Biomedical Signal Processing and Control 56 (2020) 101730

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Short communication

Decoding human brain activity with deep learning

Xiao Zheng, Wanzhong Chen*, Mingyang Li, Tao Zhang, Yang You, Yun Jiang

College of Communication Engineering, Jilin University, Changchun 130012, China

ARTICLE INFO

Article history:
Received 10 March 2019
Received in revised form 29 August 2019
Accepted 13 October 2019
Available online 28 October 2019

Keywords:
Brain-computer fusion
Deep learning
EEG
Brain decoding

ABSTRACT

Building a brain-computer fusion system that would integrate biological intelligence and machine intelligence became a research topic of great concern. Recent research has proved that human brain activity can be decoded from neurological data. Meanwhile, deep learning has become an effective way to solve practical problems.

Taking advantage of these trends, in this paper, we propose a novel method of decoding brain activity evoked by visual stimuli. To achieve this goal, we first introduce a combined long short-term memory—convolutional neural network (LSTM-CNN) architecture to extract the compact category-dependent representations of electroencephalograms (EEG). Our approach combines the ability of LSTM to extract sequential features and the capability of CNN to distil local features. Next, we employ an improved spectral normalization generative adversarial network (SNGAN) to conditionally generate images using the learned EEG features. We evaluate our approach in terms of the classification accuracy of EEG and the quality of the generated images.

The results show that the proposed LSTM-CNN algorithm that discriminates the object classes by using EEG can be more accurate than the existing methods. In qualitative and quantitative tests, the improved SNGAN performs better in the task of generating conditional images from the learned EEG representations; the produced images are realistic and highly resemble the original images.

Our method can reconstruct the content of visual stimuli according to the brain's response. Therefore, it helps to decode the human brain activity by using an image-EEG-image transformation.

1. Introduction

Recent developments in brain-computer interfaces have promise for the brain activity decoding [1]. Most studies focus on learning a latent space to classify the brain response, which is relatively easy to implement [2–5]. However, the research on understanding how the human brain works is more complex. Perhaps, we can create something more meaningful, illuminating, and complex by studying what happens in the brain. For example, we can study the brain activity caused by watching different images [6].

Decoding the human brain activity evoked by visual stimuli would have a great impact in brain-inspired computing and computer vision [7]. Many scientific communities have promoted research on analyzing how the human brain interacts with the outside world [8,9]. In some studies, deep learning was applied to decode and reconstruct the brain's visual activity [10]. Current research on reconstructing the images from the brain activity is mostly focused on functional magnetic resonance imaging (fMRI). For example, Shen et al. trained a deep neural network to reconstruct the stimulus image from the brain activity as captured by fMRI, finally, the generated images resembled the stimulus images [11,12]. Du et al. used a deep generative representation with Bayesian inference to combine an external stimulus with the human brain response via fMRI [13]. These methods depend on high sensitivity of fMRI. However, superiority of these methods is offset by the difficulty of manipulating fMRI scanners and their high cost.

To overcome these drawbacks, some researchers turned their attention to electroencephalograms (EEGs) because of their lower cost [14,15]. Additionally, EEG can provide a higher temporal resolution than fMRI. However, the EEG signal has a lower signal-to-noise ratio and lower spatial resolution, and the signal processing algorithm must achieve a higher accuracy. In 2017, PeRCeiVe Lab at the University of Central Florida developed an automated visual classifier by learning a brain activity manifold for specific visual categories with recurrent neural networks [16]. They have learned the visual category representations from EEG signals, but only obtained an average accuracy of 83%. In our opinion, it is

* Corresponding author at: College of Communication Engineering, Jilin University, Ren Min Street, 5988 Changchun, China.
E-mail addresses: zhengxiao1996@yeah.net (X. Zheng), chenwz@jlu.edu.cn (W. Chen).

https://doi.org/10.1016/j.bspc.2019.101730
1746-8094/© 2019 Elsevier Ltd. All rights reserved.

not sufficient and may affect sharpness and fidelity of generated images. Afterwards, they compared variational autoencoders and generative adversarial networks in the task of reconstructing the image [17]. Their promising results strongly demonstrate that visually relevant features extracted from EEG can effectively generate images semantically consistent with visual stimuli. However, we think that the quality of generated images can be improved.

At present, great challenges remain in decoding the human brain activity using EEG by reconstructing the visual stimuli. However, this task has a long-term research value. In our opinion, the success factors for a method of decoding and reconstructing the brain's visual responses are as follows:

1) the latent feature manifold extracted by the decoding algorithm must represent the category of visual stimuli as precisely as possible;
2) an excellent generative model must use the EEG feature manifold to learn stimuli-related image distribution;
3) a suitable evaluation method must be used to judge the trueness and sharpness of the generated images.

In this paper, we make three specific contributions:

1) we propose an LSTM-CNN architecture that consists of an LSTM network followed by a convolutional layer to extract visually related latent representations of EEG signals;
2) we employ an improved spectral normalization generative adversarial network (SNGAN) to conditionally generate images using the learned EEG features;
3) we compute the classification accuracy of the EEG latent representations and analyze the quality of the generated images. We show that our approach is superior to the existing methods.

2. Materials and methods

In this paper, we design a method to decode the brain activity evoked by visual stimuli. Our study is feasible for several reasons. First, the EEG signals are recorded when human subjects are shown images on a screen. The EEG signals convey visual information that can be used to recognize different images and understand their content. Second, EEG is a multichannel and time-domain signal with underlying noise components. It must be possible to extract low-dimensional and meaningful features to represent visual information for classification, and the features are assumed to represent the brain's visual activity. Finally, to generate images of the correct class, the EEG representations have to be distinguished among classes to some extent.

Based on the above ideas, our brain decoding architecture consists of two stages, shown in Fig. 1. The feasibility of the whole process was verified by experimental results.

The first stage—learning EEG features—aims at extracting low-dimensional EEG features from raw signals, which is implemented as a series of neural networks. Every EEG segment is associated with a specific image category, and the EEG features are available to identify image categories. The model is trained in a supervised manner by the object category of images, and a classifier for EEG features is attached at the end.

