Siamese Reconstruction Network: Accurate Image Reconstruction from Human Brain Activity by Learning to Compare

Lingyun Jiang, Kai Qiao, Linyuan Wang, Chi Zhang, Jian Chen, Lei Zeng, Haibing Bu and Bin Yan *

PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China; YunLord@outlook.com (L.J.); qiaokai1992@gmail.com (K.Q.); wanglinyuanwly@163.com (L.W.); zcboluo@163.com (C.Z.); kronhugo@163.com (J.C.); lei_zeng@sina.com (L.Z.); bhblzh@263.net (H.B.)

* Correspondence: ybspace@hotmail.com

Received: 10 September 2019; Accepted: 2 November 2019; Published: 7 November 2019

Featured Application: This study can be used in brain decoding, to calculate the brain perceived visual information.

Abstract: Decoding human brain activities, especially reconstructing human visual stimuli via functional magnetic resonance imaging (fMRI), has gained increasing attention in recent years. However, the high dimensionality and small quantity of fMRI data impose restrictions on satisfactory reconstruction, especially for the reconstruction method with deep learning requiring huge amounts of labelled samples. When compared with the deep learning method, humans can recognize a new image because our human visual system is naturally capable of extracting features from any object and comparing them. Inspired by this visual mechanism, we introduced the mechanism of comparison into deep learning method to realize better visual reconstruction by making full use of each sample and the relationship of the sample pair by learning to compare. In this way, we proposed a Siamese reconstruction network (SRN) method. By using the SRN, we improved upon the satisfying results on two fMRI recording datasets, providing 72.5% accuracy on the digit dataset and 44.6% accuracy on the character dataset. Essentially, this manner can increase the training data about from n samples to 2n sample pairs, which takes full advantage of the limited quantity of training samples. The SRN learns to converge sample pairs of the same class or disperse sample pairs of different class in feature space.

Keywords: brain decoding; visual reconstruction; functional magnetic resonance imaging; Siamese neural network; deep learning; learning to compare

1. Introduction

Scientists have always been fascinated in how the human brain works, so they have carried out studies to understand the workings of the brain [1]. An influential predictive coding hypothesis states that brain activity can be decoded to reconstruct and classify what a person has sawn [2–4]. Human brain decoding [5–7] plays an important role in brain-machine interfaces, for example helping disabled persons in expressing and motioning and promoting the developing of the brain mechanism. Since the development of functional magnetic resonance imaging (fMRI), it has become an effective tool in understanding brain activity, for example used in Brain decoding. Brain decoding can be divided into three aspects: classification, identification, and reconstruction [8]. It has made some progresses in this area, but most previous researches focused on predicting the category of image stimulus [9,10] or identifying the image stimuli from a candidate set [11]. The one important and difficult field in
neuroscience is the reconstruction of image stimuli to read out, or decode, mental content from brain activity. However, two problems restrict the accuracy of visual reconstruction. One is the confusing fMRI measurement noise and the other is the high dimensionality and limited amount of fMRI data. These two problems lead to unsatisfactory reconstruction results and even wrong categories. The first problem is caused by physical and acquisition equipment, and it can only be solved by a significant breakthrough of hardware development. Therefore, we are committed to solve the second problem.

In brain decoding, few studies have reported on perceptual image reconstruction. The current reconstruction methods mainly focus on the innovation of the method, rather than the use of limited instances. Thirion et al. [12] reconstructed simple images by an explicit inverse reconstruction approach for the same subject. Miyawaki et al. [13] achieved the reconstruction of simple visual images by combining local image bases of multiple scales. Van et al. [14] reconstructed handwritten digits of 6 and 9 from the fMRI data by employing a hierarchical generative model to learn the hierarchy of features [15]. Schoenmakers et al. [16] reconstructed handwritten letters of “BRAINS” from fMRI data by a straightforward linear Gaussian approach using the linear gauss model relying on the inversion of properly regularized encoding models. Yargholi et al. [17] reconstructed handwritten digits 6 and 9 by using Bayesian networks. Fujiwara et al. [18] proposed a model that was based on probabilistic extension of the CCA model [19] to establish mappings between the stimulus and the brain. Wang et al. [20] have achieved the good results on many tasks by the deep canonically correlated autoencoders (DCCAE) for correlation-based representation learning. It has been proven that CNN can be applied for visual reconstruction, because the features of some layers in CNN are strongly correlated with the brain activities of the visual cortex in recent studies [21–23]. Wen et al. [1] proposed encoding and decoding models based on multivariate linear regression and deconvolutional neural network [24] to describe the bi-directional relationships between the CNN and the brain and reconstruct the dynamic video frame by frame. Du et al. [25] has improved the reconstruction results by casting the reconstruction of visual stimulus as the Bayesian inference of missing view in a multiview latent variable model. Taken together, the innovation of this method could promote the better understanding of the human brain mechanism. However, for deep learning, the accurate reconstruction of image stimuli remains challenging due to the limited instances.

