User Discrimination of Content Produced by Generative Adversarial Networks

Nicholas Caporusso[1], Kelei Zhang[1], Gordon Carlson[2], Daniel Jachetta[1], Devon Patchin[1], Spencer Romeiser[1], Noah Vaughn[1] and Angela Walters[1]

[1] Department of Informatics, Fort Hays State University,
600 Park Street, 67601 Hays, United States
{n_caporusso, k_zhang4, awalters}@fhsu.edu,
{ddjachetta, djpatchin, s_romeiser, nqvaughn}@mail.fhsu.edu
[2] Institute for New Media Studies, Fort Hays State University,
600 Park Street, 67601 Hays, United States
{gscarlson}@fhsu.edu

Abstract. Artificial Intelligence (AI) is increasingly being introduced in several domains for classification and clustering of different types of existing information (e.g., text, images, audio, and video). Recently, improvements to Machine Learning (ML) and new approaches to the design and use of Neural Networks (NNs) enabled the development of algorithms that generate new content that mimics the features of the training dataset. Specifically, Generative Adversarial Networks (GANs) are particularly effective in producing content with unprecedented levels of fidelity. As a result, they can generate realistic images of people, vehicles, and nature.

In this paper, we discuss the results of a study that investigated user perception of pictures generated using GANs, with specific regard to portraits featuring faces. Specifically, our experiment involved 551 participants who were asked to classify over 7000 real images and pictures generated by ML algorithms. Our findings show that users show low accuracy in discriminating images and, thus, demonstrate the effectiveness of GANs in producing content that can be perceived as realistic.

Keywords: Artificial Intelligence · Machine Learning · Neural Networks

1 Introduction

In the recent years, Artificial Intelligence (AI) has become increasingly powerful thanks to faster hardware, to the development of new algorithms, and to the availability of open source frameworks and libraries that render AI accessible to larger audiences [1]. As a result, nowadays Machine Learning (ML) is being adopted in different domains (e.g., healthcare [2]) and tasks, such as, image [3], movement [4] and speech recognition, object detection, and autonomous vehicle navigation. In addition to solving traditional problems in the domain of AI, such as, classification and prediction, novel ML techniques aim at overcoming one of the main limitations of AI, that is, the possibility of creating new content.

Specifically, Generative Adversarial Networks (GANs) [5] are able to produce original material (e.g., images, speech, and text) with incredible fidelity compared to the training dataset. To this end, their architecture consists of two neural networks competing against one another: a generator NN creates new content based on features and parameters learned from a dataset, whereas a discriminator NN evaluates the output to ensure that its level of realism is consistent with the training set. As a result, GANs can generate high-quality, realistic images featuring non-existing individuals, original artwork that mimics the style of an artist, and text that resembles the rhetoric of a politician. Indeed, empowering machines to develop some degree of creativity and invent new content has incredible potential for applications in several industries. Simultaneously, potential misuse of novel algorithms poses challenges in terms of threats caused by highly realistic and credible profiles and fake news or content. Consequently, there is an emerging need of studying dynamics of human-machine interaction with AI-generated content to study both beneficial and malicious applications of novel Machine Learning systems.

As algorithms will produce more sophisticated and realistic replicas of user-generated content, studying individuals' behavior in the interaction with AI-generated material is crucial for improving the design of ML systems as well as for understanding mechanisms for ensuring users' security and educating them. Previous work focused on analyzing factors, such as, eye blinking, to evaluate whether users are able to distinguish fake faces in real pictures and videos [6].

2 Experimental Study

In this paper, we present a study in which we focused on users' perception of real and AI-generated content. Specifically, we investigated factors involved in the ability to distinguish between authentic and fake images with the objective of evaluating whether users are able to perceive the difference between pictures and discriminate the features that characterize images featuring real humans and AI-generated faces. To this end, we utilized the Generative Adversarial Networks dataset produced using NVIDIA's style-based generator architecture.