We have a hypothesis that the EEG signal evoked by each image can be classified correctly. The second stage—image generation—aims at producing images from the EEG features learned at the first stage. An improved SNGAN is trained to conditionally generate images. We also evaluate our approach in terms of the classification accuracy of EEG and the quality of the generated images.

2.1. EEG data acquisition and preprocessing

Our brain activity database[1] [16] consists of 128-channel EEG signals recorded from 6 participants (one female and five males), age 30–35 years, with uniform educational and cultural background; all participants received higher education. The EEG signals were recorded while the participants were shown the images as visual stimuli. The participants had been examined by a physician to exclude possible physical conditions that could affect normal brain activity. The acquisition process complied with the standards for research involving human subjects.

The visual stimuli dataset contains 2000 images of 40 different and easily discernible classes (50 images from each class) from ImageNet[2] [18]. The Tobii T60 EyeTracker was used as a monitor to display highly detailed stimulus materials through the built-in 24" HD widescreen display. The presentation software was compiled in. NET. During the recording session, each class of images was shown successively in 25 s as a batch (0.5 s per image), followed by a 10-second rest period when a black image was shown. The black image was used to "clear" the brain activity caused by watching the previous class of images. Hence, each experiment took 1400 s.

The EEG signals were collected by a 128-channel cap with active and low-noise electrodes[3] and by using four 32-channel BrainAmp DC high-precision signal amplifiers. All the EEG signals were referenced by the FCz electrode and were recorded by BrainVision[4] Recorder. The acquisition devices and software were made by BrainProducts (Germany). During the execution of the experiments, a professional technician was present to ensure that the electrode impedance is under 10 kOhm all the time. Besides, the BrainVision Analyzer 2.0 was used during the filtering process. The EEG signals were filtered by a notch filter (49–51 Hz) and a second-order band-pass filter (cut-off frequencies are 14 and 71 Hz) in runtime, which ensured that the EEG signals contained the most valuable frequency bands (beta and gamma) relevant for visual cognitive processes [19]. The frequency of sampling and data resolution were set to 1000 Hz and 16 bits, respectively.

During our experiment, the first 40 samples (0–40 ms) from each EEG sequence were discarded to minimize the disturbance from the previous stimulus image, and the following 440 samples (40–480 ms) were used. We divided the EEG dataset related to images into three parts: training set, validation set, and test set, which were divided as 80% (1600 images), 10% (200 images), and 10% (200 images), respectively. For all 6 participants, we ensured that the EEG signal recorded for the same image was assigned to the same set.

2.2. EEG visual representation learning

In this part, we intended to extract a compact visually related representation of a temporal EEG sequence at multiple channels. Many previous studies [20,21] concentrated on time-frequency domain features or statistical characteristics, ignoring the local effect and temporal sequence dynamics; therefore, only basic information on brain activity was collected. To include the local effect and sequence dynamics in our experiments, we used long short-term memory (LSTM) and a convolutional neural network (CNN) to extract sequential and local features, respectively.

The upper part of Fig. 1 shows the architecture of our EEG manifold representation model. The multichannel EEG signals are regarded as input to the *Encoder* module. It outputs a feature vector

[1] http://perceive.dieei.unict.it/index-dataset.php?name=EEG_Data
[2] http://www.image-net.org/challenges/LSVRC/2012/nonpub-downloads
[3] http://brainproducts.com/
[4] http://brainvision.com/

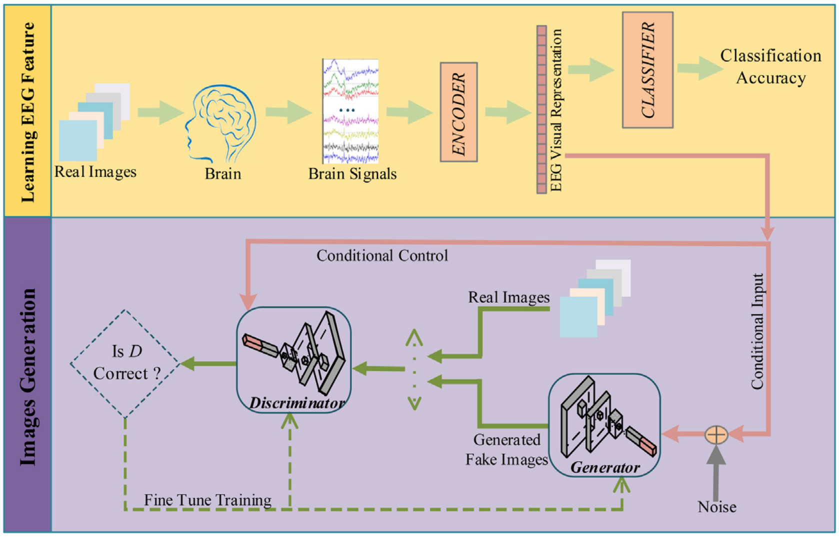


Fig. 1. Our architecture design for image generation based on EEG.

as a low-dimensional representation that summarizes the relevant attribute of the input. Theoretically, if an input EEG segment was recorded when an image was shown to the subject, the meaningful output vector would encode the brain activity at that moment. The *Encoder* module is trained in advance, and we learn the model's parameters by gradient descent through end-to-end training. In our experiments, we used several schemes for the *Encoder* module, described as follows.

1) Only LSTM (Fig. 2a): the preprocessed EEG signals are provided as an input to a LSTM layer, into which all 128 EEG channels are simultaneously fed. At each timestamp, the LSTM input size is 128, which corresponds to 128 channels of EEG data. The dimension of the hidden layer state is also set to 128. In addition, because there are 440 samples in an EEG segment, the LSTM cells are repeatedly calculated 440 times. The activation function is a rectified linear unit (ReLU) function. A fully connected linear layer (the output dimension is 40) and a softmax classifier are appended as the output layer. The output of the fully connected linear layer is considered as the EEG feature manifold for the input sequence.
2) CNN + LSTM (Fig. 2b): the encoder network includes a convolutional layer, an LSTM network, and an output layer. We try to simultaneously use the ability of LSTM to extract sequential features and the capability of CNN to distil local features. The convolutional layer is a one-dimensional CNN, whose input is 128-channel EEG segment and the output is 64-channel. The kernel size is set to 3, and the padding is true. The LSTM layer is similar to the previous architecture. The same fully connected linear layer and a softmax classifier are appended as the output layer. We call this architecture CNN-LSTM. The output of the fully connected linear layer is considered as the EEG feature manifold for the input sequence.
3) LSTM + CNN (Fig. 2c): this architecture is similar to CNN-LSTM, but the order of layers is altered to study how it affects the performance. The output length of the LSTM is equal to the number of the LSTM hidden layer state. The dimension of the LSTM hidden layer state is set to 128, so the shape of the output of LSTM is [*B**1*128] (*B* is the batch size), which is regarded the input of the first convolutional layer. We call this model LSTM-CNN. The EEG vector representation is the same as previously.

