A NEURO-AI INTERFACE FOR EVALUATING GENERATIVE ADVERSARIAL NETWORKS

Zhengwei Wang[1] *, **Qi She**[2], **Alan F. Smeaton**[3], **Tomás E. Ward**[3] **& Graham Healy**[3]
[1] School of Computer Science and Statistics, Trinity College Dublin, Dublin 1, Ireland
[2] Intel Lab, Beijing, China
[3] Insight Centre for Data Analytics, Dublin City University, Dublin 9, Ireland
zhengwei.wang@tcd.ie, qi.she@intel.com
{alan.smeaton,tomas.ward,graham.healy}@dcu.ie

ABSTRACT

Generative adversarial networks (GANs) are increasingly attracting attention in the computer vision, natural language processing, speech synthesis and similar domains. However, evaluating the performance of GANs is still an open and challenging problem. Existing evaluation metrics primarily measure the dissimilarity between real and generated images using automated statistical methods. They often require large sample sizes for evaluation and do not directly reflect human perception of image quality. In this work, we introduce an evaluation metric called **Neuroscore**, for evaluating the performance of GANs, that more directly reflects psychoperceptual image quality through the utilization of brain signals. Our results show that Neuroscore has superior performance to the current evaluation metrics in that: (1) It is more consistent with human judgment; (2) The evaluation process needs much smaller numbers of samples; and (3) It is able to rank the quality of images on a per GAN basis. A convolutional neural network (CNN) based **neuro-AI interface** is proposed to predict Neuroscore from GAN-generated images directly without the need for neural responses. Importantly, we show that including neural responses during the training phase of the network can significantly improve the prediction capability of the proposed model. Codes and data can be referred at this link: *https://github.com/villawang/Neuro-AI-Interface*.

1 INTRODUCTION

There is a growing interest in studying generative adversarial networks (GANs) in the deep learning community (Goodfellow et al., 2014). Specifically, GANs have been widely applied to various domains such as computer vision (Karras et al., 2018), natural language processing (Fedus et al., 2018), speech synthesis (Donahue et al., 2018) and time series generation (Brophy et al., 2019). Compared with other deep generative models (e.g. variational autoencoders (VAEs)), GANs are favored for effectively handling sharp estimated density functions, efficiently generating desired samples and eliminating deterministic bias (Wang et al., 2019c). Due to these properties GANs have successfully contributed to plausible image generation (Karras et al., 2018), image to image translation (Zhu et al., 2017), image super-resolution (Ledig et al., 2017), image completion (Yu et al., 2018) etc.

However, three main challenges still exist currently in the research of GANs: (1) Mode collapse - the model cannot learn the distribution of the full dataset well, which leads to poor generalization ability; (2) Difficult to train - it is non-trivial for discriminator and generator to achieve Nash equilibrium during the training; (3) Hard to evaluate - the evaluation of GANs can be considered as an effort to measure the dissimilarity between real distribution p_r and generated distribution p_g. Unfortunately, the accurate estimation of p_r is intractable. Thus, it is challenging to have a good estimation of the correspondence between p_r and p_g. Aspects (1) and (2) are more concerned with computational aspects where much research has been carried out to mitigate these issues (Li et al., 2015; Salimans et al., 2016; Arjovsky et al., 2017). Aspect (3) is similarly fundamental, however, limited literature is available and most of the current metrics only focus on measuring the dissimilarity between training and generated images. A more meaningful GANs evaluation metric that is consistent with human perceptions is paramount in helping researchers to further refine and design better GANs.

*Work done in the Insight Centre for Data Analytics, Dublin City University.

Although some evaluation metrics, e.g., Inception Score (IS), Kernel Maximum Mean Discrepancy (MMD) and Fréchet Inception Distance (FID), have already been proposed (Salimans et al., 2016; Heusel et al., 2017; Borji, 2018), their limitations are obvious: (1) These metrics do not agree with human perceptual judgments and human rankings of GAN models. A small artifact on images can have a large effect on the decision made by a machine learning system (Koh & Liang, 2017), whilst the intrinsic image content does not change. In this aspect, we consider human perception to be more robust to adversarial images samples when compared to a machine learning system; (2) These metrics require large sample sizes for evaluation (Xu et al., 2018; Salimans et al., 2016). Large-scale samples for evaluation sometimes are not realistic in real-world applications since it is time-consuming; and (3) They are not able to rank individual GAN-generated images by their quality i.e., the metrics are generated on a collection of images rather than on a single image basis.

