A Neuro-AI Interface: Learning DNNs from the Human Brain

Abstract

Deep neural networks (DNNs) are inspired from the human brain and the interconnection between the two has been widely studied in the literature. However, it is still an open question whether DNNs are able to make decisions like the brain. Previous work has demonstrated that DNNs, trained by matching the neural responses from inferior temporal (IT) cortex in monkey's brain, is able to achieve human-level performance on the image object recognition tasks. This indicates that neural dynamics can provide informative knowledge to help DNNs accomplish specific tasks. In this paper, we introduce the concept of a neuro-AI interface, which aims to use human's neural responses as supervised information for helping AI systems solve a task that is difficult when using traditional machine learning strategies. In order to deliver the idea of neuro-AI interfaces, we focus on deploying it to one of the fundamental problems in generative adversarial networks (GANs): designing a proper evaluation metric to evaluate the quality of images produced by GANs.

1. Introduction

Deep neural networks (DNNs) have successfully been applied to a number of different areas such as computer vision and natural language processing where they have demonstrated state-of-the-art results, often matching and even sometimes surpassing a human's ability. Moreover, DNNs have been studied with respect to how similar processing is carried out in the human brain, where identifying these overlaps and interconnections has been a focus of study and investigation in the literature (Cichy et al., 2016; Cichy & Kaiser, 2019; Groen et al., 2018; Kuzovkin et al., 2018; Tu et al., 2018; Batista & DiCarlo, 2018; Yamins & DiCarlo, 2016; Kriegeskorte, 2015; Kheradpisheh et al., 2016). In this research area, convolutional neural networks (CNNs) are widely studied to be compared with the visual system in human's brain because of following reasons: (1) CNNs and human's visual system are both hierarchical system; (2) Steps of processing input between CNNs and human's visual system are similar to each other e.g., in a object recognition task, both CNNs and human recognize a object based on their its shape, edge, color etc.. Work (Yamins & DiCarlo, 2016) outlines the use of CNNs approach for delving even more deeply into understanding the development and organization of sensory cortical processing. It has been demonstrated that CNNs are able to reflect the spatio-temporal neural dynamics in human's brain visual area (Cichy et al., 2016; Tu et al., 2018; Kuzovkin et al., 2018). Despite lots of work is carried out to reveal the similarity between CNNs and brain system, research on interacting between CNNs and neural dynamics is less discussed in the literature as understanding of neural dynamics in the neuroscience area is still limited.

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) and speech synthesis (Donahue et al., 2018). 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) 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 meaning-

ful 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 (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 artefact 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. The within GAN variances are crucial because it can provide the insight on the variability of that GAN.

Work (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, 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 more difficult problems in 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.

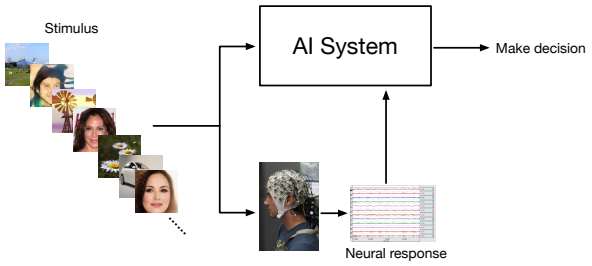


Figure 1. Schematic of neuro-AI interface. Stimuli (image stimuli used in this work) are simultaneously presented to an AI system and participants. Participants' neural responses are transferred to the AI system as supervised information for assisting the AI system make decision.

In this paper, we first demonstrate the ability of a brain-produced score (we call it 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 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) (Polich, 2007) present in EEG, via a rapid serial visual presentation (RSVP) paradigm (Wang et al., 2018a). 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).

2. Related work

Current literature has demonstrated that CNNs are able to predict neural responses in inferior temporal cortex in image recognition task (Yamins et al., 2014; Yamins & DiCarlo, 2016) via invasive BCI techniques (Waldert, 2016). Evidence shows that neural responses in inferior temporal cortex directly link the information processing during the image recognition task. Therefore, a CNN trained by predicting neural responses in inferior temporal cortex also achieves the good performance during the image recognition (Yamins et al., 2014). Comparing the traditional end-to-end machine learning system, use of DNNs for predicting neural responses in the brain favors following benefits: (1) It enables the information processing of DNNs closer to human being's brain system; (2) For some difficult tasks in real-world e.g, evaluation of image quality demonstrated in this paper, it is still challenging to design the machine learning algorithms, which teach DNNs to process the information like humans; and (3) Neural signals directly reflect the human perception and interfacing between neural responses and DNNs can be more efficient than the traditional methods regarding the area of human and AI.