Ideally, the resulting output should be a compact EEG visual representation that can discriminate brain activity. Both the *Encoder* and softmax classifier are trained using gradient descent by associating the related class label to the input signal.

2.3. Image generation using the EEG visual representation

As the lower part of Fig. 1 depicts, we develop and train our *Generator* model based on spectral normalization generative adversarial networks (SNGANs) [22]. Compared with regular generative adversarial networks, SNGANs import a novel spectral normalization technique to stabilize the *Discriminator*. The innovation of this method is that the Lipschitz constant can be controlled by constraining the spectral norm on each layer (convolutional layers, linear layer, and so on) of the discriminator.

For the *n*th layer of the network, the relationship between its input x_{n-1} and output x_n can be expressed as:

$$x_n = a_n(W_n x_{n-1} + b_n) \tag{1}$$

Where $a_n(\cdot)$ is the nonlinear activation function of the *n*th layer (ReLU is adopted), W_n is the parameter matrix, and b_n is the bias. For convenience, *b* is omitted, and Eq. 1 can be written as:

$$x_n = C_n W_n x_{n-1} \tag{2}$$

Where C_n is a diagonal matrix, which is used to represent the effect of ReLU. When the corresponding input is negative, the diagonal element is 0. And when the corresponding input is positive, the diagonal element is 1. Thus, in the case of multi-layer neural network (assuming it has N layers), the relationship of input *x* and output *f(x)* can be written as follows:

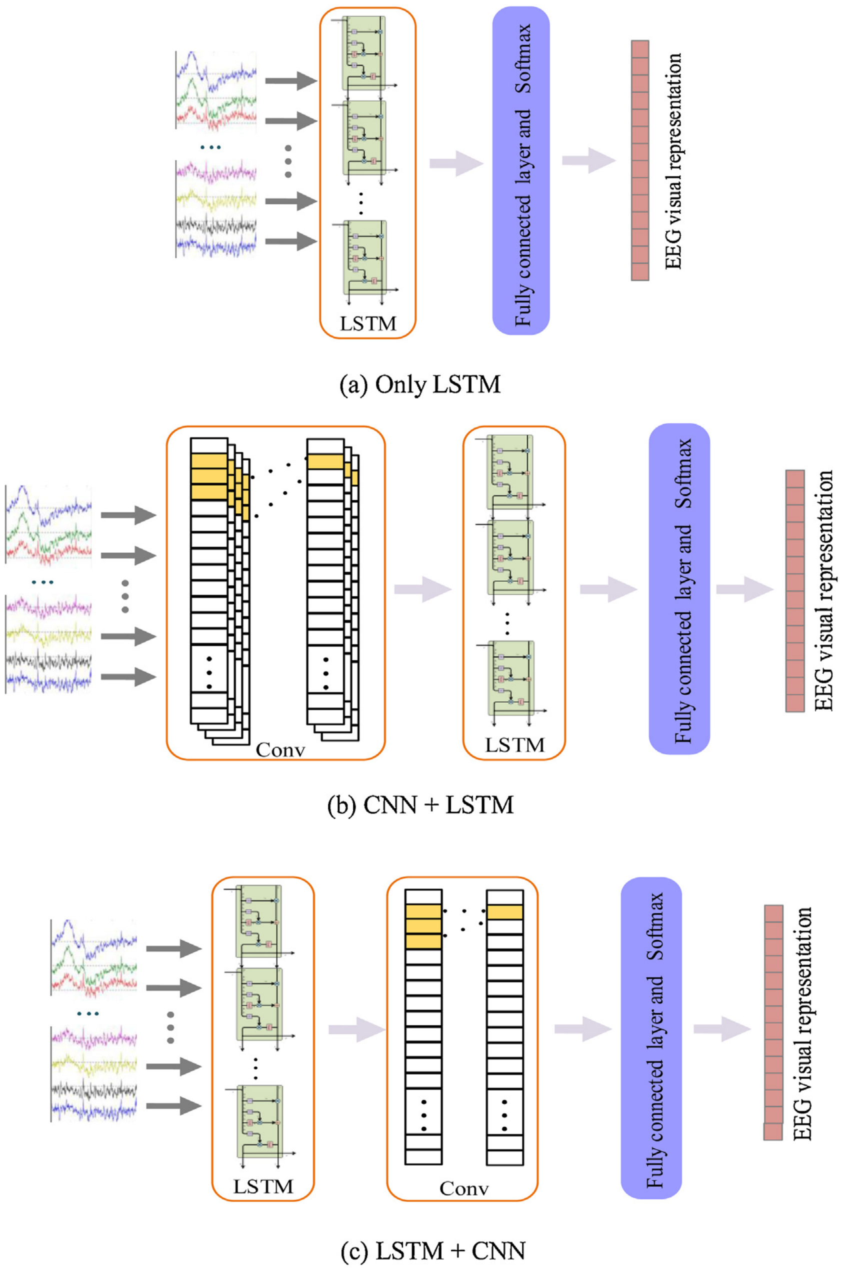
$$f(x) = C_N W_N \ldots C_1 W_1 x \tag{3}$$

Lipschitz constraint requires the gradient of *f(x)* to satisfy the following condition:

$$\left|\left|\nabla_x(f(x))\right|\right|_2 = \left|\left|C_N W_N \ldots C_1 W_1\right|\right|_2 \leq \left|\left|C_N\right|\right|_2 \left|\left|W_N\right|\right|_2 \ldots \left|\left|C_1\right|\right|_2 \left|\left|W_1\right|\right|_2 \tag{4}$$

Where the nabla sign represents differential operator, and $||W||_2$ represents the spectral norm of matrix *W*. It is defined as follows:

$$\sigma(W) := \sup \frac{\left|\left|Wy\right|\right|_2}{||y||_2} = \sup \frac{\left|\left|Wy\right|\right|_2}{||y||_2 \leq 1} \tag{5}$$



(a) Only LSTM

(b) CNN + LSTM

(c) LSTM + CNN

Fig. 2. Three schemes of Encoder module in our experiments.

