A Feature Filter for EEG Using Cycle-GAN Structure

Yue Yao(✉), Jo Plested, and Tom Gedeon

Research School of Computer Science,
The Australian National University, Canberra, Australia
u6014942@anu.edu.au

Abstract. The brain-computer interface (BCI) has become one of the most important biomedical research fields and has created many useful applications. As an important component of BCI, electroencephalography (EEG) is in general sensitive to noise and rich in all kinds of information from our brain. In this paper, we introduce a new strategy to filter out unwanted features from EEG signals using GAN-based autoencoders. Filtering out signals relating to one property of the EEG signal while retaining another is similar to the way we can listen to just one voice during a party. This approach has many potential applications including in privacy and security. We use the UCI EEG dataset on alcoholism for our experiments. Our experiment results show that our novel GAN based structure can filter out alcoholism information for 66% of EEG signals with an average of only 6.2% accuracy lost.

Keywords: Deep learning · EEG · Brain-Computer interface
Image translation · Generative adversarial nets

1 Introduction

Being an essential input signal of a Brain-Computer Interface (BCI), EEG has been harnessed in a variety of interesting and useful applications for users and has changed our life in various areas. The EEG is defined as the overall measurement of human brain electrical activity using electrodes placed on the scalp. Since it is an overall measurement, this makes EEG applicable to diverse areas like personal recognition [1], disease identification [2, 3], sleep stage classification [4], even to rebuild the picture from a person's eyes [5], and so on.

Taking personal recognition as an example, compared with fingerprint or face recognition, EEG has more advantages in identifying different people because it has a higher safety factor. For instance, if one person's fingerprint is stolen or one person's face is reconstructed by others, it is basically an irreparable problem because both fingerprint and face model are irrevocable without expensive and painful plastic surgery. But for EEG data, if it is hacked by others, users can still reset a new EEG pattern because the EEG recognizer can identify a person by both personal details and personal brain action [1].

L. Cheng et al. (Eds.): ICONIP 2018, LNCS 11307, pp. 567–576, 2018.
https://doi.org/10.1007/978-3-030-04239-4_51

But being full of information also means full of personal privacy issues for personal identification. For example, if we would like to use EEG for a personal recognition task for a bank, the only information we would like to upload is personal identity-related information and not share the full EEG with the bank, as it may contain information related to a disease or condition we may have. But since there currently does not exist a suitable information filtering algorithm, both the bank and hackers will also be able to get our other information like disease information, emotion information and so on.

But to be able to filter out unwanted features faces many difficulties. First, the property we mentioned of EEG being full of information also generally means full of noise and interference, making it hard to filter out unwanted features exactly. Second, filtering out features is not as easy as cutting off a bounding box in computer vision (CV), it is more like a transformation from the whole since EEG is not interpretable for all its features. Third, filtering out unwanted features also means we need to retain normal EEG trial properties, and we have to make sure our desired features are maintained during the operation.

As a result, we consider deep learning methods, which have achieved success in many areas like CV and natural language processing (NLP). In practice, we do not use the idea of subtracting features to filter out properties as such properties are not well-defined. Instead, we choose to generate a new EEG trial without the unwanted features but maintaining the desired features of the original EEG trial signal. So as the result, a generative adversarial network (GAN) based technique is utilized to create such an EEG signal. In this paper, we introduce GAN-based autoencoders, which is as an extension of our previous work [6]. As mentioned earlier, the feature filter of EEG is more like a style transformation. So we are inspired by the idea of Image-to-Image translation [7] introduced in the computer vision area. This approach is designed to map one image distribution to another image distribution in order to achieve a style transformation. In our paper, such a translation mechanism is used for feature filtering.

2 Related Work

EEG2Image is a work designed to transfer EEG signals to images which is derived from Bashivan's work [8]. Shown in Fig. 1, each trial of EEG is transformed to a colored image using both the time-series information and electrode location information. The transformation procedure is as follows. First, for a single trial of EEG signal, Fast Fourier Transform (FFT) is performed to extract three frequency bands, theta (4–7 Hz), alpha (8–13 Hz), and beta (13–30 Hz). Then, calculate the sum of squared absolute values for

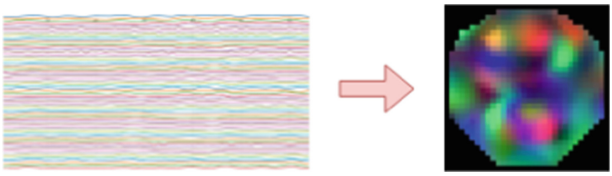

Fig. 1. EEG signal to image example [6]

each frequency band, thereby giving each electrode three scalar values to describe it. Next, using Polar Projection to project 3-D electrode position to 2-D position to create 2-D position sets in a 2-D map with three values to describe it. Then, with CloughTocher scheme to interpolate values between positions, we can produce consistent 2-D color images which reproducibly represent an EEG trial as a full color image.

