Neuroscore: A Brain-inspired Evaluation Metric for Generative Adversarial Networks

Zhengwei Wang[1*], **Qi She**[2], **Alan F. Smeaton**[1], **Tomás E. Ward**[1] and **Graham Healy**[1]

[1]Insight Centre for Data Analytics, Dublin City University, Ireland

[2]Intel Research Lab, Beijing, China

zhengwei.wang22@mail.dcu.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. Arguably the most striking results have been in the area of image synthesis. 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 the human perception of the image quality. In this work, we introduce an evaluation metric we call **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 performances 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 based brain-inspired framework is also proposed to predict Neuroscore from GAN-generated images. Importantly, we show that including neural responses during the training phase of the network can significantly improve the prediction capability of the proposed model.

1 Introduction

There is a growing interest in studying generative adversarial networks (GANs) [Goodfellow *et al.*, 2014] in the deep learning community. 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 etc. 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. 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] and 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.

Although some evaluation metrics, e.g. Inception Score, Kernel Maximum Mean Discrepancy, The Fréchet Inception Distance, 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 perturbation on images can have a large effect on the decision made by a machine learning system [Koh and Liang, 2017], whilst the intrinsic image content does not change. In this respect, 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. The

*Contact Author

within GAN variances are crucial because it can provide the insight on the variability of that GAN. In this work, we introduce a novel metric named Neuroscore to evaluate the performance of GANs, which is derived from a neuropsychological response recorded via non-invasive electroencephalography (EEG). Furthermore, we demonstrate and validate a brain-inspired framework that calculates this 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].

In detail, Neuroscore is calculated via measurement of the P300, an event-related potential (ERP) present in EEG, via a rapid serial visual presentation (RSVP) paradigm. P300 and RSVP paradigm are mature techniques in the brain-computer interface (BCI) community and have been applied in a wider variety of tasks such as image search [Gerson *et al.*, 2006], information retrieval [Mohedano *et al.*, 2015], and etc. The unique benefit of Neuroscore is that it more directly reflects the human perceptual judgment of images, which is intuitively more reliable compared to the conventional metrics in the literature [Borji, 2018]. In summary, our contributions are three-fold:

- We combine human perception research with GANs research. It is the first exploration of this research line and demonstrates that it deserves being further exploited to refine and design better GANs.
- We outline a novel experimental flow and propose a brain-inspired evaluation metric called Neuroscore, which has been demonstrated to be more effective for evaluating GANs than STOAs.
- We propose a brain-inspired framework and training strategy to generalize the use of Neuroscore, which can be directly used for GANs evaluations without recording EEG. This enables our Neuroscore to be more widely applied to real-world scenarios.

2 Related Work

Three well-known metrics are compared with Neuroscore.

Inception Score (IS) is the most widely used metric in the literature [Salimans *et al.*, 2016; Xu *et al.*, 2018; Borji, 2018]. It uses a pre-trained Inception network [Szegedy *et al.*, 2016] as an image classification model $\mathcal{M}$ to compute

$$\mathrm{IS} = \exp\left(\mathbb{E}_{\mathbf{x}\sim p_g}\left[\mathrm{KL}\left(p_{\mathcal{M}}\left(\mathrm{y}|\mathbf{x}\right)||p_{\mathcal{M}}(\mathrm{y})\right)\right]\right).$$

Where $p_{\mathcal{M}}(\mathrm{y}|\mathbf{x})$ is the label distribution of $\mathbf{x}$ that is predicted by the model $\mathcal{M}$ and $p_{\mathcal{M}}(\mathrm{y})$ is the marginal probability of $p_{\mathcal{M}}(\mathrm{y}|\mathbf{x})$ over the probability p_g. A larger inception score will have $p_{\mathcal{M}}(\mathrm{y}|\mathbf{x})$ close to a point mass and $p_{\mathcal{M}}(\mathrm{y})$ close to uniform, which indicates that the Inception network is very confident that the image belongs to a particular ImageNet category where all categories are equally represented. This suggests the generative model has both high quality and diversity.