Our study consisted of an on-line survey in which participants were presented with random picture from a sample of 20000 images (5000 real and 15000 fake) extracted from the NVIDIA dataset [7]. To this end, we developed a web-based software (publicly available at https://newmedia.fhsu.edu/projects/real-or-ai/) that prompted users with a sequence of 10 randomly-selected non-repeat images from the sample collection. The interface was set to maximize the image width to fit the screen size of users' devices. For each picture, participants were given 5 seconds to decide between two options, that is, real human or AI-generated, and to click on the appropriate button on the interface of the experimental software.

In addition to recording users' responses, the data collection tool timed each phase of the experiment and acquired the device type (i.e., desktop or mobile) and the display size (i.e., extra-small, small, medium, large, extra-large). Finally, at the end of the session, the data collection tool asked participants for demographic information (i.e., age, ethnicity, gender, and familiarity with computers) and presented a debriefing screen with the results of the experiment, to gamify the experiment and incentivize sharing it.

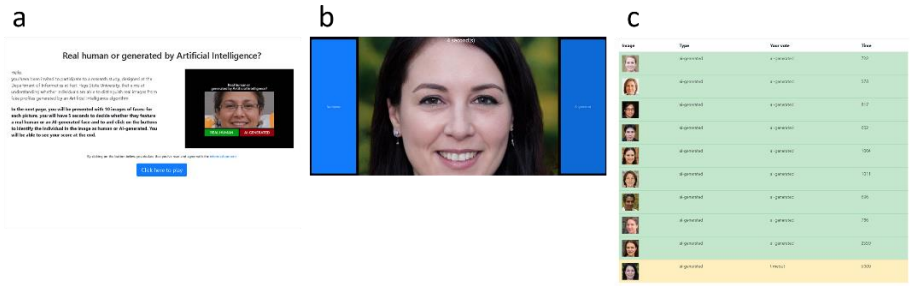

Fig. 1. The experimental software: (a) introduction page showing the instructions and informed consent, (b) trial interface in which users were presented with the image and with buttons (on the left and right sides of the display), and (c) debriefing page. Buttons had the same color to avoid bias.

3 Results and Discussion

551unique subjects participated to the study: among the ones who agreed to share their gender, 216 (39.20%) were females and 220 (39.93%) were males. Their age distribution was the following: 109 (19.78%) were 18-24, 222 (40.29%) aged 25-39, 88 (15.97%) were in the 40-54 bracket, and 34 (6.17%) were in the 55-75 range. Among the ones who agreed to share their nationality, 139 (25.22%) from Europe, 138 (25.04%) were from the US, 89 (16.15%) from China, 43 (7.80%) from other Asian countries, 19 (3.44%) identified as White Caucasian, and 83 (15.06%) were part of other groups. A total of 7137 trials were realized (an average of 12.95 trial per participant), each featuring a different picture (5668 AI-generated and 1469 actual profiles, accounting for 79.41% and 20.58% of the dataset, respectively). There were 742 cases in which the trial timed out due to several factors, including distraction, network performance causing picture rendering delay, and users not being able to identify an image before the 5-second timer expired. Timed-out trials involved 584 AI-generated images and 158 original images, accounting for 10.30% and 10.75% of the datasets, respectively. No statistically significant difference between classes was found. As a result, their values were not considered in our analysis. Consequently, our dataset involved a total of 6395 trials, involving images featuring 5084 AI-generated and 1311 real faces (79.49% and 20.50% respectively).