In deep learning, the conventional wisdom is that deep neural networks can learn from high-dimensional data, such as images or languages, but this is because they have many labeled samples to train on. These supervised learning models need large amounts of labelled data and many iterations to train the large number of parameters. However, there are always less instances in fMRI measurement. As a consequence, common deep learning methods applied to the visual reconstruction have poor effect due to the over-fitting. It is hard to obtain more instances because of equipment limitation and task complexity. Moreover, it takes a long time to acquire fMRI data. The performance of the visual reconstruction method can be improved when implementing feature mapping first and then reconstruction. Given the lack of data, it is hard to realize the end-to-end method for achieving state-of-the-art performance. Therefore, we need a method to solve the problem of limited data instances and realize less sample learning. We think that the current convolution neural networks do not make full use of these instances, because, when a batch of samples is sent to the network, the neural network has learned the average features of those inputs and updates the weights of parameters utilizing this information. However, it does not consider the relationship between different batches of samples, which might be better on characteristic expression when it is taken into consideration, just like humans always recognize images by comparing the difference of images, even though they never have a concept of a certain category of object. Aside from that, the confusion that is caused by the high dimensionality of fMRI data instances needs to be solved. We propose a fully connected network to select voxels to reduce its dimensionality to train.

To make full use of data by considering the relationship of samples, an idea hits our mind that humans have an inherent ability to acquire and recognize new patterns. When presented with image stimuli, people seem to quickly understand new concepts [26]. Humans are very good at recognizing
objects with very little direct supervision, or none at all i.e., few-shot [27] or zero-shot [28] learning. Aimed at learning the ability of humans, one-shot learning is addressed by developing domain-specific features or inference procedures, which possess highly discriminative properties for the target task. Siamese nets were first introduced in the early 1990s by Bromley and Le Cun to solve signature verification as an image matching problem [29]. The seminal work toward one-shot learning dates back to the early 2000s with the work of Li [30]. Koch et al. [31] explored a method for learning Siamese neural networks that employed a unique structure to naturally rank the similarity between inputs. The Siamese neural network consists of twin networks that accept distinct inputs, and the network is joined by an energy function at the top. In a word, the Siamese network makes the output features cluster or disperse by learning a similarity function. Otherwise, because of the input pair, it increases the training network samples to solve overfitting. It can further play a role in taking full advantage of the data. With the further development of Siamese neural networks, Hoffer and Harwood proposed the triplet network model to learn useful representations by distance comparisons [32,33]. However, it is usually used in the classification problem and utilized for visual reconstruction in this study. In our opinion, the Siamese neural network is useful in visual reconstruction by the end-to-end method when converging or dispersing the input pairs in characteristic space whether they belong to the same or different classes.

For humans, only a small amount of images or no image is needed to recognize an object. Based on the description of the object, the object can be identified based on past experience. That is the reason why human can recognize a new thing that our visual system is naturally capable of extracting features from any object and comparing them. We do not have to see what we have seen before, because we are able to compare different objects, and this is the key why we have a small sample learning ability. However, the neural network only has the ability of extracting features instead of being comparative. The comparative ability of humans is because we have prior knowledge that we can use to learn [34]. Therefore, how do we achieve this kind of fast learning? In our study, we take the comparative relationship into consideration to realize few-shot learning. We think that people recognize the image by comparing the features in the image to achieve recognition, that is, less sample learning. Children can quickly identify what is a “horse” and “zebra”, even if they have not seen them a few times because our visual cells can automatically extract the features (e.g., outlines, lighting) of images and then compare to our previous experience to identify the image [34].

