entropy

MDPI

Article

Novel Convolutional Neural Network with Variational Information Bottleneck for P300 Detection

Hongpeng Liao [1], Jianwu Xu [2] and Zhuliang Yu [1,3,*]

[1] College of Automation Science and Technology, South China University of Technology, Guangzhou 510641, China; 201720116235@mail.scut.edu.cn
[2] Guangzhou Galaxy Thermal Energy Incorporated Company, Guangzhou 510220, China; gdxingchen@126.com
[3] Pazhou Lab., Guangzhou 510330, China
[*] Correspondence: zlyu@scut.edu.cn

Abstract: In the area of brain-computer interfaces (BCI), the detection of P300 is a very important technique and has a lot of applications. Although this problem has been studied for decades, it is still a tough problem in electroencephalography (EEG) signal processing owing to its high dimension features and low signal-to-noise ratio (SNR). Recently, neural networks, like conventional neural networks (CNN), has shown excellent performance on many applications. However, standard convolutional neural networks suffer from performance degradation on dealing with noisy data or data with too many redundant information. In this paper, we proposed a novel convolutional neural network with variational information bottleneck for P300 detection. Wiht the CNN architecture and information bottleneck, the proposed network termed P300-VIB-Net could remove the redundant information in data effectively. The experimental results on BCI competition data sets show that P300-VIB-Net achieves cutting-edge character recognition performance. Furthermore, the proposed model is capable of restricting the flow of irrelevant information adaptively in the network from perspective of information theory. The experimental results show that P300-VIB-Net is a promising tool for P300 detection.

Keywords: variational information bottleneck; convolutional neural network; P300 signal detection

check for updates

Citation: Liao, H.; Xu, J.; Yu, Z. Novel Convolutional Neural Network with Variational Information Bottleneck for P300 Detection. *Entropy* **2021**, *23*, 39. https://doi.org/10.3390/e23010039

Received: 19 November 2020
Accepted: 24 December 2020
Published: 29 December 2020

1. Introduction

Brain-computer interface (BCI) provides a way for people to interact with the environment without any muscle activities, especially for people with amyotrophic lateral sclerosis, spinal cord injuries or other severe motor disabilities [1]. Event-related potentials (ERP), which is one of the important electroencephalography (EEG) signals, reflects neural activities after events. As a component of ERP, P300 is named after that positive potentials peaks occurs at about 300 ms after event-related stimuli [2]. P300 is widely used in BCI applications, like character recognition [3] and video surveillance [4].

Although P300 has been studied for long time, the detection of P300 is still challengeable in the case of low signal-to-noise ratio (SNR) due to unrelated neural activities and artifacts [5]. Lots of approaches were proposed for P300 detection [6–9]. Recently, the machine learning based methods achieved excellent performance on P300 detection [10–12]. For the traditional machine learning methods, feature extraction and classification are two of the key techniques. Principal component analysis [13], wavelet transform technique [14] were used for effective feature extraction. Support vector machine (SVM) is always used as a powerful classifier in P300 detection. In BCI Competition III [14], 17 SVM were ensembled for P300 detection and achieved the best performance. Group-sparse Bayesian linear discriminant analysis (gsBLDA) reached comparable classification accuracy, which treated signals of different channels as different groups [15].

Besides the traditional machine learning based methods, recently, deep learning models with different kinds of techniques had achieved great performance in many areas including the detection of ERP signal. A classic convolution neural network (CNN) for the detection of P300 waves was first proposed in Reference [16], that contains the spatial filter and temporal filter layers to well extract the spatial-temporal information of P300 signals. To make the CNN more robust, batch normalization and dropout were integrated into the proposed CNN, the resulting CNN is less sensitive to overfitting [17]. To further develop network which could find ERP components from data automatically, Restricted Boltzmann Machine (RBM) was utilized in ERP-Net [5]. The ERP-NET could discover all the ERP patterns contained in EEG signals. In addition, a spatial-temporal discriminative Restricted Boltzmann Machine (ST-DRBM) was further proposed [18] to learn spatial and temporal features separately and characterize the scalp distribution and temporal diversification. ST-DRBM has higher performance for ERP detection and it provides physiologically explainable results.

To further improve the robustness of network in P300 detection, in this paper, we propose a novel convolutional neural network based on variational information bottleneck. The proposed network, which is named as P300-VIB-Net, could reduce irrelevant information adaptively from the data. Hence, it is more robust against noise as well as irrelevant information. The contributions of this paper are summarized as: (1) A novel neural network architecture, P300-VIB-Net, is proposed for P300 detection. It combines the CNN as well as variational information bottleneck to make the network more robust against irrelevant information; (2) P300-VIB-Net reaches the state-of-art performance in P300 speller experiments; (3) We provide an explanation from the perspective of information theory on how the variational information bottleneck works with CNN. This also provides new insights on regularization technique.