Generative adversarial networks (GANs) are systems of two neural networks contesting with each other in a minimax game framework [9]. The GAN approach has achieved great success in the image generation area [10–12]. GANs include two main parts, namely a generator and a discriminator. The generator is mainly used to learn the distribution of the real image and produce images in order to fool the discriminator, while the discriminator needs to accept real images while rejecting generated images. Throughout this process, the generator strives to make the generated image more realistic, while the discriminator strives to identify the real image. The key part of GAN is the adversarial loss. For the image generation task, the adversarial loss is very powerful for images in one domain transformed to the other domain since this domain cannot be discriminated by simple rules.

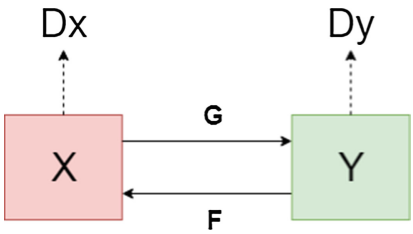

Fig. 2. CycleGAN structure

Fig. 3. Image translation example [13]

Image-to-image translation is a kind of system that can learn the mapping between an input image distribution and an output image distribution using two separate image domains [7]. Shown in Fig. 2, given a source distribution X, we are aiming to use a generative model G to map our source distribution X to target distribution Y. An example is shown in Fig. 3, though it is not perfect, the translation system has successfully transformed between the most important features between zebra and horses like the hide color. In this translation system, we do not explicitly tell the neural

network to change some features. Instead, we have the prior knowledge of two separate image distributions. As a result, it is possible for us to extract the stylistic differences between two image distributions and then directly translate them from one domain to the other domain.

Cycle-Consistent Adversarial Networks (CycleGAN) is a well-known image-to-image translation for unpaired images [13]. It overcomes the difficulty of getting paired images, and forms an autoencoder-like structure to achieve image translation. In Fig. 2, G is such a generator that generates a domain Y image from domain X, while F is the generator that generates a domain X image from the domain Y. Dx and Dy are two discriminators that are used to justify whether the coming image really belongs to domain X or domain Y, respectively. The training procedure can be separated into two symmetric parts. One is $X \rightarrow G(X) \rightarrow F(G(X))$. In this autoencoder-like loop, the training loss comes from two parts, the first is the discriminator loss which comes from Dy to judge whether G(X) is really from domain Y and the second is the reconstruction loss to judge whether F(G(X)) is the same as X or not. The other loop $Y \rightarrow F(Y) \rightarrow G(F(Y))$ is the same in principle.

But all these GAN methods are based on two hypotheses. One is that it is possible to build a strong classifier that can discriminate such features, and the second is the availability of a reliable generator that can filter out original features and rebuild target features. For the first hypothesis, if we cannot train a strong classifier in normal labeled training, it will be almost impossible for us get a strong discriminator in training, because adversarial training itself is not well designed to help train the discriminator. That is not an issue for many GAN based methods which have achieved great success in the CV area, since the most popular current datasets like MNIST [14] and CIFAR-10 [15] have already achieved more than 90% accuracy using different CNNs to serve as accurate discriminators. In contrast to CV, since the NLP area does not have a universally recognized text classification method for grammar checking, current GAN methods for NLP, like Seqgan [16] and its improved version Leakgan [17] do not have a strong discriminator to guide the generator. For our second hypothesis, we have to have a strong generator which can rebuild features. But building a strong generator is strongly related to the given type of data. For the image translation area, convolution and deconvolution-based methods are often used. The U-net [18] based method is the current state of the art [7].

Image-wise autoencoders [6] are the solution we use to meet the two hypotheses of building a GAN for EEG. An image-wise autoencoder is used to extract discriminative and robust features from EEG images. During the autoencoder training, it can reduce reconstruction loss to a very low level for the test set, making it possible to become a generator for the GAN structure. Furthermore, when we connect the features to a fully connected layer to work as a classifier, it achieves convincing results with more than 90% accuracy in the within-subject test [19], showing it has the ability to be a strong discriminator.