Figure 2 represents the experimental results as a confusion matrix: as users correctly identified AI-generated images 55.15% of the times, they were basically guessing when discriminating fake profiles, whereas they were slightly more successful (63.92%) in identifying real individuals. Furthermore, we did not find any statistically-significant correlation between accuracy and gender. Also, ethnicity does not seem to be a factor. However, we did not compare the ethnicity of the subject and that of the individual featured in the image, whereas published work on the own-gender and own-race biases [8] [9] suggests that this aspect might need further investigation.

n=6395		Actual		
		AI-generated	*Real*	
User evaluation	*AI-generated*	2804	473	3277
	Real	2280	838	3118
		5084	1311	

Fig. 2. Confusion matrix showing the actual content and participants' evaluation. Users had a total accuracy of 56.95% as they correctly identified images in 3642 trials: 2084 were AI-generated and 838 actual profiles. Users had a higher precision (85.56%) and lower recall (55.15%) in classifying pictures generated with GANs, while they had lower precision (26.87%) and higher recall (63.92%) in identifying images featuring actual humans.

Furthermore, when we analyzed the 3232 trials (50.53%) that contained information about display size, we found no statistical difference across device screens: surprisingly, the 1395 trials (43.16%) that involved desktop displays had an accuracy comparable to the 1837 trials (56.83%) that were executed on mobile devices, demonstrating the difficulty of recognizing distinctive features in images at larger resolutions.

Conversely, our findings show that age is a factor, as detailed in Figure 3, which might be negatively correlated with accuracy. Although the number of subjects in each age group influences the results, differences in correctly identifying images exist within the groups of users who are younger than 45, which confirms the presence of a trend, though we will address this concern by improving randomization in sampling and data collection.

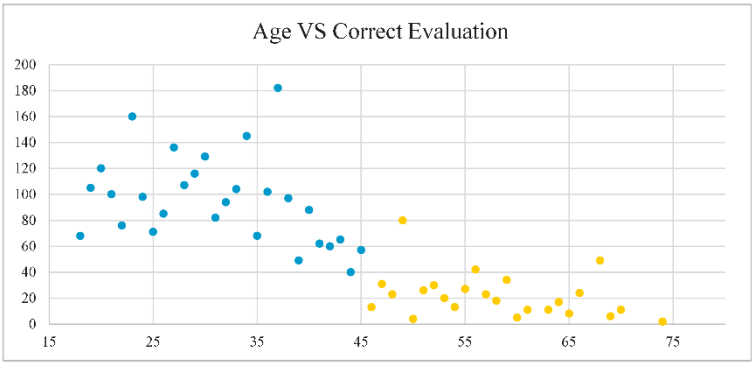

Fig. 3. Age VS correct User Evaluation. Age distribution among subjects partially explains the results, as individuals younger than 40 (331, 60,07%) were more than twice the number of people older than 40 (122 – 22,14%). Simultaneously, the 109

participants between 12 and 24 had almost the same number of correct responses of the 222 subjects aged 25-39.

Also, we found a positive correlation between hours spent per day on PC and accuracy in classifying pictures: longer times spent on the PC are associated with better ability in recognizing real and fake profiles (the correlation coefficient is 0.98). Although we enabled subjects to iterate the experiment multiple times, we were not able to collect enough repeated sessions and, thus, we do not have enough data for analyzing the presence of training effect and for evaluating users' learning curve. Nonetheless, as data collection is ongoing, we aim at addressing this issue in a follow-up paper. The data collected in our experiment is shared in an open repository: a link for downloading the dataset in Comma Separated Values (CSV) format is available on the introduction page of the experimental software.

4 Conclusion and Future Work

In the last years, ML algorithms experienced dramatic improvement, especially in the context of image classification and recognition. Furthermore, Generative Adversarial Networks introduce novel ways to leverage AI to produce content. The ultimate objective of our work is to improve content generation Machine Learning systems by taking into account user experience. To this end, in this paper, we detailed the results of the preliminary stage of a focusing on content produced using GANs, with the purpose of our study was to evaluate whether algorithms are mature enough in terms of realism and fidelity. Specifically, the objective of our work was to evaluate whether users can recognize images featuring fake faces generated by AI algorithms and distinguish them from actual humans. To this end, we designed a study that investigated user perception of images created with GANs. Our findings show that participants showed low accuracy in discriminating real from fake profiles. From our results, we can conclude that the level of fidelity of images from the NVIDIA dataset was high enough to succeed in confounding users and triggering them into thinking that they were observing real profiles. Specifically, content produced by AI generated more ambiguity and users were more likely to identify it as authentic, whereas they were more accurate with pictures featuring real humans. Although we investigated images featuring faces, a similar approach can be applied to different types of content and formats.