In this study, we propose a Siamese reconstruction network (SRN) for visual reconstruction based on the Siamese neural network. For the reconstructed images, if they are come from the same class, the relevant dimensions of similarity may be constructed in the course of learning to learn. The SRN can construct meaningful representations of what makes two objects deeply similar that go substantially beyond low-level image features; our method draws its strength from making full use of data by learning to compare when we present sample pairs to the network. In this way, the network potentially changes the distribution of characteristics in certain layers by clustering the feature vectors of same classes and enlarging the distance from the feature vectors of different classes before reconstructing the visual image. It can also increase the training samples similar to the Siamese neural network. Through the effective use of this data, end-to-end reconstruction performance will be improved. Finally, the satisfying results of two fMRI recording datasets demonstrate that our approach can be used on the limited instances of fMRI measurements for reconstructing visual images.

In summary, we make the following contributions: (1) Inspired by the visual mechanism, we introduce the mechanism of comparison into deep learning to realize better visual reconstruction via the Siamese neural network. (2) A novel Siamese reconstruction network (SRN) was developed to reconstruct the image stimuli from human fMRI. (3) Our proposed SRN achieved overwhelming superiority and exceeded by about 10% than the state of the art.
2. Materials and Methods

2.1. Experiment Data

Two public fMRI datasets were obtained from van Gerven et al. [14,35]. Dataset 1 contains 100 handwritten grayscale digits (with equal number of the class “6” and the class “9”) at a 28 × 28 pixel resolution that were taken from the training set of the MNIST database and the fMRI data from V1, V2, and V3. Dataset 2 contains 360 grayscale handwritten characters (with equal number of the class “B”, the class “R”, the class “A”, the class “N”, the class “S”, and the class “9”) at a 56 × 56 pixel resolution taken from van Gerven et al. and the fMRI data of V1 and V2 taken from three subjects [16]. These handwritten gray-scale digits and characters were presented to one subject. There are four runs interspersed with 30 s rest periods to perform 100 trials in total. In each trial, a handwritten digit or character was presented to the subject, and each block lasts 12.5 s. Blood-oxygenation-level dependent (BOLD) is obtained by the Siemens 3T MRI system. This experiment employed the EPI sequence with a repetition time (TR) of 2.5 s and isotropic voxel size of 2 × 2 × 2 mm³. The fMRI data of each image stimulus contains 3092 voxels in total from V1, V2, and V3 regions. Additionally, the detailed information regarding the fMRI data can refer to Van Gerven et al. and the two public datasets can be downloaded through http://artcogsys.com. We conducted experiments on the two datasets to test the SRN. The visual images from Dataset 2 were downscaled from 56 × 56 pixels to 28 × 28 pixels in our experiments.

2.2. Selecting Voxels

Voxel selection is an important component in fMRI brain decoding, because many voxels may have small correlation, which might not respond to the visual stimulus. It not only results in high dimensionality of fMRI data, but also overfitting during training. Although the loss of a part of the information usually occurs in a dimension reduction of voxels, the effect of overfitting can be mitigated because dimension reduction of voxels significantly increases the ratio of sample size and dimension.

Faced with the problem of how to select appropriate voxels, we proposed a method that build a encoding model mapping image stimuli to voxels, and to choose those voxels that are maximally correlated with the image stimuli by measuring the correlation with the effect of fitting namely encoding performance, as shown in Figure 1. First, we establish a classification network with five fully connected layers and then train the network, respectively, with the original images of the Dataset 1 and Dataset 2 to achieve a high accuracy rate of classification, which is 98% on the Dataset 1 and 94% on the Dataset 2. Next, the image from the training set is input to the network, and the vectors of the last but one layer are output as the high dimensional features. Finally, we chose voxels for which the model provided better predictability (encoding performance). This codifies our intuition that voxels better predicted with visual images are those to be included in the decoding model. The goodness-of-fit between model predictions and measured voxel activities was quantified while using the coefficient of determination (R²), which indicates the percentage of variance that is explained by the model. In the experiments, we computed the R² of each voxel while using 10-fold cross validation on training data, and then voxels with positive R² were selected for further analysis. Finally, we selected those voxels whose R² is at the top 200, and reduced the dimensionality of fMRI data from 3092 to 200.