The investigation of using CNNs to predict neural response from non-invasive BCI aspect is still blank in the literature. Comparing to invasively measured neural dynamics, EEG favors pros such as simple measurement, unpainful experience during recording, free to ethic argument and more easily generalized to real-world applications. However, EEG suffers challenges such as low signal quality (i.e.,

low SNR), low spatial resolution (interested neural activities span all scalp and difficult to be localized), which makes the prediction for EEG response still challenging.

With advanced machine leaching technologies applied to non-invasive BCI area, source localization and reconstruction are feasible for EEG signals today. Previously work (Wang et al., 2018b;a) have demonstrated the efficacy of using spatial filtering approaches for reconstructing P300 source ERP signals. The low SNR issue can be remedied by averaging the EEG trials. Based on this evidence, we explore the use of DNNs to predict Neuroscore when neural information is available.

3. Methodology

3.1. Neuro-AI Interface

We propose a neuro-AI interface in order to generalize the use of Neuroscore. This kind of framework interfaces between neural responses and AI systems (CNN used in this study), which uses neural responses as supervised information to train a CNN. The trained CNN is then used for predicting Neuroscore given images generated by one of the popular GAN models. Figure 2 demonstrates the schematic of neuro-AI interface 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. 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).

3.2. 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. 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.

[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.

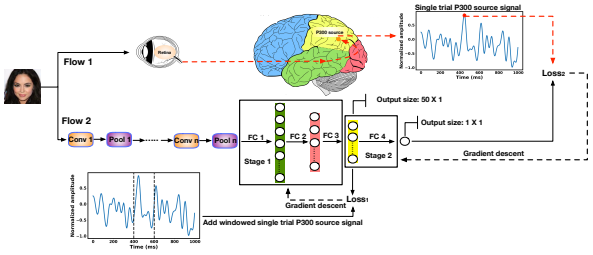


Figure 2. A neuro-AI interface and training details with adding EEG information. Our training strategy includes two stages: (1) Learning from image to P300 source signal; and (2) Learning from P300 source signal to P300 amplitude. loss_1 is the L_2 distance between the yellow layer and the single trial 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 3.2.

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

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

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 explained in Algorithm 1 and 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.

4. Results

Table 1 shows the error for each model with EEG signal, with randomized EEG signal **within each type of GAN**

Algorithm 1 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

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 1. 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. Error is defined as: $\sum_i^m |\text{Neuroscore}_{pred}^{(i)} - \text{Neuroscore}_{true}^{(i)}|$, where m $= 3$ is the number of GAN category used (DCGAN, BEGAN, PROGAN).

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. 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).

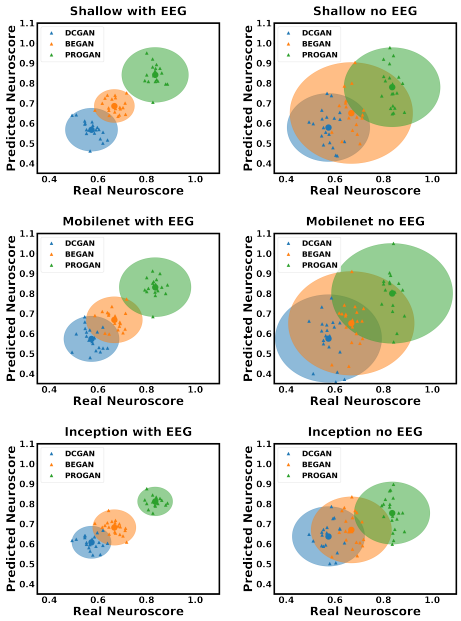


Figure 3. Scatter plot of predicted and real Neuroscore of 6 models (Shallow, Mobilenet, Inception with and without EEG for training) cross participants by 20 times repeated shuffling training and testing set. Each circle represents the cluster for a specific category. Small triangle markers inside each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

5. 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 the performance of Neuroscore can be referred in Appendix.

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. *arXiv preprint:1701.07875*, 2017.

Batista, A. P. and DiCarlo, J. J. Deep learning reaches the motor system. *Nature methods*, 15(10):772, 2018.

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

Chelazzi, L., Miller, E. K., Duncan, J., and Desimone, R. A neural basis for visual search in inferior temporal cortex. *Nature*, 363(6427):345, 1993.

Cichy, R. M. and Kaiser, D. Deep neural networks as scientific models. *Trends in cognitive sciences*, 2019.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6:27755, 2016.

Donahue, C., McAuley, J., and Puckette, M. Synthesizing audio with generative adversarial networks. *arXiv preprint:1802.04208*, 2018.

Fedus, W., Goodfellow, I., and Dai, A. M. MaskGAN: Better text generation via filling in the _ . *arXiv preprint:1801.07736*, 2018.