Which requires $Wy = my$ (if W is a p-order diagonal matrix and m is a constant number, y is a p-dimension column vector). $\sigma(W)$ is the maximum singular value of the matrix W. For the diagonal matrix C, $\sigma(C)$ is the largest element (It is 1) in the diagonal. Thus, Eq. 4 can be expressed as:

$$\left|\left|\nabla_X(f(x))\right|\right|_2 \leq \prod_{i=1}^{N} \sigma(W_i) \tag{6}$$

The upper limit of spectral norm of diagonal matrix (corresponding to ReLU) is 1. In order to make $f(x)$ satisfies Lipschitz constraint, Eq. 4 can be normalized as:

$$\left|\left|\nabla_x(f(x))\right|\right|_2 = \left|\left|C_N \frac{W_N}{\sigma(W_N)} \ldots C_1 \frac{W_1}{\sigma(W_1)}\right|\right|_2 \leq \prod_{i=1}^{N} \frac{W_i}{\sigma(W_i)} = 1 \tag{7}$$

Therefore, the Lipschitz = 1 constraint can be satisfied by dividing the network parameters by the spectral norm of the parameter matrix of this layer.

It has been proved that SNGANs have more stable performance during the training process. The objective loss functions for SNGANs are as follows:

$$\min_D loss_D = -E_{x\sim p_{data}(x)}[\log(D(x))] - E_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{8}$$

$$\min_G loss_D = -E_{z\sim p_z(z)}[\log(D(G(z)))] \tag{9}$$

In the formulas above, a *Generator* model *G()* uses random input from a noise distribution $p_z(z)$, and *G(z)* represents the process of image generating, whose result is an RGB image, whose result is an RGB image. Next, a *Discriminator* model *D()* evaluates the probability of whether its input *x* (real data or fake data) belongs to the target distribution. *D(x)* represents the discriminant result (0 or 1)

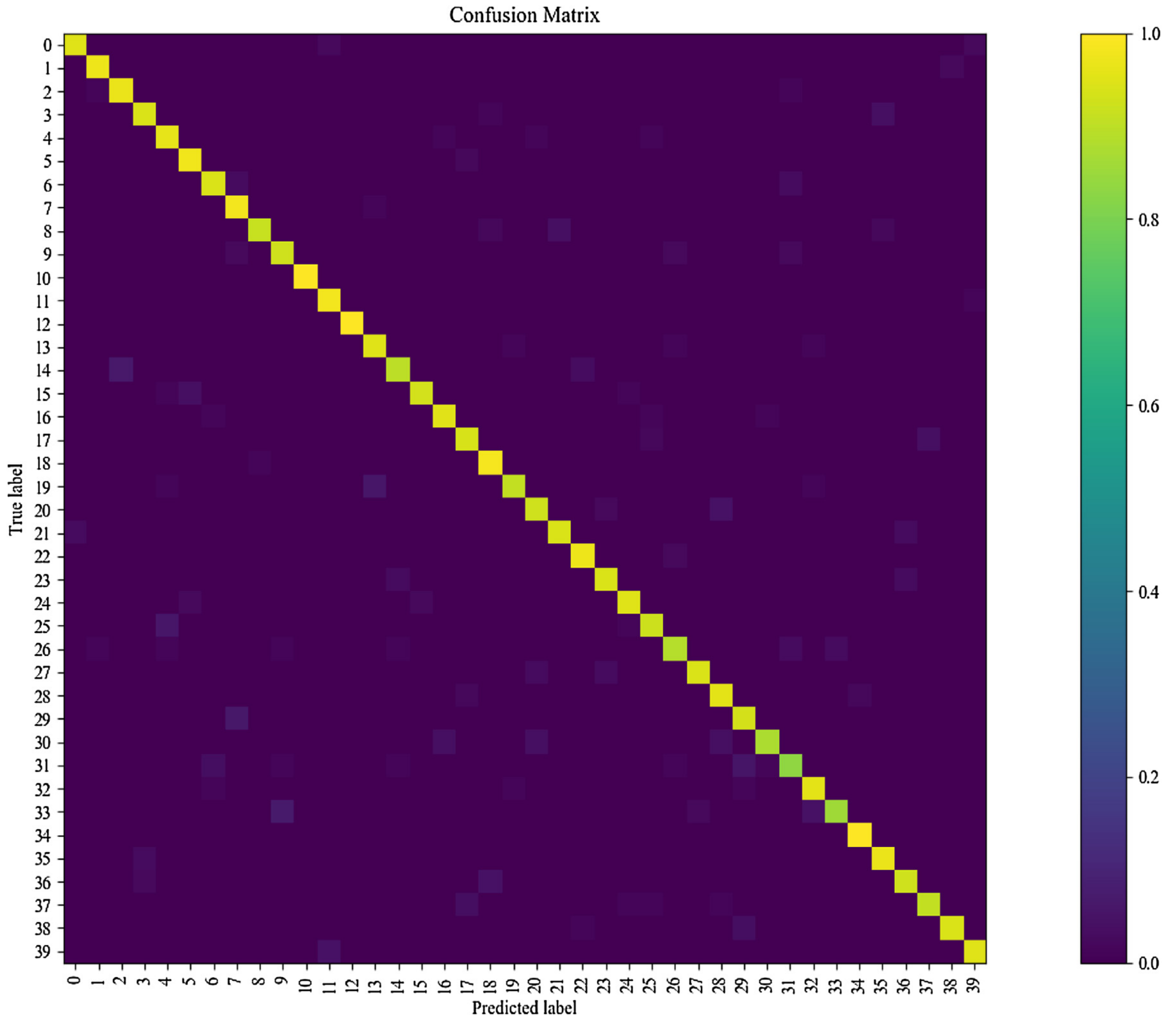


Fig. 3. The confusion matrixes of our LSTM-CNN model.