3 Methodology

3.1 UCI EEG Dataset

The dataset we use is from UCI; it is a multi-label dataset. This EEG dataset was created by the Neurodynamics Laboratory at the State University of New York. It has a total of 122 subjects with 77 diagnosed with alcoholism and 45 control subjects. Each subject has 120 separate trials [20]. If a subject is labeled with alcoholism, all 120 trials belonging to that subject will be labeled as alcoholism. The stimulus they use are several pictures selected from the Snodgrass and Vanderwart picture set. As a result, for each trial of EEG signal, there are both alcoholism and stimulus information labels.

3.2 Gan-Based Autoencoder

The Gan-based Autoencoder is mainly used for data filtering and the latent representation of this autoencoder is the filtered result we want. Our Gan-based Autoencoder is the same structure as the CycleGAN structure [13]. We call it a GAN-based Autoencoder mainly because it is principally still in a data->latent representation->original data structure and uses reconstruction loss. So in this autoencoder design, we take this latent representation as our filter result. As introduced before, the training procure can be split into two separate training loops, and each loop has two separate losses. The detailed loss definitions are as follows.

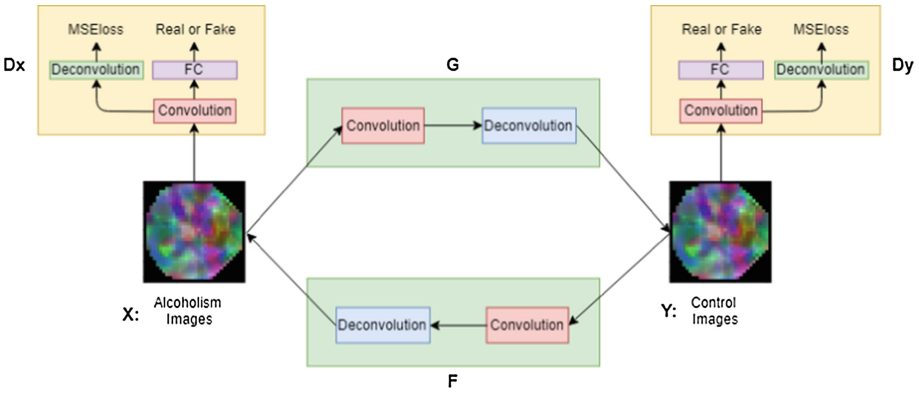

Fig. 4. Structure of GAN-based autoencoder

A. Adversarial Loss:

The adversarial loss is mainly designed to judge whether the incoming image really belongs to a certain distribution. So take loop $X \rightarrow G(X) \rightarrow F(G(X))$ for example, it is designed to map distribution X to distribution Y using generator G. The adversarial loss for this loop is defined as

$$\begin{aligned} L_{GAN}(G, D_Y, X, Y) = & E_{y \sim pdata(y)}[\log D_Y(y)] \\ & + E_{x \sim pdata(x)}[\log (1 - D_Y(G(x)))] \end{aligned} \tag{1}$$

This is a common GAN loss, where G(x) is trying to fool the discriminator D_Y to make the generated image become more similar to image distribution Y. A similar adversarial loss is introduced for loop $Y \rightarrow F(Y) \rightarrow G(F(Y))$.

B. Autoencoder Loss:

The autoencoder loss (reconstruction loss) is mainly used as a regularization term to make sure the generated image is not from random selection, because the target distribution could have multiple choices. The autoencoder loss will help the generator to choose a target image which also maintains some feature(s) from the original image in order to help reduce the reconstruction loss. Also, take loop $X \rightarrow G(X) \rightarrow F(G(X))$ for example, It is defined as:

$$L_{AL}(G, F) = E_{x \sim pdata(x)}[||F(G(x)) - x||_1] \tag{2}$$

The Autoencoder Loss is the same as common autoencoder mean squared loss to judge whether F(G(x)) is really like x or not. A similar autoencoder loss is introduced for loop $Y \rightarrow F(Y) \rightarrow G(F(Y))$ as well.