Yamins et al. (2014) demonstrates that CNN matched with neural data recorded from inferior temporal cortex (Chelazzi et al., 1993) has high performance in object recognition tasks. Given the evidence above that a CNN is able to predict the neural response in the brain and can reflect the spatio-temporal neural dynamics in the human brain visual processing area (Cichy et al., 2016; Tu et al., 2018; Kuzovkin et al., 2018), we describe a neuro-AI interface system, where human being's neural response is used as supervised information to help the AI system (CNNs used in this work) solve challenging problems in the real world. As a starting point for exploiting the idea of neuro-AI interface, we focus on utilizing it to solve one of the fundamental problems in GANs: designing a proper evaluation metric.

In this paper, we firstly introduce a brain-produced score (Neuroscore), generated from human being's electroencephalography (EEG) signals, in terms of the quality evaluation on GANs. Secondly, we demonstrate and validate a neural-AI interface (as seen in Fig. 1), which uses neural responses as supervised information to train a CNN. The trained CNN model is able to predict Neuroscore (we call the predicted Neuroscore as synthetic-Neuroscore) for images without corresponding neural responses. We test this framework via three models: Shallow convolutional neural network, Mobilenet V2 (Sandler et al., 2018) and Inception V3 (Szegedy et al., 2016). The scope of the Neuro-AI interface should not be limited in DNN and EEG signals. We think spike trains or fMRI should also be potential source signals that can be used for training AI systems; furthermore, via utilizing more temporal properties (She et al., 2018; She & Wu, 2019; Feng et al., 2019; She et al., 2019), we think the features extracted from these time series signals should be more robust and predictive via learning the temporal structure.

Neuroscore (Wang et al., 2019a) is calculated via measurement of the P300 (by averaging the single-trial P300 amplitude), an event-related potential (ERP) (Polich, 2007) present in EEG, via a rapid serial visual presentation (RSVP) paradigm (Wang et al., 2018; Healy et al., 2020; Wang, 2019; Wang et al., 2016; Healy et al., 2017). The unique benefit of Neuroscore is that it more directly reflects the perceptual judgment of images, which is intuitively more reliable compared to the conventional metrics (Borji, 2018). Details of Neuroscore can be refered in (Wang et al., 2019a).

2 METHODOLOGY

2.1 NEURO-AI INTERFACE

Figure 2 demonstrates the schematic of neuro-AI interface used in this work. Flow 1 shows that the image processed by human being's brain and produces single trial P300 source signal for each input image. Flow 2 in Fig. 2 demonstrates a CNN with including EEG signals during training stage. The convolutional and pooling layers process the image similarly as retina done (McIntosh et al., 2016). Fully connected layers (FC) 1-3 aim to emulate the brain's functionality that produces EEG signal. Yellow dense layer in the architecture aims to predict the single trial P300 source signal in 400-600 ms response from each image input. In order to help model make a more accurate prediction for the single trial P300 amplitude for the output, the single trial P300 source signal in 400-600 ms is fed to the yellow dense layer to learn parameters for the previous layers in the training step. The model was then trained to predict the single trial P300 source amplitude (red point shown in signal trail P300 source signal of Fig. 2).

2.2 TRAINING DETAILS

Mobilenet V2, Inception V3 and Shallow network were explored in this work, where in flow 2 we use these three network bones: such as Conv1-pooling layers. For Mobilenet V2 and Inception V3.