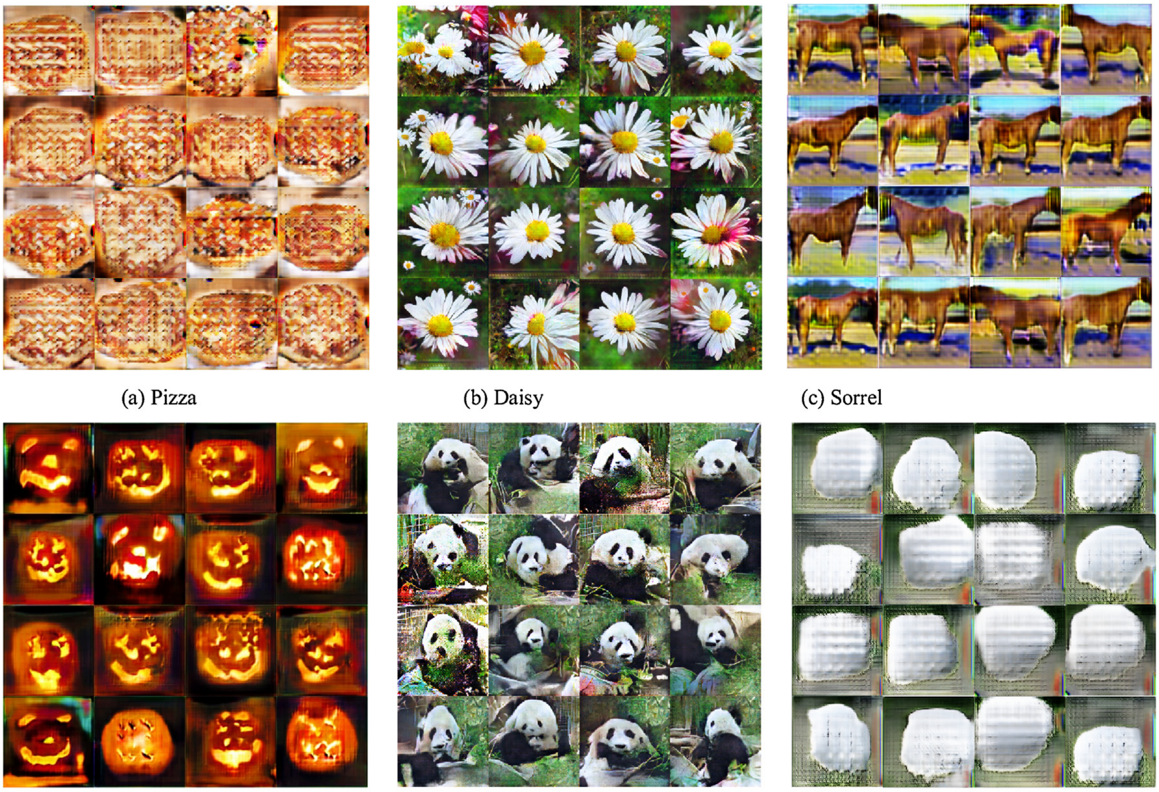
of the real image, and $D(G(z))$ represents the discriminant result of the generated image. Besides, E represents expectation operator.

In our method, we incorporate a 128-dimensional conditional vector y (which is related to the output image content) into the discriminator of SNGANs. Conditional vector y is the average EEG visual feature vector over all the images of each class, computed by the *Encoder* module as described in Section 2. With this modification, learning EEG features and image generation are linked. It allows our generator G and discriminator D to capture image features of different classes. Our generator G generates images from EEG feature vectors. By distinguishing the real images from the generated images, discriminator D compels the generator G to generate high-quality pictures that activate the corresponding activity in the human brain. Suppose the *Encoder* has been trained correctly to compute discriminable features for different image contents. Generator G and Discriminator D, two kinds of convolutional networks, will be able to capture this separability and perform accordingly. Their architecture is shown in Table 1 and is inspired by [23].

Conditional vector y appended to a 100-dimensional random noise vector z is considered as the input of our generator. The purpose of adding noise is to ensure the diversity of generated images. Next, a cluster of ResBlocks [24] upsample the input to a 3-D color image as the output. The input of our discriminator is an image of the same size, either a real or a generated image. After being conveyed to a cascade of spectral normalization ResBlocks to implement the downsampling, the size of the feature maps gradually decreases. We embed the intermediate output with conditional vector y as an inner product. A linear layer output added to the inner product is regarded as the discriminant result. Therefore, the objective loss functions for our improved SNGANs are as follows:

Table 1
Our architecture for image generation, y is a 128-dimensional conditional vector, z is a 100-dimensional stochastic noise vector, and h is the output of the previous layer.

(a) Generator
y + z $\in R^{228}$
Dense, 4*4*1024
ResBlock up 1024
ResBlock up 512
ResBlock up 256
ResBlock up 128
ResBlock up 64
BN, ReLU, 3*3 conv 3
Tanh
(b) Discriminator
RGB image $x \in R^{228*128*3}$
ResBlock down 64
ResBlock down 128
ResBlock down 256
Concat (Embed (y), h)
ResBlock down 512
ResBlock down 1024
ResBlock 1024
ReLU
Global sum pooling
Dense → 1



(a) Pizza (b) Daisy (c) Sorrel

(d) Jack-o'-lantern (e) Panda (f) Golf ball

Fig. 4. Some generation samples.

$$\min_{D} loss_D = -E_{x\sim p_{data}(x)}[\log(D(x|y))] - E_{z\sim p_z(z)}[\log(D(G(z|y)|y))] \quad (10)$$

$$\min_{G} loss_D = -E_{z\sim p_z(z)}[\log(D(G(z|y)|y))] \quad (11)$$

In the formulas, E represents expectation operator, *D(x|y)* represents the discriminant result (0 or 1) of the real image under the constraint of the conditional vector *y*, *G(z|y)* represents the process of generating an image using noise *z* and conditional vector *y*, whose result is an RGB image, and *D(G(z|y)|y)* represents the discriminant result of the generated image under the constraint of the conditional vector *y*.

