Multi-View Multi-Label Fine-Grained Emotion Decoding From Human Brain Activity

Kaicheng Fu, Changde Du, Shengpei Wang, and Huiguang He, Senior Member, IEEE

Abstract—Decoding emotional states from human brain activity play an important role in the brain–computer interfaces. Existing emotion decoding methods still have two main limitations: one is only decoding a single emotion category from a brain activity pattern and the decoded emotion categories are coarse-grained, which is inconsistent with the complex emotional expression of humans; the other is ignoring the discrepancy of emotion expression between the left and right hemispheres of the human brain. In this article, we propose a novel multi-view multi-label hybrid model for fine-grained emotion decoding (up to 80 emotion categories) which can learn the expressive neural representations and predict multiple emotional states simultaneously. Specifically, the generative component of our hybrid model is parameterized by a multi-view variational autoencoder, in which we regard the brain activity of left and right hemispheres and their difference as three distinct views and use the product of expert mechanism in its inference network. The discriminative component of our hybrid model is implemented by a multi-label classification network with an asymmetric focal loss. For more accurate emotion decoding, we first adopt a label-aware module for emotion-specific neural representations and then model the dependency of emotional states by a masked self-attention mechanism. Extensive experiments on two visually evoked emotional datasets show the superiority of our method.

Index Terms—Fine-grained emotion decoding, multi-label learning, multi-view learning, product of experts (PoEs), variational autoencoder.

I. INTRODUCTION

EMOTION decoding from visually evoked human brain activity measured by functional magnetic resonance imaging (fMRI) is an emerging research area and plays an important role in the brain–computer interfaces (BCIs). Most of the previous emotion decoding studies [1], [2], [3] considered very coarse-grained and limited emotion categories, such as positive, neutral, and negative, which cannot represent the complex emotions we experience in daily life. Recent studies [4], [5] have proven that the emotions expressed by humans constitute a high-dimensional semantic space which motivates us to study more fine-grained emotion decoding.

In addition, a more obvious limitation in previous decoding methods is that they only predicted one emotion category for one brain activity pattern [2], [3], [6], which means they only regard emotion decoding as a multiclass, rather than multi-label, decoding problem. However, there is no doubt that multiple human emotions can be elicited simultaneously by emotion stimuli. Fig. 1 shows three video screenshots of emotion-inducing movie clips used in our experimental dataset. Taking the third scene fighting as an example, we may feel angry and disgust for this uncivilized behavior while we are also surprised and confused with the reason for fighting. Besides, it is fearful and anxious for us when we see people fighting on the street, which will make us sad about the injury as well.

On the other hand, most previous emotion decoding methods adopted voxel-based decoding [6], [7], [8], which suffers from severe overfitting due to high data dimensionality, small sample size, and low signal-to-noise ratio. In order to alleviate overfitting, they barely used simple linear models [6], [9] or generalized linear models (GLMs) [8] for decoding, leading to the weak expressive ability of the model. Although some of them have involved the use of brain regions of interest (ROIs) [6] or functional cortical [8] signals for emotion decoding, they essentially are still subject to voxel-based decoding in which advanced deep learning algorithms are not easy to use.

More importantly, for visually evoked emotion decoding, neuroscience research studies have revealed the discrepancy in emotional expression between the left and right hemispheres despite the common visual input [10], [11], which can be regarded as the prior knowledge for developing models. Many previous studies have proven that exploiting the bihemisphere discrepancy is capable of improving the performance of emotion decoding by using electroencephalogram (EEG)-based data [3], [12]. They employ independent networks to learn...
features for the left and right hemispheres and then fuse them for further decoding tasks. Inspired by these, we regard the left and right hemispheres, and their difference as multiple information resources (views) to explore the interaction between the two hemispheres and take advantage of a product of experts (PoEs) mechanism [13] for feature level fusion.

Taking the above-mentioned limitations into account, we propose a novel multi-view multi-label hybrid model called multilabel classification network with brain variational autoencoder (ML-BVAE) which is, to the best of our knowledge, the first algorithm for fine-grained emotion decoding from human brain activity and can offer great potential for a fine-grained emotion BCI system. To this end, we first propose an fMRI dimensionality reduction method utilizing ROI pooling which can effectively realize noise suppression and dimensionality reduction of the original signal. Then, we raise a multi-view multi-label hybrid model which enables us to not only achieve accurate emotion decoding but also investigate the relationship between emotion categories. An overview of our experimental paradigm can be found in Fig. 2. Our main contributions can be summarized as follows.