2.3. Overview of the Proposed Siamese Reconstruction Network

To achieve accurate visual reconstruction, we proposed the SRN to fully use the data because of the limitation of samples. Inspired by humans’ ability to recognize images by learning to compare, the network uses the correlation between the pair of inputs as the pattern of brain response to difference of images. It not only considers the relationship in the same batch of samples, but also the samples in different batches. Moreover, it also increases the compactness of the feature vectors of same classes and enlarges the distance from feature vectors of extra-classes during the learning process. The Siamese neural network inspires the method. A Siamese neural network consists of twin networks that accept
distinct inputs but are joined by an energy function at the top. In this study, we have been influenced by Koch et al. [31], who proposed Siamese neural networks for one-shot image recognition. This network consists of twin networks that accept distinct inputs, but they are joined by an energy function at the top. Moreover, the network is symmetric, so that, whenever two distinct images are presented to the twin networks, the top conjoining layer will compute the same metric as if they were to present the same two images, but to the opposite twins. After learning, the network can increase the compactness of the feature vectors of the same classes and enlarge the distance from the feature vectors. Therefore, our proposed reconstruction network consists of twin sequences of fully connected layers. Additionally, the SRN accepts a pair of inputs that consisted of anchor fMRI data and target fMRI data. The anchor fMRI data and target fMRI data are selected from training dataset randomly. The anchor fMRI data is input to the main channel, and the processing of the sample needs to be reconstructed is called main channel, as shown in the upper portion of Figure 2. The target fMRI data are input to the auxiliary channel and the processing of the sample used to constrain the feature expression of network is called the auxiliary channel, as shown in the bottom portion of Figure 2. Moreover, the validation sets of fMRI data are input to the main channel, and they can be reconstructed more accurately in the trained Siamese network. Moreover, the validation sets of fMRI data are input to the main channel, and they can be more accurately reconstructed in the trained Siamese network.

We now describe the structure of the Siamese reconstruction net in detail and the specifics of the learning algorithm used in our experiments, as follows.

2.4. Model

Figure 2 summarizes the framework of our network. It only contains fully connected network with $L$ layers each with $N_l$ units, where $h_{1,l}$ represents the hidden vector in layer $l$ for the first twin,
and $h_{2J}$ denotes the same for the second twin. We use exclusively rectified linear units (ReLU) in all layers. This network consists of twin networks that accept distinct inputs but they are joined by an energy function at the last but one layer and with shared weight matrices at each layer. Additionally, the output of the last layer is measured the similarity of the original images by L2 component-wise distance. In addition, the detailed network architectures are shown as in Table 1. Let FC (K) denote the fully connected network with K filters.

<table>
<thead>
<tr>
<th>Anchor Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>FC(2048) + ReLU + Dropout</td>
</tr>
<tr>
<td>FC(1024) + ReLU + Dropout</td>
</tr>
<tr>
<td>FC(1024) + ReLU + Dropout</td>
</tr>
<tr>
<td>FC(512) + ReLU + Dropout</td>
</tr>
<tr>
<td>FC(784) + ReLU</td>
</tr>
</tbody>
</table>

**Loss Function.** In this paper, we compute two losses. One is to compute L2 component-wise distance between high dimensional features as the Siamese loss, the other is to compute L2 component-wise distance between reconstructed images and original images as MSE loss.

**Siamese Loss.** From the last but one layer, the anchor fMRI data and target fMRI data are inferred from the distinct data batches. Subsequently, we calculate the embedding of them. The embedding is represented by $f(x) \in \mathbb{R}^d$. It embeds the voxel responses $x$ into a $d$-dimensional Euclidean space. Here, we want to ensure that if the high dimension feature $f(x_a^i)$ (anchor) and $f(x_t^i)$ (target) of a specific voxels vector from same classes are closer than those from different classes. Figure 3 visualizes this.

![Figure 3](image)

*Figure 3.* The Siamese Loss minimizes the distance between an anchor and a positive target, both of which coming from same class, and maximizes the distance between the an anchor and a negative target coming from the different classes.

Let $M$ represent the minibatch size, where $i$ indexes the $i$th minibatch. Now, let $y(x_a^i, x_t^i)$ be a length $M$ vector, which contains the labels for the minibatch, where we assume $y(x_a^i, x_t^i) = 1$ whenever $x_a^i$ and $x_t^i$ are from the same character class and $y(x_a^i, x_t^i) = 0$ otherwise.