Table 1. The detailed generator structure

Encoder	Decoder
Input 32 × 32 × 3 Color Image	Input 128 × 8 × 8 Matrix
4 × 4 conv, Leaky ReLU, 4 ×4 conv, Leaky ReLU, 3 × 3 conv, Leaky ReLU, 3 × 3 conv, Leaky ReLU	4 ×4 Deconv, Leaky ReLU, 4 ×4 Deconv, Leaky ReLU, Tanh

Shown in Fig. 4, we use EEG images with alcoholism condition and then map them to an EEG image with the control condition. By doing this transformation, we aim to eliminate alcoholism information from an EEG image while still maintaining its stimulus information. Inspired by the Image-wise autoencoder, shown in Table 1, our modified version of Image-wise autoencoder is now working as our generator, and the combination of Image-wise autoencoder and one fully connected layer works as our discriminator. Adam optimizer is used with 0.0002 learning rate.

3.3 Evaluation Method

The evaluation method for GAN is a difficult problem which needs to take many factors into account [21]. For a long time after the original GAN paper was published, the generated results from GANs still needed to be judged by manual selection in the CV area. But after the critical work from Google brain, the Fréchet Inception Distance (FID) and F1 scores [21] were introduced to judge the generation quality of a GAN. Both the FID and F1 score require a strong pretrained classifier in CV, making it impossible to directly use in the bio-signal area.

Thus, we learn from the idea of using FID and Inception Score (IS) but simply use the idea of training an additional classifier to judge the classification accuracy changes. The classifier we take is still the Image-wise autoencoder with fully connected layer (FC) which is trained separately from adversarial training. In this work, we are trying to filter out alcoholism information while keeping stimulus information. So, the desired best result should be that we get a large alcoholism accuracy reduction while keeping reasonable stimulus accuracy (low stimulus accuracy reduction) through the GAN based autoencoder.

4 Results and Discussion

The picture generated by our GAN-based autoencoder is shown in Fig. 5. These are six generation examples randomly selected from all generation pairs. We can see our GAN works and makes some slight modification to the images. The fact that we cannot see interpretable features from these transformations, means the generated results from our GAN cannot be manually checked. So we turn instead to digital indicators. Here, we only evaluate whether our generated image is really removing features we do not want using the normal Image-wise autoencoder with a classification net [6]. From Fig. 6, we can see that 96.1% of the original images are correctly classified as alcoholism, which is a good result on this dataset and shows our underlying approach works. After our GAN-based autoencoder has processed these images, only 29.8% of the images are classified as alcoholism. That is 2/3 (66.3%) of images have their alcoholism information filtered out. At the same time, only 6.2% accuracy has been lost for stimulus accuracy, and its accuracy still remains well above chance, which is 20% in this case as

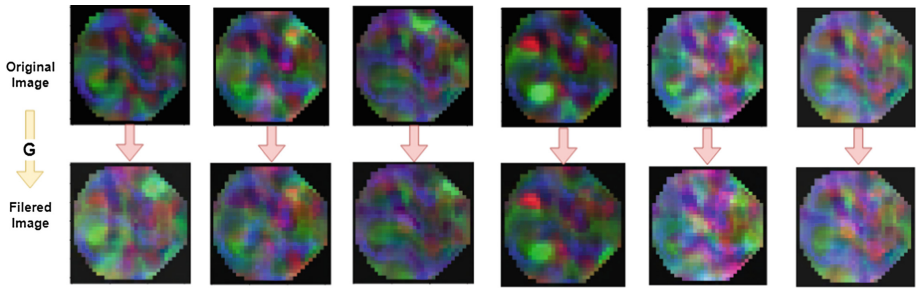

Fig. 5. GAN-based autoencoders output

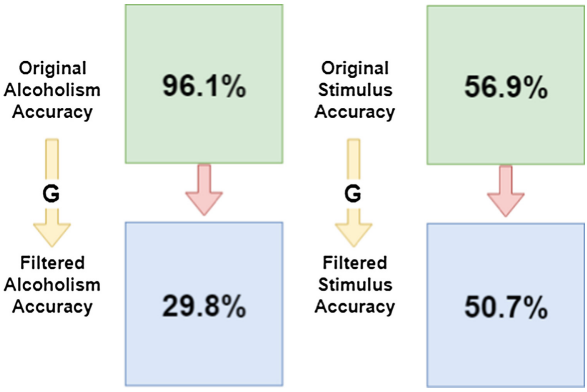

Fig. 6. GAN-based autoencoders performance

there are 5 stimulus conditions. Also, from the figure above, it seems that our GAN-based autoencoder does not change our EEG images much by eye, but it has already filtered out one feature of the original EEG image. That is a very interesting result and further study is needed to determine whether we can remove all alcoholism features while retaining stimulus features without loss of accuracy. As a summary, it turns out that our GAN-based autoencoder can filter out alcoholism information to some extent.