Kernel Maximum Mean Discrepancy (MMD) is a method for comparing two distributions, in which the test statistic is the largest difference in expectations over functions in the unit ball of a reproducing kernel Hilbert space [Gretton *et al.*, 2012]. MMD is computed as

$$\mathrm{MMD}^2(p_r, p_g) = \mathbb{E}_{\mathbf{x}_r,\mathbf{x}_r^\top,\sim p_r,\mathbf{x}_g,\mathbf{x}_g^\top\sim p_g} \left[k(\mathbf{x}_r,\mathbf{x}_r^\top) - 2k(\mathbf{x}_r,\mathbf{x}_g) + k(\mathbf{x}_g,\mathbf{x}_g^\top)\right].$$

It measures the dissimilarity between p_r and p_g for some fixed kernel function k, such as a Gaussian kernel [Li *et al.*, 2015]. A lower MMD indicates that p_g is closer to p_r, showing the GAN has better performance.

The Fréchet Inception Distance (FID) uses a feature space extracted from a set of generated image samples by a specific layer of the Inception network [Heusel *et al.*, 2017]. The feature space is modelled via a multivariate Gaussian by the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. FID is computed as

$$\mathrm{FID}(p_r, p_g) = ||\boldsymbol{\mu}_r - \boldsymbol{\mu}_g||_2^2 + \mathrm{Tr}\left(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2(\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_g)^{\frac{1}{2}}\right).$$

Similar to MMD, Lower FID is better, corresponding to more similar real and generated samples as measured by the distance between their activation distributions.

For Inception Score, the score is calculated through the Inception model [Szegedy *et al.*, 2016]. It has been shown that Inception Score is very sensitive to the model parameters [Barratt and Sharma, 2018]. Even the score produced by the same model trained using different libraries (e.g. Tensorflow, Keras, PyTorch) differ a lot from each other. It also requires a large sample size for the accurate estimation for $p_{\mathcal{M}}(\mathrm{y})$. FID and MMD both measure the similarity between training images and generated images based on the feature space [Xu *et al.*, 2018], since the pixel representations of images do not naturally support for meaningful Euclidean distances to be computed [Forsyth and Ponce, 2003]. The main concern for these two methods is whether the distributional characteristics of the feature space exactly reflects the distribution for the images [Koh and Liang, 2017].

We listed the supported features of Neuroscore and traditional metrics in Table. 1. Neuroscore can not only evaluate image quality as other metrics, but also have 3 unique characteristics, which will be demonstrated in section 5.

Feature	IS	MMD	FID	**Neuroscore**
Evaluate image quality	✓	×	✓	✓
Consistent with human	×	×	×	✓
Small sample size	×	×	×	✓
Rank images	×	×	×	✓

Table 1: Comparison between Neuroscore and other metrics.

3 Preliminaries

3.1 Generative Adversarial Networks

A generative adversarial network (GAN) has two components, the discriminator D and the generator G. Given a distribution $\boldsymbol{z} \sim p_{\boldsymbol{z}}$, G defines a probability distribution p_g as the distribution of the samples $G(\boldsymbol{z})$. The objective of a GAN is to learn the generator's distribution p_g that approximates the real data distribution p_r. Optimization of a GAN is performed with respect to a joint loss for D and G

$$\min_G \max_D \mathbb{E}_{\mathbf{x}\sim p_r}\log[D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}}\log\left[1 - D(G(\mathbf{z}))\right].$$

3.2 P300 (or P3) Component and Preprocessing

In neuroscience, the P300 ERP component is a voltage changes in the brain that occurs in response to a target stimulus [Polich, 2007], that can be measured as EEG. It reflects a participant's attention, which can be modulated by the specific instruction given to a participant. The P300 response elicited by a target stimulus is typically evident between 300ms-600ms post stimulus presentation depending on the type of task. EEG is normally recorded by using multiple channels e.g. 32 channels, which makes it difficult to estimate the P300 source amplitude. We use LDA beamformer [Treder *et al.*, 2016; Wang *et al.*, 2018b] to reconstruct P300 source signal from the recorded raw EEG epochs.