Obviously, it is notoriously difficult to balance the Generator and Discriminator when training GANs. In our experiment, there are few image samples used for the EEG acquisition, which makes it harder to train the model directly. Therefore, we design a scheme to fully use the selected images and their classes. We train our conditional SNGAN in two phases. In the first phase, we use images from ImageNet with no available EEG signals to train the SNGAN model. All conditional vectors y are set to zero vectors. After 1000 epochs, we retrain the model on the images used for EEG recording for 500 more epochs. At this moment, we provide a proper conditional vector y. Finally, we generate image samples of 128 × 128 pixels.

3. Results and discussion

This section examines the feasibility of our method for decoding the human brain activity based on EEG and deep learning. We provide the results and discussion from two aspects: 1) we analyze the performance of the *Encoder* module to evaluate how it learns to extract the latent EEG visual representation; 2) to test our improved SNGAN model, we assess the quality of EEG-driven image by two methods.

Table 2
Maximum validation accuracy (marked "Max VA") and associated test accuracy(marked "TA at max VA") for three different configurations shown in Sect. 2.2.

Model	Max VA	TA at max VA
Only LSTM	88.9%	86.8%
CNN-LSTM	88.5%	84.7%
LSTM-CNN	**95.0%**	**94.4%**

3.1. EEG representation vector: classification accuracy

Table 2 displays the achieved classification accuracy by three encoder structures. In the training stage, the learning rate is initially set to 0.001, and the *Encoder* is trained using the Adam gradient descent, with the batch size of 16. The state parameters of the model and hyperparameters of training were fine-tuned on the validation set.

Accuracy of our CNN-LSTM model is 2.1% lower than for our LSTM model. Meanwhile, our LSTM-CNN model achieves an accuracy of 7.6% higher than LSTM. The confusion matrix of our LSTM-CNN model is provided in Fig. 3. As seen, a satisfactory classification result can be reached in most of the categories. These experimental results seem to confirm that our idea was right. By combining CNN and LSTM, we are able to simultaneously harness the ability of LSTM to extract sequential features and the capability of CNN to recognize local features. However, it is necessary to point out that the order of layers plays a vital role in performance of our models.

It seems that adding a convolutional layer on the front may destroy many sequential features, and the following LSTM layer will underperform. Besides, the reason why the LSTM-CNN model performs best is that its initial LSTM layer not only harnesses the original information, but captures the context. Afterwards, the con-

Table 3
Comparison with the existing works.

Model	Accuracy	Class
SVM in [26]	82.7%	2
EEG-Net in [27]	88.0%	2
PMK in [25]	91.7%	3
RNN-based model in [29]	84.0%	40
BiLSTM + ICA + SVM in [28]	97.1%	40
Our LSTM-CNN model	94.4%	40

volutional layer will find local features using this representation of the original input, resulting in a better accuracy.

To demonstrate the effectiveness of our LSTM-CNN model, the proposed model is evaluated and compared with several existing frameworks of EEG-based visual object classification in Table 3. As seen, our method outperforms methods [25,26], and [27] in terms of the classification accuracy and categories. Note that [16] and [28] used the same dataset as this study; they used all 128 channels for the experiments and obtained the accuracy of 84.0% and 97.1%. We think that the results of [28] cannot be taken as a baseline. That study employed independent component analysis (ICA) with support vector machines (SVM) to research the EEG representation, and their method processed not the complete object-evoked EEG signals but only feature vectors. Besides, traditional machine learning classifiers have a high computational complexity, whereas the softmax classifier that is used as an output layer in our end-to-end model requires little computational time. Overall, our LSTM-CNN model has a higher analytical capability for EEG signals.

3.2. *Generated image quality*

Fig. 4 shows images of several classes generated in our experiments. We conclude that our generator can capture the distinguishing patterns. This finding confirms that the discriminator and generator are able to take advantage of the latent EEG visual features. It can be observed that for some classes (such as *Daisy* and *Pizza*) the samples are more realistic than for other classes (such as *Panda* and *Golf ball*). We assume that it depends on the complexity of the dataset. For example, the *Panda* class (n02510455) contains different postures and scenes in the real images, and it is more difficult to extract the common patterns. On the contrary, it is simple for the *Daisy* class.

To demonstrate the superiority of our method, we use the same evaluation method as in the existing papers. We generate 10,000 images per class and compute the Inception score (IS), which is suggested as an appropriate measure to evaluate the visual quality of generated images. Table 3 shows the results; it also lists the results from an existing paper [29] for a comparison. Our method achieves the average Inception score of 5.53. To the best of our knowledge, the Inception score results on ImageNet have not been published. On CIFAR-10 dataset, the best result that has been published so far is 8.22, and on STL-10 dataset, the best Inception score is 9.10 [22]. The results obtained for our dataset suggest the capability of the network to analyze the characteristics of certain classes in respect to others. Our results are still relatively low compared with CIFAR-10 and STL-10, and it should be concluded that several reasons may influence our results:

1) a higher resolution increases the difficulty of the problem: 128 × 128 in our case, 96 × 96 in STL-10, and 32 × 32 in CIFAR-10;
2) we used more classes in our study: 40 in our dataset, 10 in STL-10, and 10 in CIFAR-10;
3) we used fewer images per class: 1200 images in our case, 1300 in STL-10, and 6000 in CIFAR-10.

Table 4
Inception scores (IS) and Inception classification accuracy (IC) for each class of generated images.