The Siamese loss that is being minimized is then

$$L_s(x_a^i, x_t^i) = y(x_a^i, x_t^i) \log p(x_a^i, x_t^i) + (1 - y(x_a^i, x_t^i)) \log p(1 - y(x_a^i, x_t^i))$$

(1)

**MSE Loss.** From the last layer, the anchor network outputs a batch of 784 dimensional vectors $f(x_a^i)$, which is compared with the original image $y_i$, to make them more similar. The MSE loss where the L2 component-wise distance between vectors is computed that is being minimized is, then

$$L_m = \sum_{i=1}^{N} \left[ \|f(x_a^i) - y_i\|_2^2 \right]$$

(2)
Subsequently, we combine the Siamese loss $L_s$ and MSE loss $L_m$ together, as follows. $\alpha$ and $\beta$ are the hyperparameters, that are the weighting factors for Siamese loss and MSE loss, respectively. $w$ is the weight of network and $\lambda\|w\|_2$ is the regularization term avoiding over fitting. The total loss that is being minimized is then

$$L_t = \alpha L_s + \beta L_m + \lambda\|w\|_2$$

(3)

2.5. Implementation

In our experiments, we use Tensorflow for the implementation and test them on a NVIDIA Tesla K100 GPU cluster in Nvidia DGX station. We train the neural network for 100 epochs with a batch size of 64, with the learning rate of 0.01, being decreased by 10% every 20 steps, and use Adam as our solver.

3. Experimental Results

3.1. The Results of Reconstruction

We have trained the datasets in the network to prove the effectiveness of the proposed SRN. The reconstructed handwritten digits and characters of the proposed SRN are shown in Figures 4 and 5, respectively, where the first row denotes the presented images and the second row denotes the reconstructed images of SRN.

![Figure 4. Examples of reconstructed 10 distinct handwritten digits obtained from Dataset 1.](image1)

![Figure 5. Examples of reconstructed 12 distinct handwritten characters obtained from Dataset 2.](image2)

Overall, the images that were reconstructed by SRN have shown fine reconstruction effects for handwritten digits and characters, especially in capturing the essential features of the presented images. Although there are subtle differences in the strokes, the obtained reconstructions of handwritten digits and characters shared certain characteristics of their corresponding original images. In addition, some original images that have complex shapes and architectures are difficult to be reconstructed, but they are more recognizable to the correct class that was reconstructed by the proposed SRN.

We attribute this phenomenon to the fact that, when we train the SRN for visual reconstruction, we take the relationship of different batches into consideration, which takes advantage of data in the limitation of sample number. After the training process, the feature vectors of same classes are clustered, while the feature vectors of different classes are dispersed. Therefore, the features that were extracted by well-trained SRN from the same classes are similar. The images reconstructed based on these features have similar appearance. It is an open and difficult problem to evaluate the
quality of reconstructed images [36]. To better evaluate the reconstruction performance quantitatively, we used several standard image similarity metrics, including Pearson’s correlation coefficient (PCC), mean squared error (MSE), and structural similarity index (SSIM) [20]. However, traditional metrics, such as per-pixel MSE, not evaluating the joint statistics of the results is because they do not measure the very structure that the structured losses aim to capture. Therefore, as the auxiliary metrics, MSE and PCC do not well indicate the similarity. By contrast, SSIM can address this shortcoming by taking texture into account, so it is a more convincing measure of structural similarity.

In the experiment, we employed 10-fold cross validation to test the proposed Siamese method. We listed the PCC, MSE, and SSIM of the reconstructed images in the SRN. The results of Dataset 1 and Dataset 2 are shown in Tables 2 and 3, respectively.

<p>| Table 2. The main corresponding quantitative evaluation of structural similarity index (SSIM) reconstructed by Siamese reconstruction network (SRN) for each cross validation on Dataset 1. |
|---|---|---|---|---|---|---|---|---|---|</p>
<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRN</td>
<td>0.739</td>
<td>0.747</td>
<td>0.735</td>
<td>0.737</td>
<td>0.720</td>
<td>0.717</td>
<td>0.731</td>
<td>0.675</td>
<td>0.684</td>
</tr>
</tbody>
</table>