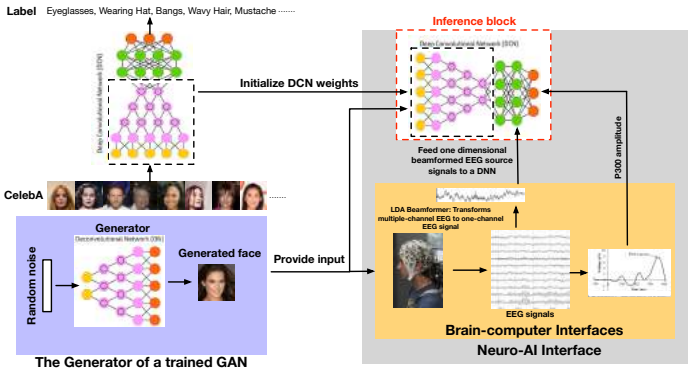

Figure 1: Schematic of neuro-AI interface. Image stimuli generated by GANs are simultaneously presented to a CNN and participants. The inference model is initialized by pretrained weights which has been trained by large scale dataset e.g., CelebA. Participants' P300 amplitude is fed to the network as ground truth and EEG responses are extracted and fed to the CNN as supervisory information for assisting the CNN predict P300 amplitude.

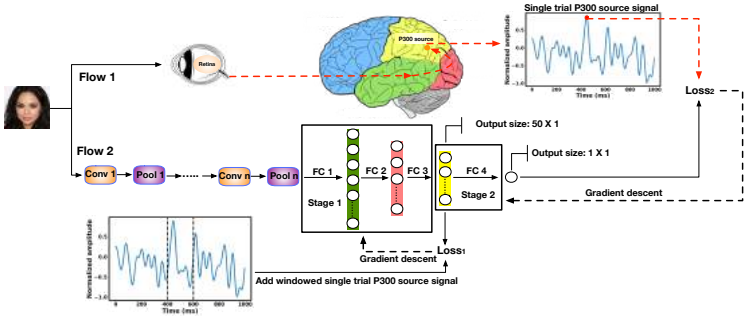

Figure 2: A neuro-AI interface and training details with adding EEG information. Our training strategy includes two stages: (1) Learning from image to P300 source signal; and (2) Learning from P300 source signal to P300 amplitude. loss_1 is the L_2 distance between the yellow layer and the single trial P300 source signal in the 400 - 600 ms corresponding to the single input image. loss_2 is the mean square error between model prediction and the single trial P300 amplitude. loss_1 and loss_2 will be introduced in section 2.2.

We used pretrained parameters from up to the FC 1 shown in Fig. 2. We trained parameters from FC 1 to FC 4 for Mobilenet V2 and Inception V3. $\boldsymbol{\theta}_1$ is used to denote the parameters from FC 1 to FC 3 and $\boldsymbol{\theta}_2$ indicates the parameters in FC 4. For the Shallow model, we trained all parameters from scratch.

We defined two stage loss function (loss_1 for single trial P300 source signal in the 400 - 600 ms time window and loss_2 for single trial P300 amplitude) as

$$\begin{aligned} \text{loss}_1(\boldsymbol{\theta}_1) &= \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{S}_i^{true} - \mathbf{S}_i^{pred}(\boldsymbol{\theta}_1)\|_2^2, \\ \text{loss}_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= \frac{1}{N}\sum_{i=1}^{N} (\text{y}_i^{true} - \text{y}_i^{pred}(\boldsymbol{\theta}_1, \boldsymbol{\theta_2}))^2, \end{aligned} \tag{1}$$

where $\mathbf{S}_i^{true} \in \mathbb{R}^{1\times T}$ is the single trial P300 signal in the 400 - 600 ms time window to the presented image, and y_i refers to the single trial P300 amplitude to each image.

The training of the models without using EEG is straightforward, models were trained directly to minimize $\text{loss}_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ by feeding images and the corresponding single trial P300 amplitude. Training with EEG information is visualized in the "Flow 2" of Fig. 2 with two stages. Stage 1

learns parameters $\boldsymbol{\theta}_1$ to predict P300 source signal while stage 2 learns parameters $\boldsymbol{\theta}_2$ to predict single trial P300 amplitude with $\boldsymbol{\theta}_1$ fixed.