Gargiulo, G., Calvo, R. A., Bifulco, P., Cesarelli, M., Jin, C., Mohamed, A., and van Schaik, A. A new EEG recording system for passive dry electrodes. *Clinical Neurophysiology*, 121(5):686–693, 2010.

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

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Groen, I. I., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., and Baker, C. I. Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *Elife*, 7:e32962, 2018.

Hartmann, K. G., Schirrmeister, R. T., and Ball, T. EEG-GAN: Generative adversarial networks for electroencephalographic brain signals. *arXiv preprint:1806.01875*, 2018.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. 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.

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

Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., and Masquelier, T. Deep networks can resemble human feed-forward vision in invariant object recognition. *Scientific reports*, 6:32672, 2016.

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

Kriegeskorte, N. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1:417–446, 2015.

Kuzovkin, I., Vicente, R., Petton, M., Lachaux, J.-P., Baciu, M., Kahane, P., Rheims, S., Vidal, J. R., and Aru, J. Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. *Communications biology*, 1(1):107, 2018.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., 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.

Li, Y., Swersky, K., and Zemel, R. Generative moment matching networks. In *International Conference on Machine Learning*, pp. 1718–1727, 2015.

McIntosh, L., Maheswaranathan, N., Nayebi, A., Ganguli, S., and Baccus, S. Deep learning models of the retinal response to natural scenes. In *Advances in Neural Information Processing Systems*, pp. 1369–1377, 2016.

Mohedano, E., McGuinness, K., Healy, G., O'Connor, N. E., Smeaton, A. F., Salvador, A., Porta, S., and Giró-i Nieto, X. Exploring EEG for object detection and retrieval. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pp. 591–594. ACM, 2015.

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

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training

GANs. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520. IEEE, 2018.

She, Q., Chen, G., and Chan, R. H. Evaluating the small-world-ness of a sampled network: Functional connectivity of entorhinal-hippocampal circuitry. *Scientific reports*, 6: 21468, 2016.

She, Q., Gao, Y., Xu, K., and Chan, R. H. Reduced-rank linear dynamical systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Tu, T., Koss, J., and Sajda, P. 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.

Waldert, S. Invasive vs. non-invasive neuronal signals for brain-machine interfaces: will one prevail? *Frontiers in neuroscience*, 10:295, 2016.

Wang, Z., Healy, G., Smeaton, A. F., and Ward, T. E. 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, 2018a.

Wang, Z., Healy, G., Smeaton, A. F., and Ward, T. E. Spatial filtering pipeline evaluation of cortically coupled computer vision system for rapid serial visual presentation. *Brain-Computer Interfaces*, 5(4):132–145, 2018b.

Xu, Q., Huang, G., Yuan, Y., Guo, C., Sun, Y., Wu, F., and Weinberger, K. Q. An empirical study on evaluation metrics of generative adversarial networks. *arXiv preprint:1806.07755*, 2018.

Yamins, D. L. and DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356, 2016.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. Generative image inpainting with contextual attention. *arXiv preprint:1801.07892*, 2018.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint:1703.10593v6*, 2017.

Appendix

5.1. Neuroscore performance

Figure 4 shows the *averaged reconstructed P300 signal* across all participants (using LDA beamformer) in the RSVP experiment. It should be noted here that the *averaged reconstructed P300 signal* is calculated as the difference between averaged target trials and averaged standard trials after applying the LDA beamformer method. The solid lines in Figure 4 are the means of the averaged reconstructed P300 signals for each image category (across 12 participants) while the shaded areas represent the standard deviations (across participants). It can be seen that the averaged reconstructed P300 (across participants) clearly distinguishes between different image categories.

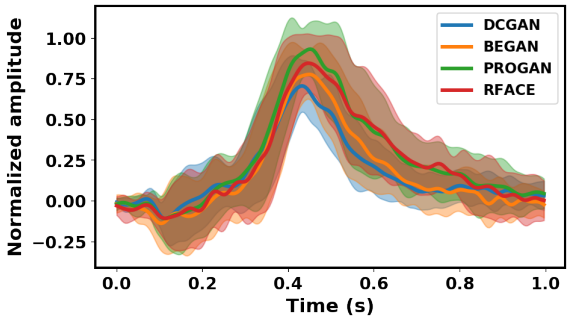


Figure 4. Averaged reconstructed (via LDA beamformer) P300 signal across 12 participants in this study.

In order to statistically measure this correlative relationship, we calculated the Pearson correlation coefficient and p-value (two-tailed) between Neuroscore and BE accuracy and found ($\mathbf{r(48) = -0.767, p = 2.089e-10}$). We also did the Pearson statistical test and bootstrap on the correlation between Neuroscore and BE accuracy (human judgment performance) only for GANs i.e., DCGAN, BEGAN and PROGAN. Pearson statistic is (**r(36)=-0.827, p=4.766e-10**) and the bootstrapped $\mathbf{p \leq 0.0001}$.