<p>| Table 3. The main corresponding quantitative evaluation of SSIM reconstructed by SRN for each cross validation on Dataset 2. |
|---|---|---|---|---|---|---|---|---|</p>
<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRN</td>
<td>0.422</td>
<td>0.430</td>
<td>0.483</td>
<td>0.457</td>
<td>0.442</td>
<td>0.465</td>
<td>0.444</td>
<td>0.433</td>
<td>0.459</td>
</tr>
</tbody>
</table>

Next, we presented the quantitative results of the two test datasets based on 10-fold cross validation as compared with the several state-of-the-art methods. As shown in Table 4, in both Dataset 1 and Dataset 2 the PCC and MSE of our proposed Siamese method were slightly weaker than those of the DGMM [20]. When compared with the latest methods, the Siamese reconstruction net provided about 8% higher accuracy on the most important SSIM metric in Dataset 1 and 10% higher accuracy in Dataset 2. The better performance is caused by the fact that converging or dispersing input pairs in feature space according to whether they belong to the same or different classes and making full use of limited training samples by learning to compare. The high values of MSE and PCC should not be a cause of concern, because it is not so important for pixel-level comparison due to the complex noise in the fMRI data and limited samples. Moreover, humans do not care much on detailed information at the pixel level and the structure according to attention mechanism. This is the reason why we should disperse the characteristics of extra-classes to distinguish the class. Consequently, the two metrics only serve as auxiliary measures and our proposed method performed better overall.

<p>| Table 4. Performance of several image reconstruction methods on the test sets. Results were averaged over 10-fold cross validation. |
|---|---|---|---|</p>
<table>
<thead>
<tr>
<th>Datasets</th>
<th>Algorithms</th>
<th>PCC</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset 1</td>
<td>Miyawaki et al. [13]</td>
<td>0.767</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>BCCA [18]</td>
<td>0.411</td>
<td>0.119</td>
</tr>
<tr>
<td></td>
<td>DCCAE-A [20]</td>
<td>0.548</td>
<td>0.074</td>
</tr>
<tr>
<td></td>
<td>DGMM [25]</td>
<td>0.803</td>
<td>0.037</td>
</tr>
<tr>
<td></td>
<td>FCN</td>
<td>0.594</td>
<td>0.065</td>
</tr>
<tr>
<td></td>
<td>SRN</td>
<td>0.736</td>
<td>0.045</td>
</tr>
<tr>
<td>Dataset 2</td>
<td>Miyawaki et al.</td>
<td>0.481</td>
<td>0.067</td>
</tr>
<tr>
<td></td>
<td>BCCA</td>
<td>0.348</td>
<td>0.128</td>
</tr>
<tr>
<td></td>
<td>DCCAE-A</td>
<td>0.354</td>
<td>0.073</td>
</tr>
<tr>
<td></td>
<td>DGMM</td>
<td>0.498</td>
<td>0.058</td>
</tr>
<tr>
<td></td>
<td>FCN</td>
<td>0.386</td>
<td>0.071</td>
</tr>
<tr>
<td></td>
<td>SRN</td>
<td>0.461</td>
<td>0.068</td>
</tr>
</tbody>
</table>
We set up the comparative experiment of single fully connect network (FCN) in order to further illustrate the effect of Siamese reconstruction net, as shown in the Table 4. The Siamese reconstruction net provided about 12% higher accuracy on the most important SSIM metric in Dataset 1 and 13% higher accuracy in Dataset 2 when compared with the FCN methods.

The images that were reconstructed by the Siamese reconstruction net on each dataset performed poorer or better than the existing methods. We attribute this phenomenon to the fact that, because of the handwritten digits having simpler characteristics and less variety, the image similarity and its quantitative evaluation are significantly superior to those of characters’. Therefore, the images reconstructed by the Siamese reconstruction net can hardly be improved in Dataset 1. In addition, there are several complex characteristics and a variety of the obtained reconstructions of handwritten characters, which increase the difficulty in reconstruction, but improve the accuracy of distinction by Siamese reconstruction net.

3.2. The Clustering of Visual Voxel Responses

Giving more account for the efforts of the Siamese reconstruction net, we output the high dimensional features of last but one layer. Subsequently, those features come from different classes are mapped to 2-D space by T-SNE [37]. As shown in Figure 6, below, the red points represent the features of samples coming from class “9” and the green points represent the features of samples coming from the class “6”.

![Figure 6](image_url) The effect on projection of clustering and dispersing characteristics.