1) We propose an effective fMRI dimensionality reduction method called ROI pooling which can suppress noise and alleviate overfitting when complex models are used for emotion decoding.

2) We design a tailored BVAE with the PoE mechanism for learning expressive multi-view neural representations while taking the bihemisphere discrepancy into account, which is of great importance to emotion decoding studies.

3) We develop a novel multi-label classification network to exploit the relationship between brain activity and emotion labels and the dependency across emotion labels. For the first time, we realize multi-label emotion decoding from brain activity.

4) Sufficient experiments on two visually evoked emotional datasets demonstrate that ML-BVAE outperforms the compared methods on several evaluation metrics and can achieve fine-grained emotion decoding of up to 80 emotions. Our source code is available in https://github.com/KaichengFu1997/ML-BVAE.

II. RELATED WORK

A. Emotion Decoding From Brain Activity

In recent years, decoding emotional states from brain activity has been widely studied. Most existing studies performed emotion decoding based on either fMRI or EEG signals. In fMRI-based emotion decoding, [14] decoded four affective dimensions with logistic regression with percent signal change features; [6] used linear regression (LR) to make predictions separately for each of the 34 emotion categories, and each of the 14 affective dimensions based on the responses of ROI voxels; [8] utilized support vector machine (SVM) to build the relationship between four emotion categories and brain activity in the auditory and motor cortex. Although they have made some progress in emotion decoding, they can only use simple linear models limited by high-dimensional features. In EEG-based emotion decoding, [15] used dynamical graph convolutional networks (GCNs) to model the multichannel EEG features and performed EEG emotion classification; [16] and [17] used contrastive learning and domain adaptation, respectively, for cross-subject EEG-based emotion recognition; [3] and [12] both took the discrepancy of bihemisphere into consideration for emotion decoding from EEG-based emotional datasets. Despite the complex models and scenarios, they still performed multiclass emotion decoding with coarse-grained emotion categories. Compared with EEG-based methods, fMRI-based studies can perform more fine-grained emotion decoding due to its higher spatial resolution but easily suffers from overfitting. Our method belongs to fMRI-based emotion decoding and we perform
dimensionality reduction to the original fMRI data which can alleviate overfitting when complex models, such as deep neural networks (DNNs) are used. We also achieve multi-label emotion decoding which means decoding a set of emotional states simultaneously.

B. Multi-Label Emotion Classification

Recently, researchers have noticed the complexity of human emotion expression and have committed to studying the problem of multi-label emotion classification. Fei et al. [18] proposed latent emotion memory (LEM) for learning latent emotion distribution and utilized bidirectional GRU to learn emotion coherence so as to realize the multi-label emotion classification of text data. Wang et al. [20] exploited graph neural networks (GNNs) with heterogeneous hierarchical message passing for multimodal multi-label emotion classification with textual, visual, and acoustic modalities. However, these studies all belong to coarse-grained emotion classification which involves only 12 emotion categories at most [20]. It is difficult for these studies to truly reflect the complex emotion categories of humans. Based on this disadvantage, [21] built a fine-grained emotion dataset with 24 types of emotions using a large collection of tweets; [22] collected another fine-grained emotion dataset named GoEmotions with 58k English Reddit comments, which involves 27 emotion categories. Furthermore, [23] proposed a model called Seq2Emo with a bidirectional decoder for multi-label emotion classification on GoEmotions. However, the abovementioned studies are all based on multimedia data, rather than recognizing human emotions based on physiological signals which are also known as emotion decoding. Different from these studies, for the first time, we achieve fine-grained emotion decoding from human brain activity with fMRI data, which can decode up to 80 emotion categories.

C. Multi-View Multi-Label Learning

Multi-label learning has many applications in text categorization [24], bioinformatics [25], Web mining [26], and so on. In recent years, multi-label learning has also been used in semantic decoding from brain activity [27], [28], [29]. However, the study of multi-label emotion decoding is still lacking. Besides, in multi-view learning, the information in some views is useful to handle the weakness of other views. Furthermore, multi-view learning can naturally be embedded into multi-label learning tasks to improve classification performance. Models in multi-view multi-label learning can be divided into generative models and discriminative models. Sun and Zong [30] proposed a Gaussian mixture VAE placing a conditional Bernoulli mixtures distribution on the labels y, which belongs to a probabilistic generative model. Zhang et al. [31] used matrix factorization to uncover the latent patterns among different views for more accurate multi-label classification. Wu et al. [32] exploited shared subspace for fusing multi-view representations and also took view-specific information into consideration. The above two models are both attributed to the discriminative model. Besides, [33] proposed deep hybrid models which can bridge the discriminative and generative models in which the former often attain higher predictive accuracy, while the latter is more strongly regularized and can obtain more expressive representations. The discriminative component of the hybrid model pays more attention to the predicted probability \( p(y|x) \), rather than the distribution of the label \( p(y) \). Inspired by this, we propose a multi-view multi-label hybrid model for learning expressive neural representations and predicting multiple emotional states accurately.