In our future work, we will realize multiple iterations over the dataset: by collecting several responses for each image, we will be able to rank pictures by realism based on users' perception. Then, we will administer pictures from the ranked dataset to subjects, and we utilized an eye-tracking system to acquire and analyze fixations, in order to identify the key features utilized by individuals to distinguish between real and fake image. Subsequently, we will use the data to identify features that characterize fake images and to help improve both the generator and the discriminator components of the Generative Adversarial Network.

References

1. Caporusso, N., Helms, T. and Zhang, P. 2019, July. A Meta-Language Approach for Machine Learning, to appear in 2nd International Conference on Human Factors in Artificial Intelligence and Social Computing. In International Conference on Applied Human Factors and Ergonomics. Springer. To be published.
2. De Pace, A., Galeandro, P., Trotta, G.F., Caporusso, N., Marino, F., Alberotanza, V. and Scardapane, A., 2017, April. Synthesis of a Neural Network Classifier for Hepatocellular Carcinoma Grading Based on Triphasic CT Images. In Recent Trends in Image Processing and Pattern Recognition: First International Conference, RTIP2R 2016, Bidar, India, December 16–17, 2016, Revised Selected Papers (Vol. 709, p. 356). Springer. doi:10.1007/978-3-319-60483-1_13.
3. Bevilacqua, V., Uva, A.E., Fiorentino, M., Trotta, G.F., Dimatteo, M., Nasca, E., Nocera, A.N., Cascarano, G.D., Brunetti, A., Caporusso, N. and Pellicciari, R., 2016, December. A Comprehensive Method for Assessing the Blepharospasm Cases Severity. In International Conference on Recent Trends in Image Processing and Pattern Recognition (pp. 369-381). Springer, Singapore. doi: 10.1007/978-981-10-4859-3_33
4. Bevilacqua, V., Trotta, G.F., Loconsole, C., Brunetti, A., Caporusso, N., Bellantuono, G.M., De Feudis, I., Patruno, D., De Marco, D., Venneri, A. and Di Vietro, M.G., 2017, July. A RGB-D Sensor Based Tool for Assessment and Rating of Movement Disorders. In International Conference on Applied Human Factors and Ergonomics (pp. 110-118). Springer, Cham. doi: 10.1007/978-3-319-60483-1_12
5. Bevilacqua, V., Trotta, G.F., Brunetti, A., Caporusso, N., Loconsole, C., Cascarano, G.D., Catino, F., Cozzoli, P., Delfine, G., Mastronardi, A. and Di Candia, A., 2017, July. A Comprehensive Approach for Physical Rehabilitation Assessment in Multiple Sclerosis Patients Based on Gait Analysis. In International Conference on Applied Human Factors and Ergonomics (pp. 119-128). Springer, Cham. doi: 10.1007/978-3-319-60483-1_13.
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).
7. Li, Y., Chang, M.C., Farid, H. and Lyu, S., 2018. In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. arXiv preprint arXiv:1806.02877.
8. Karras, T., Laine, S. and Aila, T., 2018. A style-based generator architecture for generative adversarial networks. arXiv preprint arXiv:1812.04948.
9. Wright, D.B. and Sladden, B., 2003. An own gender bias and the importance of hair in face recognition. Acta psychologica, 114(1), pp.101-114.
10. Johnson, K.J. and Fredrickson, B.L., 2005. “We all look the same to me” Positive emotions eliminate the own-race bias in face recognition. Psychological science, 16(11), pp.875-881.