Briefly, given a target EEG epoch $\mathbf{X}_i \in \mathbb{R}^{C\times T}$ and a standard EEG epoch $\mathbf{K}_i \in \mathbb{R}^{C\times T}$ (C is the number of channels and T is time points in each EEG epoch). The optimization problem for the LDA beamformer is to find a projection vector $\mathbf{w} \in \mathbb{R}^{C\times 1}$ that solves the optimization problem:

$$\min_{\mathbf{w}} \mathbf{w}^\top \mathbf{\Sigma} \mathbf{w} \ \text{ s.t.} \mathbf{w}^\top \mathbf{p} = 1, \tag{1}$$

where $\mathbf{\Sigma} \in \mathbb{R}^{C\times C}$ is the EEG epoch covariance matrix ($\mathbf{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i \mathbf{X}_i^\top$, N is number of trials) and $\mathbf{p} \in \mathbb{R}^{C\times 1}$ is the spatial pattern difference between target and standard condition [Treder *et al.*, 2016]. The closed-form solution is

$$\mathbf{w} = \mathbf{\Sigma}^{-1}\mathbf{p}(\mathbf{p}^\top \mathbf{\Sigma}^{-1}\mathbf{p})^{-1}. \tag{2}$$

The source signal of each single trial $\mathbf{s}$ can be obtained as

$$\mathbf{s} = \mathbf{w}^\top \mathbf{X}_i = (\mathbf{p}^\top \mathbf{\Sigma}^{-1}\mathbf{p})^{-1}\mathbf{p}^\top \mathbf{\Sigma}^{-1}\mathbf{X}_i, \tag{3}$$

where $\mathbf{s} \in \mathbb{R}^{1\times T}$. Hence, LDA beamformer enables transformation of multi-channel EEG epochs to single-channel EEG epochs facilitating more robust measurement of the P300.

4 Methodology

4.1 Neuroscore

We used a rapid serial visual presentation (RSVP) paradigm [Wang *et al.*, 2016, 2018a] to elicit the P300 ERP. Our experimental procedure is illustrated in [Wang *et al.*, 2018c]. We average the single-trial P300 amplitude (as Neuroscore) to mitigate the background EEG noise [Polich, 2007], which renders a stable measurement of the EEG response to a typical type of stimulus. In general, our Neuroscore is calculated via two steps: (1) Reconstruct P300 source signal from raw EEG; (2) Average the P300 amplitude of each reconstructed single trial source signal across trials (see Algorithm 1).

The proposed Neuroscore reflects a human's perceptual response to different GANs via EEG measurements, thus it is consistent with the human perceptual judgment to GANs.

4.2 Brain-inspired Framework

We propose a brain-inspired framework in order to generalize the use of Neuroscore. This kind of framework is used for predicting Neuroscore given images generated by one of the popular GAN models. Figure. 1 demonstrates the brain-inspired framework used in this work[1]. 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. 1 demonstrates the brain-inspired deep convolutional neural network (CNN) framework. 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. 1).

Algorithm 1 Calculation of Neuroscore

Input:

- $\mathbf{X} \in \mathbb{R}^{N\times C\times T}$ is the EEG signal corresponding to the target stimulus, where N is the number of target trials, C is the number of channels, and T is the number of time points.
- $\mathbf{K} \in \mathbb{R}^{M\times C\times T}$ is the EEG signal corresponding to the standard stimulus, M is number of standard trials, C is number of channels, T is number of time points. The target and standard EEG trials are already explained in section 3.2.