2. Deep Learning Based on Variational Information Bottleneck

Deep neural networks can be explained in the information-theoretical framework [19] by information bottleneck (IB) that aims to find the short code for input which maintains the maximum information about output with mutual information [20]. The neural network with variational information bottleneck (VIB) show less overfitting and adversarial robustness [21]. Recently, variational discriminator bottleneck (VDB) with IB gets an important improvement in imitation learning, adversarial inverse reinforcement learning, and generative adversarial network (GAN) [22]. Information dropout was generalized by dropout based on IB, making better use of architectures with limited capacity [23]. In this section, an introduction on IB is presented as follows.

2.1. Information Bottleneck Principle

Relevant information in input data $\mathbf{x} \in \mathbf{X}$ is defined as the information that signal $\mathbf{x}$ provides about output data $\mathbf{y} \in \mathbf{Y}$. Signal coding focuses on discovering the representation $\mathbf{Z}$ of $\mathbf{X}$, as known as code or hidden variables, keeping the most information about $\mathbf{Y}$, which is measured by mutual information $I(\mathbf{Z}, \mathbf{Y})$ between $\mathbf{Z}$ and $\mathbf{Y}$

$$I(\mathbf{Z}, \mathbf{Y}) = \int_{\mathbf{z}} \int_{\mathbf{y}} p(\mathbf{z}, \mathbf{y}) \log \frac{p(\mathbf{z}, \mathbf{y})}{p(\mathbf{y}) p(\mathbf{z})} d\mathbf{y} d\mathbf{z}. \tag{1}$$

It is obvious that $I(\mathbf{Z}, \mathbf{Y})$ achieves the maximal value by taking the identity coding of input data as $\mathbf{Z} = \mathbf{X}$. This identity encoding is not a useful representation of the processed data. Hence, in practice, the constraint $I(\mathbf{X}, \mathbf{Z}) \leq I_c$ is imposed as the 'bottleneck', where I_c is a constant, restricting the information from $\mathbf{X}$ to $\mathbf{Z}$. This suggests the objective

$$\max \ I(\mathbf{Z}, \mathbf{Y}) \ s.t. \ I(\mathbf{X}, \mathbf{Z}) \leq I_c. \tag{2}$$

By introducing a Lagrange multiplier β, the above problem can be formulated as minimizing the function below to get the ideal representation,

$$-I(\mathbf{Z},\mathbf{Y})+\beta I(\mathbf{X},\mathbf{Z}). \tag{3}$$

Minimizing the first term enhances the transfer of information from the intermediate coding variable **Z** to output variable **Y**, while minimizing the second term limits the transfer of information from the input variable **X** to the intermediate coding variable **Z**. We can find a suitable β to preserve minimal information from **X** to **Z** and the information in **Z** is sufficient to predict **Y**.

The information bottleneck principle discussed above defines an optimal representation and holds the most distinctive information about **Y** in **X**. However, the computation about mutual information in information bottleneck principle is always hard except some very restrictive cases. How to simplify the calculation of problem in (3) is always an important problem in practice.

2.2. Variational Information Bottleneck

To solve the computational problem in IB, two significant improvements were proposed in the variational information bottleneck [21]. Firstly, variational inference is applied to build an upper bound of the function of IB. Secondly, the objective function can be optimized by stochastic gradient descent with the reparameterization trick [24]. Deep neural networks can be used for parameterization of distributions.

For $I(\mathbf{Z},\mathbf{Y})$ defined in (1), since $p(\mathbf{y}|\mathbf{z})$ is difficult to obtain in practice, we use a variational approximation $q(\mathbf{y}|\mathbf{z})$ to approximate $p(\mathbf{y}|\mathbf{z})$. Since the Kullback-Leibler divergence is non-negative, that is, $KL(p(\mathbf{y}|\mathbf{z})||q(\mathbf{y}|\mathbf{z}))\geq 0$, we have

$$\int_{\mathbf{y}} p(\mathbf{y}|\mathbf{z})\log p(\mathbf{y}|\mathbf{z})d\mathbf{y} \geq \int_{\mathbf{y}} p(\mathbf{y}|\mathbf{z})\log q(\mathbf{y}|\mathbf{z})d\mathbf{y}. \tag{4}$$