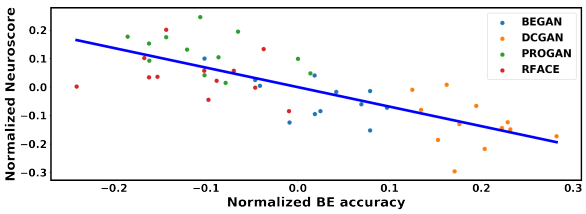


Figure 5. Correlation between NS and BE accuracy. Neuroscore and BE are both mean centered within each participant.

5.2. Comparison to other evaluation metrics

Three traditional methods are also employed to evaluate the GANs used in this study. Table 2 shows the scores from the three traditional metrics, Neuroscore and human judgment for three GANs. To be consistent with other metrics (smaller score indicates better GAN performance), we use 1/Neuroscore for comparison. It can be seen that all three methods are consistent with each other and they rank the GANs in the same order of PROGAN, DCGAN and BEGAN from high to low performance. By comparing the three traditional evaluation metrics to the human, it can be seen that they are not consistent with human judgment of GAN performance. It should be remembered that Inception Score is able to measure the quality of the generated images (Salimans et al., 2016) while the other two methods cannot do so. However, Inception Score still rates DCGAN as outperforming BEGAN. Our proposed Neuroscore is consistent with human judgment.

Methods	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
1/Neuroscore	1.715	1.479	1.195
Human	**0.995**	**0.824**	**0.705**

Table 2. Score comparison for each GAN category. Three conventional scores: Inception Score (IS), Maximum Mean Discrepancy (MMD), Fréchet Inception Distance (FID), and Neuroscore are compared with each other. Lower score indicates better performance of GAN.

5.3. Performance of neuro-AI interface

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. 6. 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 perception; (2) Much smaller sample size is required for evaluation; (3) Higher Neuroscore exactly indicates better image quality while IS

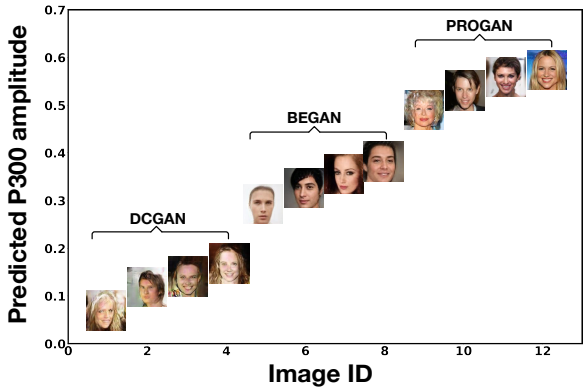
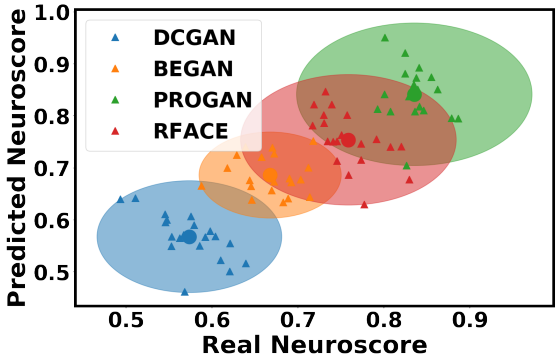
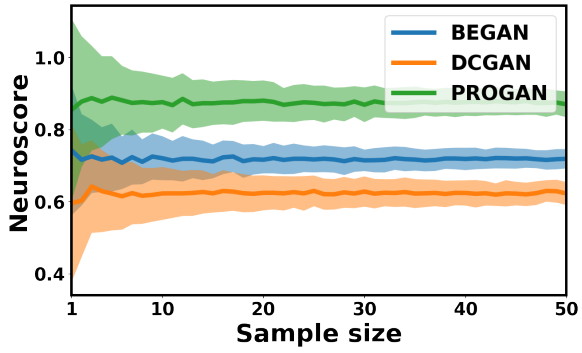


Figure 6. P300 for each single image predicted by the proposed neuro-AI interface in our paper. Higher predicted P300 indicates the better image quality.

does not.



(a)



(b)

Figure 7. Performances of the CNN model trained by including EEG response. (a). Scatter plot between predicted and real Neuroscore. EEG corresponding to real face (RFACE) has been included to test the generalization of the architecture. (b). Neuroscore of different evaluated sample size for each type of GAN. 200 repeated measurements have been made by randomly shuffling the image samples.

We also included the RFACE images in our generalization test. Figure 7(c) demonstrates that the predicted Neuroscore 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.7(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.