Output: Neuroscore

$\mathbf{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i \mathbf{X}_i^\top + \frac{1}{M}\sum_{i=1}^{M} \mathbf{K}_i \mathbf{K}_i^\top$
for t_i in [400 ms, 600 ms] **do**
  $\mathbf{p} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_{i,t_i} - \frac{1}{M}\sum_{i=1}^{M} \mathbf{K}_{\boldsymbol{i,t_i}}$
  $\mathbf{w} = \mathbf{\Sigma}^{-1}\mathbf{p}(\mathbf{p}^\top \mathbf{\Sigma}^{-1}\mathbf{p})^{-1}$
  $\mathrm{J}_{t_i} \leftarrow \mathbf{w}^\top \mathbf{\Sigma} \mathbf{w}$
  $\mathrm{W}_{t_i} \leftarrow \mathbf{w}$
end for
$t_{optimal}$=argmin$_{t_i}$J
$\mathbf{w}_{optimal}$=W$_{t_{optimal}}$
t$_{P300}$=[t$_{optimal}$ - 100 ms, t$_{optimal}$ + 100 ms] ▷ *This is time window being detected for P300.*
for $i = 1 : N$ **do**
  $\mathbf{s} = \mathbf{w}_{optimal}^\top \mathbf{X}_i$
  $a = \max(\mathbf{s}_{t_{p300}})$
  $\mathrm{A}_i \leftarrow$ a
end for
Neuroscore = $\frac{1}{N}\sum_{i=1}^{N} \mathrm{A}_i$

4.3 Training Details

Mobilenet V2, Inception V3 and Shallow network were explored in this work, where in flow 2 we use these three net-

[1]We understand that human being's brain system is much more complex than what we demonstrated in this work and the flow in the brain is not one-directional [She *et al.*, 2016, 2018]. Our framework can be further extended to be more biologically plausible.

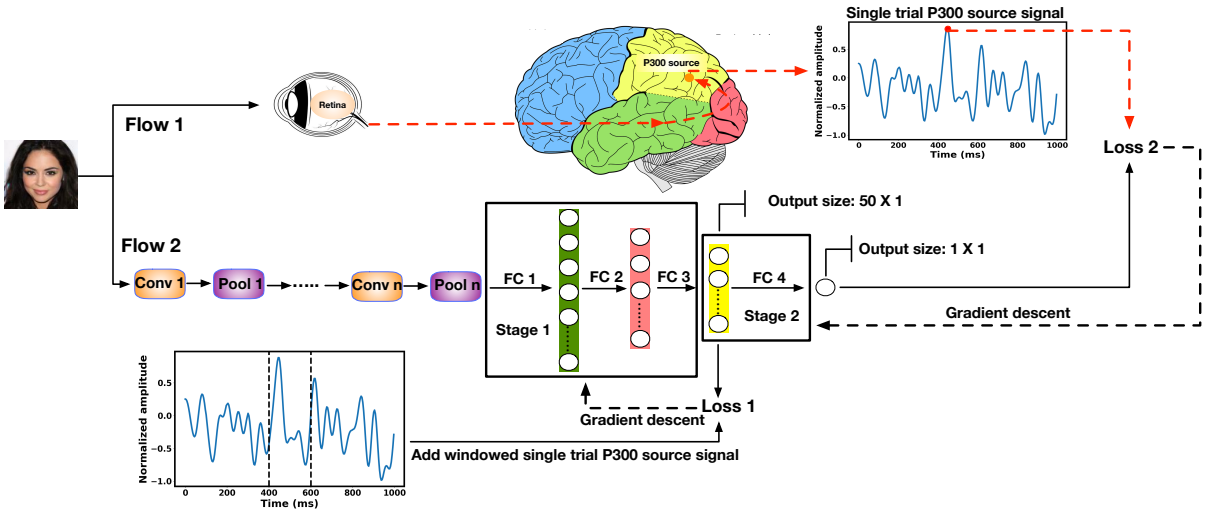

Figure 1: Brain-inspired framework and training details with adding EEG information. Our training strategy includes two stages: (1) Learning from image to P300 source signal; (2) Learning from P300 source signal to P300 amplitude. loss_1 is the L_2 distance between the yellow layer and the single trial 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 4.3.

work bones: such as Conv1-pooling layers. For Mobilenet V2 and Inception V3. We used pretrained parameters from up to the FC 1 shown in Fig. 1. 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 added EEG to the model because we first want to find a function $f(\chi) \rightarrow \mathbf{s}$ that maps the images space χ to the corresponding single trial P300 source signal $\mathbf{s}$. This prior knowledge can help us to predict the single trial P300 amplitude in the second learning stage.