3 RESULTS

Table 1 shows the error for each model with EEG signal, with randomized EEG signal **within each type of GAN** and without EEG. All models with EEG perform better than models without EEG, with much smaller errors and variances. Statistic tests between model with EEG and without EEG are also carried out to verify the significance of including EEG information during the training phase. One-way ANOVA tests (P-value) for each model with EEG and without EEG are stated as: $P_{Shallow} = 0.003$, $P_{Mobilenet} = 0.012$ and $P_{Inception} = 5.980e-05$. Results here demonstrate that including EEG during the training stage helps all three CNNs improve the performance on predicting the Neuroscore. The performance of models with EEG is ranked as follows: Inception-EEG, Mobilenet-EEG, and Shallow-EEG, which indicates that deeper neural networks may achieve better performance in this task.

Model		Error mean(std)
Shallow net	Shallow-EEG	**0.209 (±0.102)**
	Shallow-EEG$_{\mathrm{random}}$	0.348 (±0.114)
	Shallow	0.360 (±0.183)
Mobilenet	Mobilenet-EEG	**0.198 (±0.087)**
	Mobilenet-EEG$_{\mathrm{random}}$	0.404 (±0.162)
	Mobilenet	0.366 (±0.261)
Inception	Inception-EEG	**0.173 (±0.069)**
	Inception-EEG$_{\mathrm{random}}$	0.392 (±0.057)
	Inception	0.344 (±0.149)

Table 1: Errors of 9 models for cross participants ("-EEG" indicates models are trained with paired EEG, "-EEG$_{\mathrm{random}}$" refers to EEG trials which are randomized in the loss_1 **within each type of GAN**). Results are averaged by shuffling training/testing sets for 20 times. Error is defined as: $\sum_i^m |\mathrm{Neuroscore}_{pred}^{(i)} - \mathrm{Neuroscore}_{true}^{(i)}|$, where $\mathrm{m} = 3$ is the number of GAN category used (DCGAN, BEGAN, PROGAN) (Radford et al., 2015; Berthelot et al., 2017; Karras et al., 2017).

Table 2 shows the comparison between synthetic-Neuroscore and three traditional scores. To be consistent with all the scores (smaller score indicates better GAN), we used 1/IS and 1/synthetic-Neuroscore for comparisons in Table 2. It can be seen that people rank the GAN performance as PROGAN $>$ BEGAN $>$ DCGAN. All three synthetic-Neuroscores produced by the three models with EEG are consistent with human judgment while the other three conventional scores are not (they all indicate that DCGAN outperforms BEGAN).

Metrics		DCGAN	BEGAN	PROGAN
1/IS		0.44	0.57	0.42
MMD		0.22	0.29	0.12
FID		63.29	83.38	34.10
Proposed Methods	1/Shallow-EEG	**1.60**	**1.39**	**1.14**
	1/Mobilenet-EEG	**1.71**	**1.29**	**1.20**
	1/Inception-EEG	**1.51**	**1.34**	**1.24**
Human (BE accuracy)		**0.995**	**0.824**	**0.705**

Table 2: Three conventional scores: IS, MMD, FID, and synthetic-Neuroscore produced by three models with EEG for each GAN category. A lower score indicates better performance of the GAN. Bold text indicates the consistency with human judgment (BE) accuracy.

4 CONCLUSION

In this paper, we introduce a neuro-AI interface that interacts CNNs with neural signals. We demonstrate the use of neuro-AI interface by introducing a challenge in the area of GANs i.e., evaluate the quality of images produced by GANs. Three deep network architectures are explored and the results demonstrate that including neural responses during the training phase of the neuro-AI interface improves its accuracy even when neural measurements are absent when evaluating on the test set. More details of Neuroscore can be found in the recent work (Wang et al., 2019b).

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint:1701.07875*, 2017.

David Berthelot, Thomas Schumm, and Luke Metz. BEGAN: boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

Ali Borji. Pros and cons of GAN evaluation measures. *arXiv preprint:1802.03446*, 2018.

Eoin Brophy, Zhengwei Wang, and Tomas E Ward. Quick and easy time series generation with established image-based GANs. *arXiv preprint arXiv:1902.05624*, 2019.

Leonardo Chelazzi, Earl K Miller, John Duncan, and Robert Desimone. A neural basis for visual search in inferior temporal cortex. *Nature*, 363(6427):345, 1993.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6:27755, 2016.

Chris Donahue, Julian McAuley, and Miller Puckette. Synthesizing audio with generative adversarial networks. *arXiv preprint:1802.04208*, 2018.