Therefore,

$$\begin{aligned}
I(\mathbf{Z},\mathbf{Y}) \quad &= \int_{\mathbf{z}} p(\mathbf{z})\int_{\mathbf{y}} p(\mathbf{y}|\mathbf{z})\log\frac{p(\mathbf{y}|\mathbf{z})}{p(\mathbf{y})}d\mathbf{y}d\mathbf{z} && (5)\\
&\geq \int_{\mathbf{z}} p(\mathbf{z})\int_{\mathbf{y}} p(\mathbf{y}|\mathbf{z})\log\frac{q(\mathbf{y}|\mathbf{z})}{p(\mathbf{y})}d\mathbf{y}d\mathbf{z} && (6)\\
&= \int_{\mathbf{z}}\int_{\mathbf{y}} p(\mathbf{z},\mathbf{y})\log q(\mathbf{y}|\mathbf{z})d\mathbf{y}d\mathbf{z}+H(\mathbf{Y}), && (7)
\end{aligned}$$

where $H(\mathbf{Y}) = -\int_{\mathbf{z}} p(\mathbf{z})\log p(\mathbf{z})d\mathbf{z}$ is independent of the optimization, and it can be neglected. Recall Markov assumption about joint distribution $p(\mathbf{X},\mathbf{Y},\mathbf{Z})$ in Reference [21], which is $p(\mathbf{Z}|\mathbf{X},\mathbf{Y}) = p(\mathbf{Z}|\mathbf{X})$, corresponding to the Markov chain$Y \leftrightarrow X \leftrightarrow Z$, we have $p(\mathbf{z},\mathbf{y}) = \int_{\mathbf{x}} p(\mathbf{x})p(\mathbf{y}|\mathbf{x})p(\mathbf{z}|\mathbf{x})d\mathbf{x}$. Consequently,

$$I(\mathbf{Z},\mathbf{Y}) \geq \int_{\mathbf{z}}\int_{\mathbf{y}}\int_{\mathbf{x}} p(\mathbf{x})p(\mathbf{y}|\mathbf{x})p(\mathbf{z}|\mathbf{x})\log q(\mathbf{y}|\mathbf{z})d\mathbf{x}d\mathbf{y}d\mathbf{z}. \tag{8}$$

Similarly, we can get the upper bound $I(\mathbf{X},\mathbf{Z})$ of the second term in (3), because of $KL(p(\mathbf{z})||a(\mathbf{z}))\geq 0$, where $a(\mathbf{z})$ is the variational approximation of $p(\mathbf{z})$, which is

$$\int_{\mathbf{z}} p(\mathbf{z})\log p(\mathbf{z})d\mathbf{z} \geq \int_{\mathbf{z}} p(\mathbf{z})\log a(\mathbf{z})d\mathbf{z}. \tag{9}$$

Therefore,

$$I(\mathbf{X},\mathbf{Z}) = \int_{\mathbf{z}}\int_{\mathbf{x}} p(\mathbf{z},\mathbf{x}) \log \frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{x})p(\mathbf{z})} d\mathbf{x}d\mathbf{z} \tag{10}$$

$$= \int_{\mathbf{z}}\int_{\mathbf{x}} p(\mathbf{z}|\mathbf{x})p(\mathbf{x}) \log \frac{p(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})} d\mathbf{x}d\mathbf{z} \tag{11}$$

$$\leq \int_{\mathbf{x}} p(\mathbf{x}) KL(p(\mathbf{z}|\mathbf{x})||a(\mathbf{z}))d\mathbf{x}. \tag{12}$$

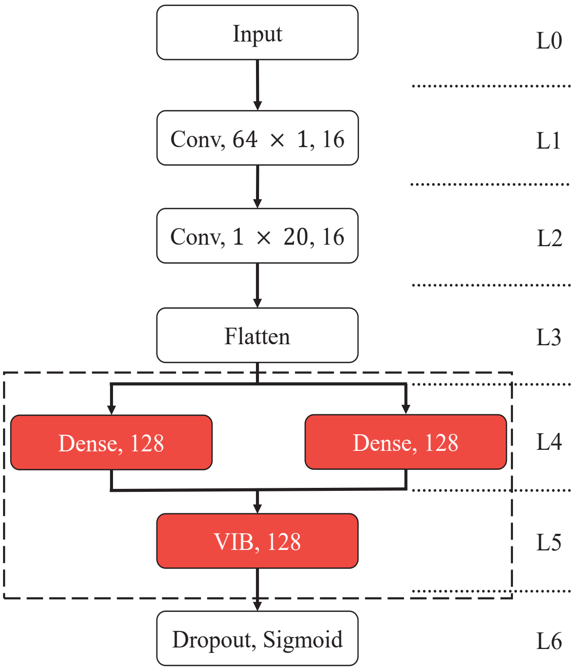


Figure 1. Details of network architecture. Pivotal information of each layer is shown in solid boxes. Crucial part of the proposed model is shown in dashed box.