We compared the performance of the models with and without EEG for training. 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

$$\begin{aligned} \text{loss}_1(\boldsymbol{\theta}_1) &= \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{S}_i^{true} - \mathbf{S}_i^{pred}(\boldsymbol{\theta}_1)\|_2^2, \\ \text{loss}_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= \frac{1}{N}\sum_{i=1}^{N}(\mathrm{y}_i^{true} - \mathrm{y}_i^{pred}(\boldsymbol{\theta}_1, \boldsymbol{\theta_2}))^2, \end{aligned} \tag{4}$$

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 explained in Algorithm 2 and visualized in the "Flow 2" of Fig. 1 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.

Algorithm 2 Two training stages with EEG information.

Stage 1: Training parameters $\boldsymbol{\theta}_1$.
Input: Images and averaged P300 signal $\mathbf{S}_i^{true}$.
for number of training iterations **do**
  Update $\boldsymbol{\theta}_1$ by descending its stochastic gradient: $\nabla_{\boldsymbol{\theta}_1}\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{S}_i^{true} - \mathbf{S}_i^{pred}(\boldsymbol{\theta}_1)\|_2^2$
end for
Stage 2: Freezing $\boldsymbol{\theta}_1$, training parameters $\boldsymbol{\theta}_2$.
Input: Images and single trial P300 amplitude y_i^{true}.
for number of training iterations **do**
  Update $\boldsymbol{\theta}_2$ by descending its stochastic gradient: $\nabla_{\boldsymbol{\theta}_2}\frac{1}{N}\sum_{i=1}^{N}(\mathrm{y}_i^{true} - \mathrm{y}_i^{pred}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2))^2$
end for

5 Results

5.1 EEG Improves Model Performance

Individual Participant Performance. Three models have been validated for each individual participant as shown in Fig. 2. It can be seen that all three models trained with EEG outperform the models without EEG with smaller error and variances across almost all the individual subjects. Few failures here might result from the number of EEG trials of individual participant is not sufficient enough for training deep networks to learn the mapping function $f(\chi)$ from image to EEG.

Cross Participant Performance. Table. 2 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.

Adding EEG information reduces error in all three models (as the same error shown in Fig. 2), which are 0.151, 0.168 and **0.171** for Shallow, Mobilenet, and Inception respectively. This indicates that the Inception model bene-

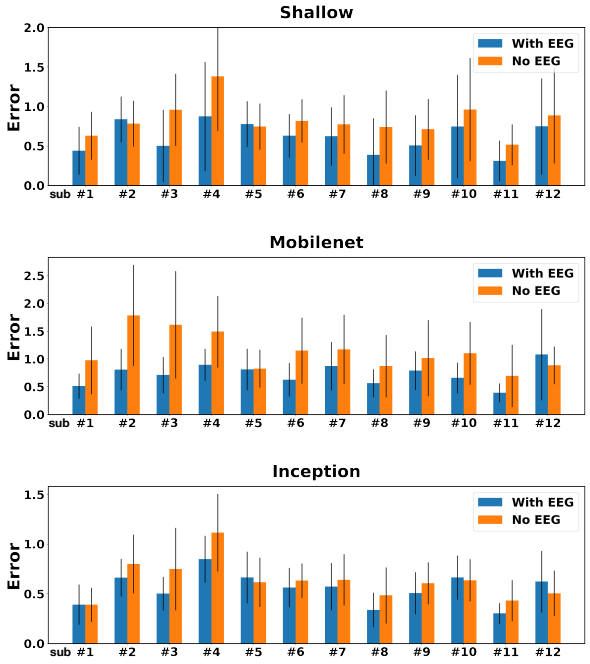

Figure 2: Error of 3 models with and without EEG. Error is defined as: $\sum_i^m |\text{Neuroscore}_{pred}^{(i)} - \text{Neuroscore}_{true}^{(i)}|$, where $\text{m} = 3$ is the number of GAN category used (DCGAN, BEGAN, PROGAN, 12 participants) and Neuroscore is obtained by averaging single trial P300 amplitude. A smaller value indicates better performance.