William Fedus, Ian Goodfellow, and Andrew M Dai. MaskGAN: Better text generation via filling in the _. *arXiv preprint:1801.07736*, 2018.

Fan Feng, Rosa HM Chan, Xuesong Shi, Yimin Zhang, and Qi She. Challenges in task incremental learning for assistive robotics. *IEEE Access*, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Graham Healy, Zhengwei Wang, Cathal Gurrin, Tomas Ward, and Alan F Smeaton. An EEG image-search dataset: a first-of-its-kind in IR/IIR. NAILS: neurally augmented image labelling strategies. 2017.

Graham Healy, Zhengwei Wang, Tomas Ward, Alan Smeaton, and Cathal Gurrin. Experiences and insights from the collection of a novel multimedia EEG dataset. In *International Conference on Multimedia Modeling*, pp. 475–486. Springer, 2020.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *arXiv preprint:1812.04948*, 2018.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pp. 1885–1894, 2017.

Ilya Kuzovkin, Raul Vicente, Mathilde Petton, Jean-Philippe Lachaux, Monica Baciu, Philippe Kahane, Sylvain Rheims, Juan R Vidal, and Jaan Aru. Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. *Communications biology*, 1(1):107, 2018.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 105–114. IEEE, 2017.

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pp. 1718–1727, 2015.

Lane McIntosh, Niru Maheswaranathan, Aran Nayebi, Surya Ganguli, and Stephen Baccus. Deep learning models of the retinal response to natural scenes. In *Advances in Neural Information Processing Systems*, pp. 1369–1377, 2016.

John Polich. Updating P300: an integrative theory of P3a and P3b. *Clinical Neurophysiology*, 118 (10):2128–2148, 2007.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520. IEEE, 2018.

Qi She and Anqi Wu. Neural dynamics discovery via gaussian process recurrent neural networks. *arXiv preprint arXiv:1907.00650*, 2019.

Qi She, Yuan Gao, Kai Xu, and Rosa HM Chan. Reduced-rank linear dynamical systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Qi She, Fan Feng, Xinyue Hao, Qihan Yang, Chuanlin Lan, Vincenzo Lomonaco, Xuesong Shi, Zhengwei Wang, Yao Guo, Yimin Zhang, et al. Openloris-object: A dataset and benchmark towards lifelong object recognition. *arXiv preprint arXiv:1911.06487*, 2019.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Tao Tu, Jonathan Koss, and Paul Sajda. Relating deep neural network representations to EEG-fMRI spatiotemporal dynamics in a perceptual decision-making task. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1985–1991, 2018.

Zhengwei Wang. *Cortically coupled image computing*. PhD thesis, Dublin City University, 2019.

Zhengwei Wang, Graham Healy, Alan F Smeaton, and Tomas E Ward. An investigation of triggering approaches for the rapid serial visual presentation paradigm in brain computer interfacing. In *2016 27th Irish Signals and Systems Conference*, pp. 1–6. IEEE, 2016.

Zhengwei Wang, Graham Healy, Alan F Smeaton, and Tomas E Ward. A review of feature extraction and classification algorithms for image RSVP based BCI. *Signal Processing and Machine Learning for Brain-machine Interfaces*, pp. 243–270, 2018.

Zhengwei Wang, Graham Healy, Alan F Smeaton, and Tomas E Ward. Use of neural signals to evaluate the quality of generative adversarial network performance in facial image generation. *Cognitive Computation*, pp. 1–12, 2019a.

Zhengwei Wang, Qi She, Alan F Smeaton, Tomas E Ward, and Graham Healy. Synthetic-Neuroscore: Using a neuro-AI interface for evaluating generative adversarial networks. *arXiv preprint arXiv:1905.04243*, 2019b.

Zhengwei Wang, Qi She, and Tomas E Ward. Generative adversarial networks in computer vision: A survey and taxonomy. *arXiv preprint arXiv:1906.01529*, 2019c.

Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Q. Weinberger. An empirical study on evaluation metrics of generative adversarial networks. *arXiv preprint:1806.07755*, 2018.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. *arXiv preprint:1801.07892*, 2018.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint:1703.10593v6*, 2017.