Using empirical data distribution $p(\mathbf{x},\mathbf{y}) = \frac{1}{N}\sum_{n=1}^{N} \delta_{x_n}(\mathbf{x})\delta_{y_n}(\mathbf{y})$, where N is the number of samples, we can write the upper bound as

$$\frac{1}{N}\sum_{n=1}^{N}[-\int_{\mathbf{z}} p(\mathbf{z}|\mathbf{x}_n) \log q(\mathbf{y}_n|\mathbf{z})d\mathbf{z} \\ +\beta \mathrm{KL}(p(\mathbf{z}|\mathbf{x}_n)||a(\mathbf{z}))]. \tag{13}$$

The first term is in the form of a cross-entropy loss function. The second term can be regarded as a regularization term. $a(\mathbf{z})$ is the distribution we assume, usually a standard normal distribution. $p(\mathbf{z}|\mathbf{x})$ is an encoder, which transforms $\mathbf{X}$ into $\mathbf{Z}$. Suppose the encoder is of the form $p(\mathbf{z}|\mathbf{x}) = N(\mathbf{z}|f_e^{\mu}(\mathbf{x}), f_e^{\Sigma}(\mathbf{x}))$, where f_e is a neural network which outputs both the mean μ and covariance matrix Σ, we can use reparameterization trick to generate $\mathbf{z} = f(\mathbf{x}, \epsilon)$ which is a deterministic function of $\mathbf{x}$ and Guassian random variable ϵ. Since the noise ϵ is independent of parameters of the model, it is easy to take gradients in the training process. If our choice of $p(\mathbf{z}|\mathbf{x})$ and $a(\mathbf{z})$ allows computation of an analytic Kullback-Leibler divergence, we can get further simplified objective function in training.

2.3. P300-VIB-Net

In this paper, we proposed a new neural network for EEG classification. The proposed network as shown in Figure 1 is based on VIB and classic convolutional network architecture which is widely used in P300 detection problem. Parameters in the middle of convolution layers represents kernel size and parameter at the end of convolution layers is the number of kernels. The first few layers L0, L1, L2, and L3 of the model are similar to the layers of traditional convolutional network for P300 detection. L0 is the input layer. The size of input data to L0 is the number of channels multiplied by the length of signals in the time domain. L1 plays a role as a spatial filter to get the best combinations of signals from all electrodes. L2 serves as a temporal filter as well as a sub-sampler, which extracts the most important time-domain features. In L3, the feature map matrices, which are the outputs of L2, are flattened into vectors and input to the following fully connected layers.

In L4, there are two different fully connected networks that take the output of L3 as input to generate the mean and variance of encoder $p(\boldsymbol{z}|\boldsymbol{x})$ as

$$p(\boldsymbol{z}|\boldsymbol{x})= N(\boldsymbol{z}; \boldsymbol{\mu}, e^{\hat{\boldsymbol{\sigma}}}) \tag{14}$$

$$(\boldsymbol{\mu}, \hat{\boldsymbol{\sigma}})= \text{NeuralNet}_{\boldsymbol{\phi}}, (\boldsymbol{x}) \tag{15}$$

where $\text{NeuralNet}_{\phi}(\boldsymbol{x})$ represents the layers L0-L1-L2-L3-L4, ϕ is parameters of the network layers, $\boldsymbol{\mu}$ and $\hat{\boldsymbol{\sigma}}$ correspond to the output of two fully connected networks in L4. In order to guarantee the non-negativeness of covariance, the exponential of $\hat{\boldsymbol{\sigma}}$ is used to represent the variance of $\boldsymbol{z}$.

In L5, reparameterization tricks is applied for easy calculating of gradients. Firstly, we produce $\boldsymbol{\epsilon}$ by standard normal distribution function. Then, we generate $\boldsymbol{z}$ with $\boldsymbol{\mu}$, $\hat{\boldsymbol{\sigma}}$ and $\boldsymbol{\epsilon}$ as

$$\boldsymbol{\epsilon}\sim N(\boldsymbol{0}, \boldsymbol{I}) \tag{16}$$

$$\boldsymbol{z}= \boldsymbol{\mu} + e^{\hat{\boldsymbol{\sigma}}} \odot \boldsymbol{\epsilon}. \tag{17}$$

In L6, in order to avoid overfitting, the dropout [25] is used. Dropout is a universally used technique in deep learning, which makes the existence of any particular hidden unit untrustworthy and cuts down the co-adaptation of neurons, alleviating overfitting in neural networks at the cost of increased training time. By generating a binary vector $\boldsymbol{r}$ whose elements follow Bernoulli distribution with p as the parameter representing the means drop rate in dropout, we have the output as

$$\boldsymbol{r}\sim \text{Bernoulli}(p) \tag{18}$$

$$\boldsymbol{y}= f(\boldsymbol{r} * \boldsymbol{z} + \boldsymbol{b}), \tag{19}$$

where f is the sigmoid function $f(x) = \frac{1}{1+e^{-x}}$ that represents the P300 signal detection probability.