III. METHODOLOGY

A. Problem Definition

Formally speaking, we have a multi-view multi-label dataset with \( N \) samples \( D = \{ (x^l_i, x^r_i, y_i) \}_{1 \leq i \leq N} \), where \( x^l_i \in \mathbb{R}^{D_l} \) and \( x^r_i \in \mathbb{R}^{D_r} \) are feature vectors of the left and right hemisphere views, respectively, \( y_i \) is the difference view between the left and right hemispheres (i.e., \( x^l_i - x^r_i \)), \( y_i \in \mathbb{R}^c \) is label vector in which \( y_{ic} \in \{0, 1\} \). \( y_{ic} = 1 \) means that emotion label \( c \) is relevant to the brain activity \( i \). We introduce shared latent variables \( z \), which can be regarded as multi-view neural representations and are also capable of predicting multiple emotion labels. For simplicity, we rewrite the joint of three views \( (x^l, x^r, x^d) \) as \( x \). The task of our hybrid model ML-BVAE is to learn a joint model \( p(x, y, z) \) from \( D \). We assume that \( p(x, y, z) \) has a parametric form specified by the decomposition

\[
p(x, y, z) = p(y|x, z)p(x, z)
\]

in which \( p(x, z) \) and \( p(y|x, z) \) are, respectively, the latent variables generative component and the discriminative component of ML-BVAE. In practice, we assume that the latent variables \( z \) contain all the information of \( x \). Therefore, we simplify the discriminative component as \( p(y|z) \).

The standard approach for training ML-BVAE is to maximize the marginal likelihood

\[
\log p(x, y) = \log \int_z p(x, y, z) = \log \int_z p(y|x, z)p(x, z).
\]

For the intractable integral, we apply variational inference to obtain a tight lower bound of the marginal likelihood with Jensen’s inequality

\[
\log \int_z p(y|x, z)p(x, z)
= \log \int_z p(y|z)p(x, z)
= \log \int_z \frac{p(y|z)p(x, z)}{q(z|x)}q(z|x)
\geq \mathbb{E}_{q(z|x)}[\log p(x, z) - \log q(z|x) + \log p(y|z)]
= \mathcal{L}_{\text{ELBO}} + \mathbb{E}_{q(z|x)}[\log p(y|z)]
\]

in which \( q(z|x) \) is the variational distribution. The lower bound of ML-BVAE has two terms: \( \mathcal{L}_{\text{ELBO}} \) denotes the evidence lower bound (ELBO) of the generative component, while \( \mathbb{E}_{q(z|x)}[\log p(y|z)] \) can measure the classification error of the
Then, we concatenate the features of ROI in each hemisphere to voxels in each subvolume as the feature of this subvolume, according to its coordinates, then split the volume evenly into brain areas (ROIs). In order to further extract the features of ROI represents a voxel. Voxels in this ROI are placed in a 3-D volume which is split evenly into several subvolumes. Then, the brain activity of voxels in each subvolume is averaged.

discriminative component (the larger the value, the smaller the classification error).

After training, we can obtain the multi-label classification probability by

\[
p(y|x) = p(y|z)p(z|x) \approx p(y|z)q(z|x). \tag{4}
\]

B. Overview

The proposed approach involves three key components: fMRI dimensionality reduction with ROI pooling (see Fig. 3), multi-view neural representations learning with BVAE, and multi-label learning with emotion-specific neural representations, in which the second and the third components shown in Fig. 4 serve as the generative and the discriminative component of ML-BVAE, respectively. In the following, we will introduce them separately.

C. fMRI Dimensionality Reduction With ROI Pooling

Using an fMRI signal directly for voxel-based decoding will introduce considerable noise and easily cause overfitting due to high data dimensionality, small sample size, and low signal-to-noise ratio [35, 36, 37]. Therefore, we first use brain atlas [34], [38] to divide the whole brain cortical into multiple brain areas (ROIs). In order to further extract the features of each ROI, we place the voxels of each ROI in a 3-D volume according to its coordinates, then split the volume evenly into several subvolumes and calculate the average brain activity of voxels in each subvolume as the feature of this subvolume, as illustrated in Fig. 3 (take eight subvolumes for example). Then, we concatenate the features of ROI in each hemisphere to obtain ROI pooling features \(x^l\) and \(x^r\). This method can realize noise suppression and dimensionality reduction of the input signal. We set the number of features in each ROI as an optional hyperparameter named \(N_{ROIF}\).