The figure shows the effect of clustering and dispersing characteristics. After using the Siamese reconstruction net, the features coming from the different class tend to be relatively dispersive, and those coming from same class tend to be relatively clustering. However, the clustering effect of the same classes is not as ideal as we thought. We analyzed that the proposed method should not cluster the features of the same class in visual reconstruction as closer as classification problems, because the purpose of the network is to accurately reconstruct the visual images. The original images, even in the same class, have complex shapes and various structures, so that, when training, it will not increase the compactness of the feature vectors of the same class and therefore enlarge the distance from the feature vectors between the extra-classes.

4. Discussion

**Advantage.** When compared to current convolution neural networks, the Siamese neural network, which accepts sample pairs, can not only make use of every samples, but also the relationship between
every two samples during training process, thus may work better than the plain deep network. In this study, we introduced SRN into the visual reconstruction and improved the effort of reconstruction through converging or dispersing the input pairs in feature space according to whether they belong to the same or different classes. Essentially, the operation, called “learning to compare”, can be regarded as one regularization term, thus the potential to other visual decoding task that usually encounters the limited training samples.

**Selective Reconstruction.** Given the perplexing fMRI measurement noise and high dimensionality of limited data instances, it is important and challenging to remove those ineffective voxels to reduce the dimensionality of voxels and avoid interference, while keeping the necessary information. In the study, through selecting valuable voxels as many previous studies, we obtained improved reconstruction. However, some information was also removed during the selection, and making full use of acquired voxels is hard to achieve. Thus, how to keep necessary information and avoid interference in the developing the computation model may need to attract more attention.

**Inspiration of Human Visual Mechanism.** Humans can recognize a new image without having seen lots of images, because the human visual system is naturally capable of extracting features from objects and comparing them based on the past experience. Therefore, we introduced the comparative relationship to the deep neural network and improve the reconstruction effect. There are some other intriguing mechanisms in human visual system, such as visual attention that human can quickly locate important target areas and conduct detailed analysis. For the further development of brain decoding, we can introduce other human visual mechanisms to the deep neural network.

5. Conclusions

This study first proposed SRN based on Siamese neural network for the visual reconstruction, inspired by humans’ inherent ability to recognize one image while using few samples by comparing between images. There are always less instances in fMRI measurement, while these supervised learning models need large amounts of labelled data to train the large number of parameters, so common deep learning methods that are applied to the visual reconstruction have a poor effect due to the over-fitting. To make full use of data, the proposed method takes the relationship of sample pairs into account through changing the distributions in feature space by clustering the feature vectors of similar classes and enlarging the distance from the feature vectors of different classes during learning. Given the high dimensionality of limited data instances, it is essential to select voxels to reduce the dimensionality of fMRI data and learn the mapping from fMRI data to extract features in visual reconstruction. First, we employed the extracted features based on trained classification networks to estimate the correlation of each voxel to accomplish the selection of voxels. Second, we trained the reconstruction network with the selected voxels. Based on qualitative and quantitative results, the proposed Siamese reconstruction net performs considerably better in all two datasets in the most important SSIM metric, providing 8% and 10% higher accuracy in Dataset 1 and Dataset 2 as compared to the state-of-the-art methods. These results have demonstrated that our proposed Siamese reconstruction net architecture plays a role in clustering the characters of the same class and it showed effectiveness in visual reconstruction. To the best of our knowledge, this study is the first to propose a mechanism of comparison for dramatically clustering the characters based on Siamese neural network in visual image reconstruction. Visual reconstruction was realized by an end-to-end method instead of mapping the voxel responses to the feature space. In this work, we did not explore the full range of possibilities that Siamese reconstruction net potentially enables. Our future work includes applications of our method to fMRI mental illness classification problem.

**Author Contributions:** Conceptualization, L.J., H.B. and B.Y.; Formal analysis, L.W.; Methodology, L.J., K.Q. and C.Z.; Software, L.J. and J.C.; Validation, J.C. and L.Z.; Writing-original draft, L.J.; Writing-review & editing, K.Q., L.W., H.B. and B.Y.

**Funding:** This work was funded by the National Key R&D Program of China under grant 2017YFB1002502 and National Natural Science Foundation of China (No. 61701089 and No. 61601518).
**Conflicts of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**References**