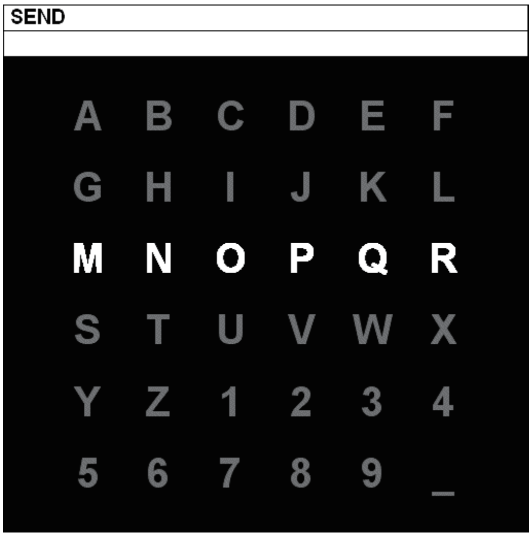
In the P300-VIB-Net, $p(\mathbf{z}|\mathbf{x})$ is parameterized by L0-L1-L2-L3-L4-L5, which encodes $\mathbf{X}$ into intermediate representation $\mathbf{Z}$. We suppose that $a(\mathbf{z})$ is $N(\mathbf{0}, \mathbf{I})$, $q(\mathbf{y}|\mathbf{z})$ is parameterized by L6 as described in (19). After taking the analytical result of KL divergence as [24], the loss function can be formulated as

$$\begin{aligned}\text{Loss} =& \frac{1}{N}\sum_{n=1}^{N} - \int_{\mathbf{z}} p(\mathbf{z}|\mathbf{x}_n) \log q(\mathbf{y}_n|\mathbf{z}) d\mathbf{z} \\ & -\frac{\beta}{2}\sum_{j=1}^{J}[1 + log((\sigma_j)^2) - (\mu_j)^2 - (\sigma_j)^2],\end{aligned} \tag{20}$$

where J is the size of hidden variable $\mathbf{z}$, μ_j and σ_j are j-th elements to generate $\mathbf{z}$. $\mathbf{y}_n$ are labels given in the datasets. The cross entropy of probabilities $\mathbf{y}$ and labels $\mathbf{y}_n$ will be calculated during the learning process to optimize (20).

3. Experimental Results

In this section, the experimental results of the proposed network on P300 speller paradigm will be presented. The proposed model, P300-VIB-Net, will be compared with other state-of-art algorithms to show the effectiveness of the proposed method.



SEND

A	B	C	D	E	F
G	H	I	J	K	L
M	N	O	P	Q	R
S	T	U	V	W	X
Y	Z	1	2	3	4
5	6	7	8	9	_

Figure 2. P300 speller interface in brain-computer interface (BCI) Competition III [26].

3.1. P300 Speller Paradigm

The occurrence of P300 is related to the human's reaction to the stimulus. P300 is relatively obvious and easy to observe among all ERP components. Thus, P300 is considered to reflect the process of receiving stimulation. Subjects usually are shown with a random sequence of target and non-target stimuli based on the oddball paradigm. Generally, the smaller the probability of the appearance of the target stimulus, the greater the magnitude of P300, from which we can find the target that subjects focus on.

Data set II of BCI competition III which is widely used as a benchmark data set for P300 detection is used as the test data set. The P300 speller paradigm of data set II was described in Reference [26]. It was based on the principle that flashing characters on the screen that subjects focus on will stimulate the presence of P300. The stimulation graphical user interface consists a 6 × 6 matrix, including 36 characters, as shown in Figure 2. Subjects are asked to focus on one character given in a prompter on top of the matrix. Every row and every column of the matrix flash once at the rate of 5.7 Hz randomly in each epoch. There are 2 of 12 intensifications that contain the target character at the intersection of the row and the column. Therefore, the target character can be detected by distinguishing P300 and non-P300 signals of every flashing. In other words, the classification of 36 classes is transformed into binary classification problem. 30 P300 and 150 non-P300 signals are obtained after repeating 15 times for each character. There are 85 characters in the training set and 100 characters in the test set of subject A and subject B.

3.2. Data Preprocessing

The 64 channel EEG data was collected with sampling rate 240 Hz. Data preprocessing consists four steps. First of all, the time window is chosen as 0$\sim$670 ms after flashing. Each sample is of size 64 × 160. Secondly, data is bandpass filtered by a 4-th order Chebyshev type I filter with bandwidth 0.1–20 Hz. Thirdly, every sample is normalized to be zero mean and unit variance. Last but not the least, since data in training sets and test sets is unbalanced, we duplicate the P300 signals 4 times to keep the data sets balance.

3.3. Experiments of P300-VIB-Net

The network model used in experiments is shown in Figure 1. Compared to traditional CNN network for P300 detection, the most significant modification is layers L4 and L5, which are the kernel components of VIB network presented in the dashed box.

Table 1. Character recognition rate of different models.

