaptive discriminator augmentation (Karras, Aittala, et al., 2020). After training with 6M images, we achieved a Fréchet Inception Distance score of 7.76. A preliminary pilot study revealed that unconstrained training sets were too difficult for our human participants to learn within the allotted time. Therefore, we selected the 144 easiest images within each category (benign/malignant) and image source (real/GAN), based on the distance of each image's VGG16 features (Simonyan & Zisserman, 2015) from a hyperplane separating benign and malignant categories. See Figure 1 for some examples.

Participants

We recruited 64 Prolific workers. Participants were randomly assigned to one of two groups (Real/GAN). Subjects with a d' of less than 0.5 on the last block of training (see below) were excluded from the study, leaving 27 and 23 subjects in groups Real and GAN, respectively.

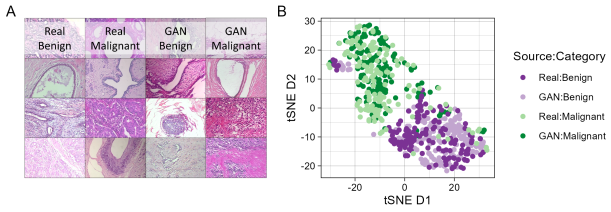

Figure 1: **(A)** Examples of real and GAN-generated samples. **(B)** tSNE embedding of the VGG16 features for the images used in the experiment.

Procedure

Training After instructions, participants received 4 blocks of 64 trials (32 trials per category). On each trial, a sample was presented in the centre of the screen with the prompt "Is this sample benign (B) or malignant (M)?". After pressing a response key, participants received visual feedback for 1 s ("Correct!" or "Error!") and after a 1 s blank screen, the next trial started. Groups Real and GAN were trained with BreakHis and GAN-generated samples, respectively.

Generalization test Both groups completed a final 64-trial block in the absence of feedback, containing 32 real and 32 GAN-generated samples (16 per category within each source).

Results

We characterised performance via sensitivity (d') and bias (β), and estimated Bayesian mixed-effects models in brms (Bürkner, 2018) for each index. We assessed the reliability of the estimated differences by quantifying the percentage of the posteriors' HDI contained within a region of practical equivalence (ROPE) spanning ±0.1 SDs (Kruschke, 2018).

Training Participants learned to categorize the samples (Figure 2A), increasing their sensitivity across training blocks (b = 0.40, 95% CI = [0.28, 0.53], 0% in ROPE), with no reliable group differences (b = -0.06, 95% CI = [-0.53, 0.42], 30% in ROPE). There was no reliable increase in bias nor group effects on it (b = 0.05, 95% CI = [-0.04, 0.13], 47% in ROPE, and b = -0.07, 95% CI = [-0.19, 0.05], 32% in ROPE, respectively).

Generalization test We quantified the generalization gap by calculating difference scores for sensitivity and bias measures from tests with real and GAN-generated samples (Figure 2B)[1]. These difference scores were positive if sensitivity/bias was higher for samples of the training image source (e.g., real samples for group Real), and negative if those measures were lower for samples of the opposite image source (e.g., GAN-generated samples for group Real). Our analysis revealed no reliable differences in sensitivity across image sources for either group (b = -0.06, 95% CI = [-0.33, 0.20], 37% in ROPE and b = -0.11, 95% CI = [-0.50, 0.28], 25% in ROPE, for group Real and GAN, respectively). There were no reliable differences in bias either (b = 0.41, 95% CI = [-0.08, 0.90], 11% in ROPE and b = -0.37, 95% CI = [-1.07, 0.30], 18% in ROPE, for group Real and GAN, respectively).

[1] Participants were as sensitive and unbiased in categorizing samples from their training image source (not shown).

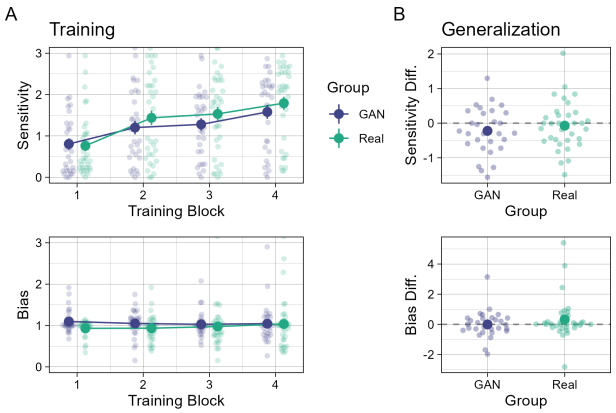

Figure 2: Sensitivity (d') and bias (β) during **(A)** training and **(B)** the generalization test.

Conclusions

Our GAN was remarkably proficient at capturing the image statistics of the histology samples present in the BreakHis dataset (Figure 1B). Human participants successfully learned categories from GAN-generated samples (Figure 2A) and, most critically, were able to transfer category representations acquired with either real or GAN-generated samples across image sources (Figure 2B). This generalization occurred with no appreciable loss, suggesting a strong correspondence between the category representations that humans and GANs learn when presented with the same stimuli.

Thus, the current findings support the use of category representations in GANs as a tool to study the corresponding representations in humans, opening exciting venues for future work. For instance, it would be possible to rearrange category membership of the visually complex histology samples to revisit findings established using categorization tasks built upon visually impoverished stimuli (Ashby & Valentin, 2018). In a more applied setting, the close correspondence in class-specific representations in human and GANs might provide the opportunity to devise training regimes that speed up the acquisition of, or deepen existing expertise in radiologists (or other medical experts), by generating tasks with high control over sample difficulty (e.g., Roads, Xu, Robinson, & Tanaka, 2018).

Acknowledgments

V.N. was funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (EP/Y026489/1). This research was undertaken using the supercomputing facilities at Cardiff University operated by Advanced Research Computing at Cardiff (ARCCA) on behalf of

the Cardiff Supercomputing Facility and the HPC Wales and Supercomputing Wales (SCW) projects.

References

Ashby, F. G., & Valentin, V. V. (2018). The categorization experiment: Experimental design and data analysis. In J. T. Wixted (Ed.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (pp. 1–41). Hoboken, NJ, USA: John Wiley & Sons, Inc. doi: 10.1002/9781119170174.epcn508

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. doi: 10.32614/RJ-2018-017

Goetschalckx, L., Andonian, A., & Wagemans, J. (2021). Generative adversarial networks unlock new methods for cognitive science. *Trends in Cognitive Sciences*, *25*(9), 788–801. doi: 10.1016/j.tics.2021.06.006

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2* (pp. 2672–2680). Cambridge, MA, USA: MIT Press.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., & Aila, T. (2020). Training generative adversarial networks with limited data. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (pp. 12104–12114). Red Hook, NY, USA: Curran Associates Inc.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020, June). Analyzing and improving the image quality of StyleGAN. In (pp. 8107–8116). IEEE Computer Society. doi: 10.1109/CVPR42600.2020.00813

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. doi: 10.1177/2515245918771304

Mirza, M., & Osindero, S. (2014, November). Conditional Generative Adversarial Nets. *arXiv:1411.1784 [cs, stat]*. (arXiv: 1411.1784)

Roads, B. D., Xu, B., Robinson, J. K., & Tanaka, J. W. (2018). The easy-to-hard training advantage with real-world medical images. *Cognitive Research: Principles and Implications*, *3*(1), 38. doi: 10.1186/s41235-018-0131-6

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International conference on learning representations.*

Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016, July). A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, *63*(7), 1455–1462. doi: 10.1109/TBME.2015.2496264