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

Table 2: Errors of 9 models for cross participants ("-EEG" indicates models are trained with paired EEG, "-EEG$_{\text{random}}$" 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.

fits the most when adding EEG information in the training stage. 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. We used the randomized EEG signal here as a baseline to see the efficacy of adding EEG to produce better Neuroscore output. When randomizing the EEG, it shows that the error for each three model increases significantly. For Mobilenet and Inception, the error of the randomized EEG is even higher than those without EEG in the training stage, demonstrating that the EEG information in the training stage is crucial to each model.

Figure. 3 shows that the models with EEG information have a stronger correlation between predicted Neuroscore and real Neuroscore. The cluster (blue, orange, and green circles) for each category of the model trained with EEG (left column) is more separable than the cluster produced by model without EEG (right column). This conveys with EEG for training models: (1) Neuroscore is more accurate; and (2) Neuroscore is able to rank the performances of different GANs, which cannot be achieved by other metrics [Borji, 2018].

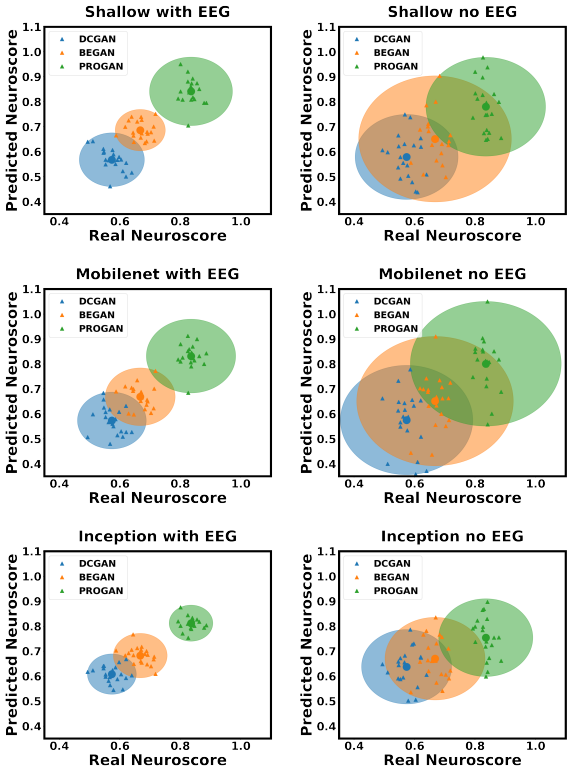

Figure 3: Scatter plot of predicted and real Neuroscore of 6 models (Shallow, Mobilenet, Inception with and without EEG for training) cross participants by 20 times repeated shuffling training and testing set. Each circle represents the cluster for a specific category. Small triangle markers inside each cluster correspond to each shuffling process. The dot at the center of each cluster is the mean.

5.2 Neuroscore Aligns with Human Perceptions

Figure. 5(a) shows the correlation between real Neuroscore and human judgment (BE accuracy) according to three GANs: BEGAN, DCGAN, and PROGAN. The statistical test demonstrates the strong correlation between those two variables. This indicates that Neuroscore can be used to evaluate GANs as it reflects human perceptual judgment.

We have already demonstrated the Neuroscore derived from raw EEG is consistent with the human perception. We are going to demonstrate the same property of Neuroscore predicted from the brain-inspired framework. We compare the Neuroscore with three widely used evaluation metrics. The ultimate goal of GANs is to generate images that are indistinguishable from real images by human beings. Therefore, consistency between an evaluation metric and human

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
Ours	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 3: Three conventional scores: Inception Score (IS), Maximum Mean Discrepancy (MMD), Fréchet Inception Distance (FID), and Neuroscore produced by three models with EEG for each GAN category. A lower score indicates better performance of GAN. Neuroscore is consistent with human judgments.

perception is a critical requirement for the metric to be considered good. Table. 3 shows the comparison between Neuroscore and three traditional scores. To be consistent with all the scores (smaller score indicates better GAN), we used 1/IS and 1/Neuroscore for comparisons in the Table. 3. It can be seen that human ranks the GAN performance as: PROGAN $>$ BEGAN $>$ DCGAN. All three Neuroscores produced by three models with EEG are consistent with human judgment while other three conventional scores are not (they all think that DCGAN outperforms BEGAN).