Subjects	**Epochs**	**Models**								
		CNN [16]	**MCNN-1** [16]	**E-SVM** [14]	**BN3** [17]	**PST-DRBM** [18]	**ERP-Net** [5]	**SAE-ESVM** [27]	**CM-CW-CNN-ESVM** [28]	**Proposed**
A	1	16	18	16	22	**24**	22	21	22	15
	5	61	61	72	73	**75**	**75**	72	64	66
	10	86	79	83	86	**90**	**90**	**90**	86	88
	15	97	97	97	98	98	99	**100**	99	**100**
B	1	35	39	35	**47**	43	42	40	37	32
	5	79	77	75	76	79	77	**80**	**80**	72
	10	91	92	91	95	94	**96**	90	95	94
	15	92	94	96	95	98	98	98	**99**	**99**
Avg	1	25.5	28.5	25.5	**34.5**	33.5	32	30.5	29.5	23.5
	5	70	69	73.5	74.5	**77**	76	76	72	69
	10	88.5	85.5	87	90.5	92	**93**	90	90.5	91
	15	94.5	95.5	96.5	96.5	98	98.5	99	99	**99.5**

We can get the probability y_i by P300-VIB-Net to determine whether the signal contains P300 component when one row or one column flashes. We can get the coordinates of target character by

$$i_x = \arg \max_{1 \leq i \leq 6} y_i \tag{21}$$

$$i_y = \arg \max_{7 \leq i \leq 12} y_i, \tag{22}$$

where i is the index of row or column range in $[1, 12]$, y_i represents the probability that the signal is P300 while the ith row or column in the matrix is intensificated, i_x and i_y represent the row and column index with most likely P300 signals. The target character is the one at the intersection of i_x-th row and i_y-th column.

In Table 1, the character recognition rate of subject A and B with 1, 5, 10, and 15 epochs are presented. The results of P300-VIB-Net and other models including E-SVM, CNN, MCNN-1, BN3, ERP-Net, PST-DRBM, SAE-ESVM, CM-CW-CNN-ESVM, are presented for comparison. There are two algorithms that combine traditional machine learning and deep learning and achieve impressive recognition performance. Sparse autoencoder (SAE) is used for deep feature extraction and ESVM is used for classification in SAE-ESVM. While in CM-CW-CNN-ESVM, high-level features are extracted by CNN. After that, Fisher ratio (F-ratio) is used to select these features to get optimal features. However, when the number of repeat epochs is relatively small, BN3, PST-DRBM, and ERP-Net achieve the best performance in terms of average character recognition rate. With increasing number of repeat epochs, the character recognition rate of P300-VIB-Net becomes the highest one.

3.4. The Role of VIB Term

As mentioned above, the VIB term represents the mutual information between the input data and intermediate code. Minimizing the VIB term means restricting the information from input data $\mathbf{X}$ to intermediate representation $\mathbf{Z}$ as well as output variable $\mathbf{Y}$. Maximization the cross-entropy between $\mathbf{Z}$ and $\mathbf{Y}$ will force the information flowing from input to predict output. Therefore, when these two processes work simultaneously, the whole loss makes the model focus only on information that is related to output in input data. The resulting model is less disturbed by information that is not related to the output. Hence, VIB could improve the generalization performance. This can be verified by finding the relationship between β and character recognition rate.

The results in Figures 3 and 4 are obtained by manually adjusting β to show the effects of β on the character recognition rate. As shown in Figure 3, the most suitable value of β is 0.01, which make the network attains the best character recognition rate. When increasing β larger than 0.01, the character recognition rate is gradually reduced. This can be explained that with bigger β, more and more information including discriminative information are blocked from input to intermediate code (feature). With extremely large value of β, the information are totally blocked and the character recognition totally failed.

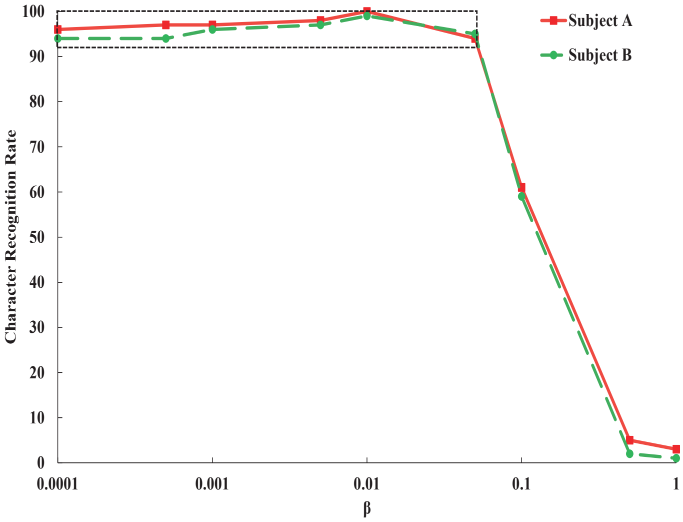


Figure 3. Variation of character recognition rate of subject A and subject B with the changing β of variational information bottleneck (VIB) regularization term.