Class	IS		IC	
	[29]	Our method	[29]	Our method
German shepherd (n02106662)	4.91	**5.26**	0.23	**0.31**
Egyptian cat (n02124075)	4.45	**5.11**	0.29	**0.34**
Lycaenid butterfly (n02281787)	5.03	**5.51**	0.37	**0.40**
Sorrel (n02389026)	5.86	**6.03**	0.62	**0.74**
Capuchin (n02492035)	**4.99**	4.95	0.41	**0.45**
Elephant (n02504458)	5.35	**5.41**	0.57	**0.59**
Panda (n02510455)	6.35	**7.02**	**0.72**	0.70
Anemone fish (n02607072)	6.11	**6.37**	0.81	**0.85**
Airliner (n02690373)	**6.20**	5.83	**0.86**	0.85
Broom (n02906734)	4.76	**5.01**	0.35	**0.38**
Canoe (n02951358)	4.59	**4.72**	0.24	**0.31**
Cellphone (n02992529)	5.17	**5.62**	0.31	**0.38**
Mug (n03063599)	4.62	**5.47**	0.23	**0.25**
Convertible (n03100240)	4.54	**4.55**	**0.34**	0.29
Desktop PC (n03180011)	5.81	**6.24**	**0.61**	0.60
Digital watch (n03197337)	4.54	**5.17**	0.51	**0.54**
Electric guitar (n03272010)	4.91	**5.23**	0.32	**0.60**
Electric locomotive (n03272562)	4.88	**5.06**	0.24	**0.27**
Espresso maker (n03297495)	**5.33**	4.99	**0.32**	0.28
Folding chair (n03376595)	4.88	**5.26**	0.27	**0.36**
Golf ball (n03445777)	5.06	**6.02**	0.28	**0.41**
Piano (n03452741)	4.47	**5.02**	0.22	**0.28**
Iron (n03584829)	4.32	**4.38**	0.23	**0.39**
Jack-o'-lantern (n03590841)	6.64	**7.07**	**0.91**	0.87
Mailbag (n03709823)	5.51	**5.49**	**0.49**	0.44
Missile (n03773504)	5.87	**6.32**	0.54	**0.57**
Mitten (n03775071)	5.10	**5.39**	0.36	**0.51**
Mountain bike (n03792782)	4.86	**6.05**	0.33	**0.40**
Mountain tent (n03792972)	4.70	**5.36**	0.30	**0.47**
Pyjama (n03877472)	4.21	**4.66**	0.20	**0.33**
Parachute (n03888257)	4.59	**4.87**	0.38	**0.49**
Pool table (n03982430)	**4.68**	4.56	0.35	**0.43**
Radio telescope (n04044716)	5.08	**5.23**	0.37	**0.46**
Reflex camera (n04069434)	4.64	**5.05**	0.29	**0.31**
Revolver (n04086273)	4.55	**4.71**	0.26	**0.33**
Running shoe (n04120489)	4.31	**4.67**	0.22	**0.31**
Banana (n07753592)	6.28	**6.56**	0.83	**0.87**
Pizza (n07873807)	5.87	**7.24**	**0.79**	**0.79**
Daisy (n11939491)	5.81	**7.28**	0.74	**0.81**
Bolete (n13054560)	5.37	**6.29**	0.60	**0.66**
All	*5.07*	***5.53***	*0.43*	***0.49***
P-value of Mann-Whitney U test	***0.006***		***0.032***	

Furthermore, for the purpose of confirming whether the images generated upon a given condition can show similar information as the corresponding real images, we mix 1000 generated images per class together to classify and compute the classification accuracy through the Inception network (IC). As shown in Table 3, per-class Inception classification accuracy is computed, and the average Inception classification accuracy of 0.49 is higher than in the existing paper. It should be noted that the probability of random guess in our case is 2.5%. Hence, to a certain extent, it confirms that the generated images are realistic.

Besides, Mann—Whitney U test was employed to evaluate whether a significant difference exists between our results and the results in [29]. The computed p-values of IS and IC are 0.006 and 0.032, respectively, which are lower than 0.05. Thus, our method significantly outperforms the method in [29].

These data suggest that human brain activity can be successfully analyzed, and our approach can accelerate the development of an automatic visual classifier of EEG signals. However, the results for some classes are worse than for other classes. We may explain it by observing the image dataset: the lower performance is related to classes with a higher intra-class variance (Table 4).

Although our results are still relatively low compared with other areas of image processing, we show that the generated images can help in automatic classification of EEG. It should be concluded that several reasons may influence our results:

1) The number of images used to record EEG signals is not enough.
2) Because of the restrictions on experimental conditions, the network models are not trained well enough.
3) Our conditional vectors are noisy. We use the EEG features instead of image class labels, unlike in traditional supervised learning.

4. Conclusion

In this paper, we propose a method of reading the mind on the visual level based on deep learning. It consists of two phases: 1) an LSTM-CNN model is designed to extract the EEG visual representation; 2) using the learned EEG features, an improved SNGAN network is used to conditionally generate images that depict the same visual categories as the stimuli. Our results find that the proposed LSTM-CNN algorithm is able to reach a competitive performance in discriminating the object classes using EEG. In qualitative and quantitative tests, the improved SNGAN can generate images according to the EEG signals evoked by visual stimuli.

Therefore, our method can reconstruct the contents of a visual stimulus according to the brain's response. The method can achieve an image-EEG-image transformation. Our promising results demonstrate that the human brain activity in visual recognition can be decoded and applied to automated visual classification. Decoding and reconstructing the brain activity can be considered as a critical step toward the research in machine learning, computer vision, and brain-inspired computing. In the future, we plan to develop more effective deep learning methods to differentiate EEG signals evoked by images of different categories. Additionally, we plan to attempt to reconstruct the original image, not to generate an image describing the same visual category as the stimuli.

Declaration of Competing Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

We sincerely appreciate the editors and reviewers for their valuable suggestions and questions. This work was supported by the Science and technology development project of Jilin province, China [grant numbers 20190302034GX]; the Fundamental Research Funds for the Central Universities of China [grant numbers 451170301193].

References

[1] O.R. Pinheiro, L.R.G. Alves, J.R.D. Souza, EEG signals classification: motor imagery for driving an intelligent wheelchair, IEEE. Lat. Am. Trans. 16 (1) (2018) 254–259, http://dx.doi.org/10.1109/TLA.2018.8291481.

[2] L. Wang, M.A. Ronald, P.V.D. Johannes, B.A.M.A. Johan, A broadband method of quantifying phase synchronization for discriminating seizure EEG signals, Biomed. Signal Process. Control 52 (2019) 371–383, http://dx.doi.org/10.1016/j.bspc.2018.10.019.