5.3 Neuroscore Needs Much Smaller Samples

The number of samples for evaluations is crucial in real-world applications considering computational efficiency and efforts for labeling. Traditional metrics need a large sample size to capture the underlying statistical properties of the real and generated images [Salimans *et al.*, 2016; Xu *et al.*, 2018]. In practice, it should prefer the metric is not very sensitive to the sample size. i.e. the small sample size can also make a good estimation. Figure. 5(b) shows that Neuroscore converges stably at around 20 presentations of a specific image (for signal-enhancement purposes), which is much less than the thousands of images required by traditional methods [Borji, 2018; Xu *et al.*, 2018]. This is due to the fact that the P300 becomes stable when dozens of EEG trials corresponding to one category are available.

5.4 Neuroscore Can Rank Images

Another property of using Neuroscore is the ability to track the quality of an individual image. Traditional evaluation metrics are unable to score each individual image for two reasons: (1) They need large-scale samples for evaluation; (2) Most methods (e.g. MMD and FID) evaluate GANs based on the dissimilarity between real images and generated images so they are not able to score the generated image one by one. For our proposed method, the score of each single image can also be evaluated as a single trial P300 amplitude. We demonstrate that using the predicted single trial P300 amplitude to observe the single image quality in Fig. 4. This property provides Neuroscore with a novel capability that can observe the variations within a typical GAN. Although Neuroscore and IS are generated from deep neural networks. Neuroscore is more suitable than IS for evaluating GANs in that: (1) It is more explainable than IS as it is a direct reflection of human

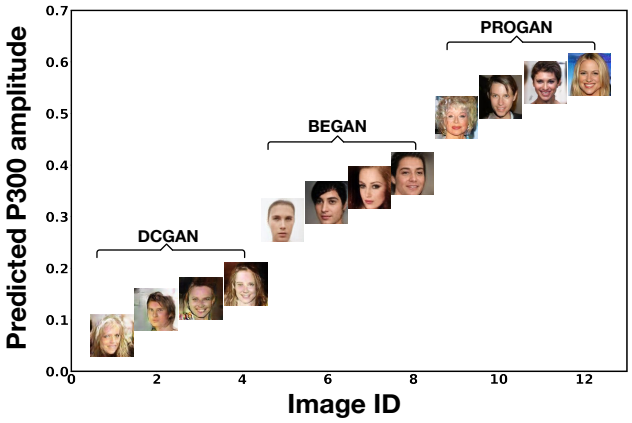

Figure 4: P300 for each single image predicted by the proposed brain-inspired framework in our paper. Higher predicted P300 indicates the better image quality.

perception; (2) Much smaller sample size is required for evaluation; (3) Higher Neuroscore exactly indicates better image quality while IS does not.

5.5 Generalization of Neuroscore

We also included the RFACE images in our generalization test. Figure. 5(c) demonstrates that the predicted Neuroscore

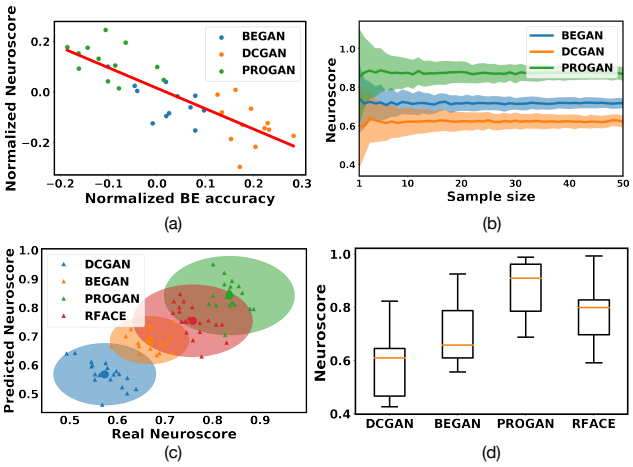

Figure 5: Performances of Neuroscore. (a). Correlation between real Neuroscore and behavioral (BE) accuracy (human judgment in behavioral task) across 12 participants. Neuroscore and BE are both mean centered within each participant. The red line is the linear regression fitting between Neuroscore and BE accuracy. **Pearson statistics:** $r(36) = -0.828, p = 4.766e-10$. (b). Neuroscore of different evaluated sample size for each type of GAN. 200 repeated measurements have been made by randomly shuffling the image samples. (c). Scatter plot between predicted and real Neuroscore. EEG corresponding to real face (RFACE) has been included to test the generalization of the architecture. (d). Boxplot of Neuroscore (from EEG signals) for each image category across 12 participants.