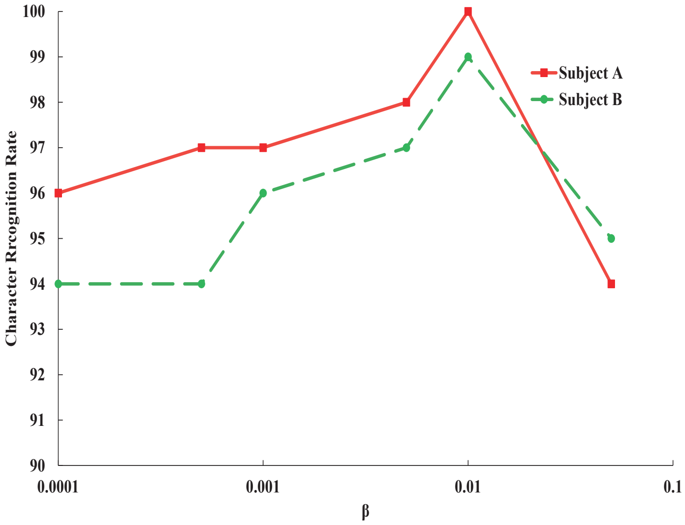


Figure 4. Variation of character recognition rate of subject A and subject B with the weight β changing around optimal value.

To present a more clear illustration, the curve in the dashed box in Figure 3 is magnified and shown in Figure 4. When we increase β from 0.0001 to 0.01, the character recognition rate increases, which indicates that restriction of mutual information between input signals X and code Z could improve the performance of character recognition by blocking label irrelevant information from input data to feature vector. From these results, we could find that whether the weight β is too large or too small, the classification performance will seriously be degraded.

4. Conclusions

Event-related potentials detection is an important problem in BCI research. The low signal-to-noise ratio of EEG signal makes the detection of ERP challengeable. A novel convolutional neural network based on VIB is proposed for P300 detection in this paper. With VIB regularization term added to the traditional cross-entropy loss, the information flowing from input data to intermediate code could be controlled and the label irrelevant information is removed from intermediate variables (features). The experimental results demonstrate that P300-VIB-Net could achieve state-of-art performance in the P300 speller character recognition problem. VIB constraint in P300-VIB-Net enhances the generalization performance of the model. On the other hand, the performance of P300-VIB-Net will deteriorate when the amount of data is relatively small, because it's difficult to estimate information with small amount of data. In our future work, we will explore models based on VIB with other problems in BCI.

Author Contributions: Investigation, H.L.; methodology, Y.Z. and H.L.; project administration, J.X.; software, H.L.; writing-original draft, H.L.; writing-review & editing, supervision, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grants 61836003, 61573150 and 61573152. International Cooperation open Project of State Key Laboratory of Subtropical Building Science, South China University of Technology (2019ZA01).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: http://www.bbci.de/competition/ii/#datasets.

Acknowledgments: The authors acknowledge all of the anonymous reviewers for their constructive comments that helped to improve the quality of this paper.

Conflicts of Interest: As for Guangzhou Galaxy Thermal Energy Incorporated Company, the authors declare no conflict of interest. The company provides no financial support for the research of this article. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Birbaumer, N.; Cohen, L.G. Brain–computer interfaces: Communication and restoration of movement in paralysis. *J. Physiol.* **2007**, *579*, 621–636. [CrossRef] [PubMed]
2. Farwell, L.A.; Donchin, E. Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalogr. Clin. Neurophysiol.* **1988**, *70*, 510–523. [CrossRef]
3. Lin, Z.; Zhang, C.; Zeng, Y.; Tong, L.; Yan, B. A novel P300 BCI speller based on the Triple RSVP paradigm. *Sci. Rep.* **2018**, *8*, 3350. [CrossRef] [PubMed]
4. Jotheeswaran, J.; Sushama, A.S.; Pippal, S. Hybrid video surveillance systems using P300 based computational cognitive threat signature library. *Procedia Comput. Sci.* **2018**, *145*, 512–519. [CrossRef]
5. Li, J.; Yu, Z.L.; Gu, Z.; Wu, W.; Li, Y.; Jin, L. A hybrid network for ERP detection and analysis based on restricted Boltzmann machine. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2018**, *26*, 563–572. [CrossRef] [PubMed]
6. Zhang, J.; Xu, Y.; Yao, L. P300 detection using Boosting Neural Networks with application to BCI. In Proceedings of the ICME International Conference on Complex Medical Engineering, Beijing, China, 23–27 May 2007; pp. 886–891.
7. Kong, W.; Guo, S.; Long, Y.; Peng, Y.; Zeng, H.; Zhang, X.; Zhang, J. Weighted extreme learning machine for P300 detection with application to brain computer interface. *J. Ambient. Intell. Humaniz. Comput.* **2018**, 1–11. [CrossRef]
8. Meng, H.; Wei, H.; Yan, T.; Zhou, W. P300 Detection with Adaptive Filtering and EEG Spectrogram Graph. In Proceedings of the 2019 IEEE International Conference on Mechatronics and Automation (ICMA), Tianjin, China, 4–7 August 2019; pp. 1570–1575.
9. Rosenfeld, J.P. P300 in detecting concealed information and deception: A review. *Psychophysiology* **2020**, *57*, e13362. [CrossRef]
10. Morabbi, S.; Keyvanpour, M.; Shojaedini, S.V. A new method for P300 detection in deep belief networks: Nesterov momentum and drop based learning rate. *Health Technol.* **2019**, *9*, 615–630. [CrossRef]
11. Shojaedini, S.; Morabbi, S.; Keyvanpour, M. A New Method to Improve the Performance of Deep Neural Networks in Detecting P300 Signals: Optimizing Curvature of Error Surface Using Genetic Algorithm. *J. Biomed. Phys. Eng.* **2020**. Available online: https://jbpe.sums.ac.ir/article_46648_d5c552438ba7f346d14990e6e0cc0869.pdf (accessed on 10 November 2020).