[3] X. Songyun, L. Chang, W. You, Z. Juanli, D. Xu, A hybrid BCI (brain-computer interface) based on multi-mode EEG for words typing and mouse control, J. X. Univ. Technol. 34 (2) (2016) 245–249.

[4] G.R. Muller-Putz, G. Pfurtscheller, Control of an electrical prosthesis with an ssvep-based BCI, IEEE Trans. Biomed. Eng. 55 (1) (2008) 361–364, http://dx.doi.org/10.1109/TBME.2007.897815.

[5] T. Zhang, W. Chen, LMD based features for the automatic seizure detection of EEG signals using SVM, IEEE Trans. Neural Syst. Rehabil. Eng. 25 (8) (2017) 1100–1108, http://dx.doi.org/10.1109/TNSRE.2016.2611601.

[6] M.V. Peelen, P.E. Downing, The neural basis of visual body perception, Nat. Rev. Neurosci. 8 (8) (2007) 636–648, http://dx.doi.org/10.1038/nrn2195.

[7] H.P. Op de Beeck, K. Torfs, J. Wagemans, Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway, J. Neurosci. 28 (40) (2008) 10111–10123, http://dx.doi.org/10.1523/JNEUROSCI.2511-08.2008.

[8] B.N. Pasley, S.V. David, N. Mesgarani, A. Flinker, S.A. Shamma, N.E. Crone, R.T. Knight, E.F. Chang, Reconstructing speech from human auditory cortex, PLoS Biol. 10 (1) (2012), e1001251, http://dx.doi.org/10.1371/journal.pbio.1001251.

[9] Z. Kourtzi, N. Kanwisher, Cortical regions involved in perceiving object shape, J. Neurosci. 20 (9) (2000) 3310–3318, http://dx.doi.org/10.1016/S1058-2746(96) 80198-4.

[10] Y. Yuan, G. Xun, Q. Suo, K. Jia, A. Zhang, Wave2Vec: deep representation learning for clinical temporal data, Neurocomputing 324 (2019) 31–42, http://dx.doi.org/10.1016/j.neucom.2018.03.074.

[11] T. Horikawa, Y. Kamitani, Generic decoding of seen and imagined objects using hierarchical visual features, Nat. Commun. 8 (2015) 15037, http://dx.doi.org/10.1038/ncomms15037.

[12] G. Shen, T. Horikawa, K. Majima, Y. Kamitani, Deep image reconstruction from human brain activity, bioRxiv (2018) 240317, http://dx.doi.org/10.1101/240317.

[13] C. Du, H. He, Sharing deep generative representation for perceived image reconstruction from human brain activity, Proceedings of the International Joint Conference on Neural Networks (2017) 1049–1056, http://dx.doi.org/10.1109/IJCNN.2017.7965968.

[14] M. Nakanishi, Y. Wang, Y.T. Wang, Y. Mitsukura, T.-P. Jung, A high-speed brain speller using steady-state visual evoked potentials, Int. J. Neural Syst. 24 (06) (2014), 1450019, http://dx.doi.org/10.1142/S0129065714500191.

[15] H. Cecotti, A. Graser, Convolutional neural networks for p300 detection with application to brain-computer interfaces, IEEE Trans. Pattern Anal. Mach. Intell. 33 (3) (2011) 433–445, http://dx.doi.org/10.1109/TPAMI.2010.125.

[16] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, M. Shah, Deep learning human mind for automated visual classification, Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition (2017) 4503–4511, http://dx.doi.org/10.1109/CVPR.2017.479.

[17] I. Kavasidis, S. Palazzo, C. Spampinato, D. Giordano, Brain2Image: converting brain signals into images, Proceedings of the 2017 ACM Multimedia Conference (2017) 1809–1817, http://dx.doi.org/10.1145/3123266.3127907.

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet large scale visual recognition challenge, Int. J. Comput. Vision 115 (3) (2015) 211–252, http://dx.doi.org/10.1007/s11263-015-0816-y.

[19] T.A. Pedley, Electroencephalography: Basic Principles, Clinical Applications, and Related Fields, third ed., Williams and Wilkins, Baltimore, 1993.

[20] B. Kaneshiro, M.P. Guimaraes, H.S. Kim, A.M. Norcia, P. Suppes, A representational similarity analysis of the dynamics of object processing using single-trial EEG classification, PLoS One 10 (8) (2015), e0135697, http://dx.doi.org/10.1371/journal.pone.0135697.

[21] A.X. Stewart, A. Nuthmann, G. Sanguinetti, Single-trial classification of EEG in a visual object task using ICA and machine learning, J. Neurosci. Methods 228 (10) (2014) 1–14, http://dx.doi.org/10.1016/j.jneumeth.2014.02.014.

[22] T. Miyato, T. Kataoka, M. Koyama, Y. Yoshida, Spectral normalization for generative adversarial networks, International Conference on Learning Representations (2018).

[23] T. Miyato, M. Koyama, CGANs with projection discriminator, International Conference on Learning Representations (2018).

[24] K. He, X. Zhang, S. Ren, Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016) 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.

[25] A. Kapoor, P. Shenoy, D. Tan, Combining Brain Computer Interfaces With Vision for Object Categorization, Proc, CVPR, Anchorage, Alaska, USA, 2008.

[26] R. El-Lone, M. Hassan, A. Kabbara, R. Hleiss, Visual Objects Categorization Using Dense EEG: a Preliminary Study, Proc. ICABME, Beirut, Lebanon, 2015, pp. 112–118.

[27] V. Parekh, R. Subramanian, D. Roy, C.V. Jawahar, An EEG-based Image Annotation System, Proc. NCVPRIPG, Mandi, India, 2017, pp. 303–313.

[28] A. Fares, S. Zhong, J. Jiang, Region Level Bi-directional Deep Learning Framework for EEG-based Image Classification, Proc. BIBM, Madrid, Spain, 2018, pp. 368–373.

[29] S. Palazzo, C. Spampinato, I. Kavasidis, D. Giordano, Generative adversarial networks conditioned by brain signals, Proceedings of the IEEE International Conference on Computer Vision (2017) 3430–3438, http://dx.doi.org/10.1109/ICCV.2017.369.