5 Limitation and Future Work

The first limitation is in using accuracy only as performance evaluation. This is an issue because there could be various ways to reduce accuracy like adding random noise or adversarial attacks [22]. One potential solution to this is to check whether such methods can achieve the same performance as GAN-based autoencoder. The second limitation is that there is still a 6.2% accuracy drop in stimulus classification. One possible solution is to try to add a stimulus discriminator to provide a penalty for stimulus information loss. But since the stimulus classifier is currently far from a strong classifier. Our 56.9% is reasonable where chance is 20%, but cannot really be called 'strong'. Thus, the result of adding a stimulus discriminator is not predictable. The third point is future work for the generator, the U-net structure should be tried since it is the current state of the art method for image translation.

6 Conclusion

Removing or filtering features out of EEG signals is difficult, but we have shown some excellent initial results. This approach can lead to many useful applications, such as privacy protection. An example could be where a hospital stores only the medical condition related EEG signal, but the bank stores only personal identification part of an EEG (assuming a future ATM collects EEG for greater security). This paper introduces

GAN-based autoencoders, which transfer the feature filtering task to an image translation task. The experiment results show that our GAN-based autoencoder can filter out a large proportion of unwanted features while mostly keeping desired features, as evaluated by using accuracy drops. Limited by time, the potential of these models is not fully revealed, with further adjustment and fine-tuning, the performance could be increased.

References

1. Kumari, P., Vaish, A.: Brainwave based user identification system: a pilot study in robotics environment. Robot. Auton. Syst. **65**, 15–23 (2015)
2. Truong, N.D., Nguyen, A.D., Kuhlmann, L., Bonyadi, M.R., Yang, J., Kavehei, O.: A Generalised Seizure Prediction with Convolutional Neural Networks for Intracranial and Scalp Electroencephalogram Data Analysis. arXiv preprint arXiv:1707.01976 (2017)
3. Thodoroff, P., Pineau, J., Lim, A.: Learning robust features using deep learning for automatic seizure detection. In: Machine Learning for Healthcare Conference, pp. 178–190 (2016)
4. Ebrahimi, F., Mikaeili, M., Estrada, E., Nazeran, H.: Automatic sleep stage classification based on EEG signals by using neural networks and wavelet packet coefficients. In: 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2008. EMBS 2008, pp. 1151–1154. IEEE (2008)
5. Palazzo, S., Spampinato, C., Kavasidis, I., Giordano, D., Shah, M.: Generative adversarial networks conditioned by brain signals. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3410–3418 (2017)
6. Yao, Y., Plested, J., Gedeon, T.: Deep Feature Learning and Visualization for EEG Recording Using Autoencoders. Submited to International Conference on Neural Information Processing (ICONIP) 2018 12 (2018)
7. Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arXiv preprint arXiv:1611.07004 (2017)
8. Bashivan, P., Rish, I., Yeasin, M., Codella, N.: Learning representations from EEG with deep recurrent-convolutional neural networks. arXiv preprint arXiv:1511.06448 (2015)
9. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
10. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
11. Liu, M.-Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Advances in Neural Information Processing Systems, pp. 700–708 (2017)
12. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan. arXiv preprint arXiv:1701.07875 (2017)
13. Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv preprint arXiv:1703.10593 (2017)
14. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998). Wiley-IEEE Press, Indianapolis, Indiana
15. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images (2009)
16. Yu, L., Zhang, W., Wang, J., Yu, Y.: SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. arXiv preprint arXiv:1609.05473 (2016)
17. Guo, J., Lu, S., Cai, H., Zhang, W., Yu, Y., Wang, J.: Long Text Generation via Adversarial Training with Leaked Information. arXiv preprint arXiv:1709.08624 (2017)

18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
19. Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J.: Eegnet: a compact convolutional network for eeg-based brain-computer interfaces. arXiv preprint arXiv:1611.08024 (2016)
20. Li, Y., Dzirasa, K., Carin, L., Carlson, D.E.: Targeting EEG/LFP synchrony with neural nets. In: Advances in Neural Information Processing Systems, pp. 4623–4633 (2017)
21. Lucic, M., Kurach, K., Michalski, M., Gelly, S., Bousquet, O.: Are gans created equal? a large-scale study. arXiv preprint arXiv:1711.10337 (2017)
22. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)