12. Oralhan, Z. 3D input convolutional neural networks for P300 signal detection. *IEEE Access* **2020**, *8*, 19521–19529. [CrossRef]
13. Kaper, M.; Meinicke, P.; Grossekathoefer, U.; Lingner, T.; Ritter, H. BCI competition 2003-data set IIb: Support vector machines for the P300 speller paradigm. *IEEE Trans. Biomed. Eng.* **2004**, *51*, 1073–1076. [CrossRef] [PubMed]
14. Rakotomamonjy, A.; Guigue, V. BCI competition III: Dataset II-ensemble of SVMs for BCI P300 speller. *IEEE Trans. Biomed. Eng.* **2008**, *55*, 1147–1154. [CrossRef] [PubMed]
15. Yu, T.; Yu, Z.; Gu, Z.; Li, Y. Grouped automatic relevance determination and its application in channel selection for P300 BCIs. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2015**, *23*, 1068–1077. [CrossRef]
16. Cecotti, H.; Graser, A. Convolutional neural networks for P300 detection with application to brain-computer interfaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 433–445. [CrossRef] [PubMed]
17. Liu, M.; Wu, W.; Gu, Z.; Yu, Z.; Qi, F.; Li, Y. Deep learning based on Batch Normalization for P300 signal detection. *Neurocomputing* **2018**, *275*, 288–297. [CrossRef]
18. Li, J.; Yu, Z.L.; Gu, Z.; Tan, M.; Wang, Y.; Li, Y. Spatial-temporal discriminative restricted Boltzmann machine for event-related potential detection and analysis. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2019**, *27*, 139–151. [CrossRef] [PubMed]
19. Tishby, N.; Zaslavsky, N. Deep learning and the information bottleneck principle. In Proceedings of the 2015 IEEE Information Theory Workshop (ITW), Jeju Island, Korea, 11–15 October 2015; pp. 1–5.
20. Tishby, N.; Pereira, F.C.; Bialek, W. The Information Bottleneck Method. *arXiv*, **2000**, arXiv:physics/0004057. Available online: https://arxiv.org/abs/physics/0004057 (accessed on 24 April 2000).
21. Alemi, A.A.; Fischer, I.; Dillon, J.V.; Murphy, K. Deep Variational Information Bottleneck. *arXiv*, **2017**, arXiv:1612.00410. Available online: https://arxiv.org/abs/1612.00410 (accessed on 23 October 2019).
22. Peng, X.B.; Kanazawa, A.; Toyer, S.; Abbeel, P.; Levine, S. Variational Discriminator Bottleneck: Improving Imitation Learning, Inverse RL, and GANs by Constraining Information Flow. *arXiv*, **2019**, arXiv:1810.00821. Available online: https://arxiv.org/abs/1810.00821 (accessed on 29 December 2018).
23. Achille, A.; Soatto, S. Information dropout: Learning optimal representations through noisy computation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2897–2905. [CrossRef]
24. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv*, **2013**, arXiv:1312.6114. Available online: https://arxiv.org/abs/1312.6114 (accessed on 1 May 2014).
25. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
26. Donchin, E.; Spencer, K.M.; Wijesinghe, R. The mental prosthesis: Assessing the speed of a P300-based brain-computer interface. *IEEE Trans. Rehabil. Eng.* **2000**, *8*, 174–179. [CrossRef] [PubMed]
27. Kundu, S.; Ari, S. P300 based character recognition using sparse autoencoder with ensemble of SVMs. *Biocybern. Biomed. Eng.* **2019**, *39*, 956–966. [CrossRef]
28. Kundu, S.; Ari, S. P300 based character recognition using convolutional neural network and support vector machine. *Biomed. Signal Process. Control.* **2020**, *55*, 101645. [CrossRef]