is still correlated with the real Neuroscore when adding the RFACE images and the model ranks the types of images

as: PROGAN>RFACE>BEGAN>DCGAN, which is consistent with the Neuroscore that has been measured directly from participants shown in Fig.5(d).

Compared to traditional evaluation metrics, Neuroscore is able to score the GAN based on very few image samples, relatively. Recording EEG in the training stage could be the limitation of generalizing Neuroscore to evaluate a new GAN. However, the use of dry electrode EEG recording system [Gargiulo *et al.*, 2010] can accelerate and simplify the data acquisition significantly. Moreover, GANs enable the possibility of synthesizing the EEG [Hartmann *et al.*, 2018], which has wide applications in Brain-machine interface research.

6 Conclusion

In this paper, we outline Neuroscore and provide a brain-inspired framework to calculate a synthetic Neuroscore for evaluating the performance of GANs. Three deep network architectures are explored and the results demonstrate that including neural responses during the training phase of the brain-inspired network improves its accuracy even when neural measurements are absent when evaluating on the test set. We compared our Neuroscore measure to traditional evaluation metrics and demonstrated the unique advantages of Neuroscore: (1) It is consistent with human perception; (2) It requires a much smaller number of samples for calculation; and (3) It can rank individual images in terms of quality within a specific GAN.

References

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

Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint:1801.01973*, 2018.

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

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.

David A Forsyth and Jean Ponce. A modern approach. *Computer vision: A Modern Approach*, pages 88–101, 2003.

Gaetano Gargiulo, Rafael A Calvo, Paolo Bifulco, Mario Cesarelli, Craig Jin, Armin Mohamed, and André van Schaik. A new EEG recording system for passive dry electrodes. *Clinical Neurophysiology*, 121(5):686–693, 2010.

Adam D Gerson, Lucas C Parra, and Paul Sajda. Cortically coupled computer vision for rapid image search. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(2):174–179, 2006.

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*, pages 2672–2680, 2014.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Kay Gregor Hartmann, Robin Tibor Schirrmeister, and Tonio Ball. EEG-GAN: Generative adversarial networks for electroencephalographic brain signals. *arXiv preprint:1806.01875*, 2018.

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*, pages 6626–6637, 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*, pages 1885–1894, 2017.

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*, pages 105–114. IEEE, 2017.

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pages 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*, pages 1369–1377, 2016.

Eva Mohedano, Kevin McGuinness, Graham Healy, Noel E O'Connor, Alan F Smeaton, Amaia Salvador, Sergi Porta, and Xavier Giró-i Nieto. Exploring EEG for object detection and retrieval. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 591–594. ACM, 2015.

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

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*, pages 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*, pages 4510–4520. IEEE, 2018.

Qi She, Guanrong Chen, and Rosa HM Chan. Evaluating the small-world-ness of a sampled network: Functional connectivity of entorhinal-hippocampal circuitry. *Scientific reports*, 6:21468, 2016.

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

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*, pages 2818–2826, 2016.

Matthias S Treder, Anne K Porbadnigk, Forooz Shahbazi Avarvand, Klaus-Robert Müller, and Benjamin Blankertz. The LDA beamformer: Optimal estimation of erp source time series using linear discriminant analysis. *Neuroimage*, 129:279–291, 2016.

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*, pages 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*, pages 243–270, 2018.

Zhengwei Wang, Graham Healy, Alan F Smeaton, and Tomas E Ward. Spatial filtering pipeline evaluation of cortically coupled computer vision system for rapid serial vi-

sual presentation. *Brain-Computer Interfaces*, 5(4):132–145, 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. *arXiv preprint arXiv:1811.04172*, 2018.

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.

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.