D. Multi-View Neural Representations Learning With BVAE

In BVAE, we assume that the left hemisphere view \(x^l\), the right hemisphere view \(x^r\), and the difference view \(x^d\) can be generated by the common latent variables \(z\), which can be regarded as the common features among the three distinct views. That is, we assume a generative model of the form

\[
p_\theta(x, z) = p(z)p_\theta(x^l|z)p_\theta(x^r|z)p_\theta(x^d|z) \tag{5}
\]

where \(\theta = [\theta_1, \theta_r, \theta_d]\). Then, the three views ELBO, written as \(\mathcal{L}_{ELBO_m}\), is equal to

\[
\begin{align*}
&\mathbb{E}_{q_\phi}[\lambda_l \log p_\theta(x^l|z) + \lambda_r \log p_\theta(x^r|z) + \lambda_d \log p_\theta(x^d|z)] \\
&- \beta \text{KL}[q_\phi(z|x)||p(z)] \tag{6}
\end{align*}
\]

where the \(L_{AE}\) term denotes the reconstruction error of the three views and the \(D_{KL}\) term regularizes the encoder by minimizing the Kullback–Leibler (KL) divergence between the joint approximate posterior \(q_\phi(z|x)\) and the prior \(p(z)\). In practice, \(\beta\) is slowly annealed to 1 [39, 40] to form a valid ELBO. \(\lambda_l, \lambda_r, \lambda_d\) are hyperparameters that balance the reconstruction error among the three views.

The key issue for training BVAE is to specify the joint approximate posterior distribution \(q_\phi(z|x)\). According to the conditional independence assumptions in the generative model, the true joint posterior distribution can be written as

\[
p(z|x) \propto \frac{p(x|z)p(z)}{p(x)} = \frac{p(z)p(x^l|z)p(x^r|z)p(x^d|z)}{p(x)}
\]

\[
= \frac{p(z)p(x^l|z)p(x^l)}{p(x)p(z)} = \frac{p(x^l|z)p(x^l)p(x^r|z)p(x^r)}{p(x)p(z)}
\]

\[
= \frac{p(z)p(x^l|z)p(x^l)p(x^r|z)p(x^d|z)}{p^2(z)}
\]

\[
\propto \frac{p(z)^2p(x^l|z)p(x^r|z)p(x^d|z)}{p(z)}
\]

which means the joint posterior is a product of individual posterior from each view, with an additional quotient by the square of prior. For numerical stability, if we approximate the single-view true posterior with \(q_{\phi_l}(z|x^l) = \tilde{q}_{\phi_l}(z|x^l)p(z)\), \(q_{\phi_r}(z|x^r) = \tilde{q}_{\phi_r}(z|x^r)p(z)\), and \(q_{\phi_d}(z|x^d) = \tilde{q}_{\phi_d}(z|x^d)p(z)\), where \(\tilde{q}_{\phi_l}(z|x^l)\), \(\tilde{q}_{\phi_r}(z|x^r)\), and \(\tilde{q}_{\phi_d}(z|x^d)\) are the underlying inference network of each view, we can avoid the quotient term by

\[
p(z|x) \approx \frac{p(z|\tilde{q}_{\phi_l}(z|x^l)\tilde{q}_{\phi_r}(z|x^r)\tilde{q}_{\phi_d}(z|x^d))}{p^2(z)}
\]

\[
= \frac{[\tilde{q}_{\phi_l}(z|x^l)p(z)][\tilde{q}_{\phi_r}(z|x^r)p(z)][\tilde{q}_{\phi_d}(z|x^d)p(z)]}{p^2(z)}
\]

\[
= p(z)\tilde{q}_{\phi_l}(z|x^l)\tilde{q}_{\phi_r}(z|x^r)\tilde{q}_{\phi_d}(z|x^d). \tag{8}
\]

That is, the encoder of BVAE can be written as

\[
q_\phi(z|x) = p(z)\tilde{q}_{\phi_l}(z|x^l)\tilde{q}_{\phi_r}(z|x^r)\tilde{q}_{\phi_d}(z|x^d) \tag{9}
\]

in which \(\phi = \{\phi_l, \phi_r, \phi_d\}\). In other words, the joint approximately posterior distribution is a product of prior and three single-view posterior distributions called PoE [41]. When these four distributions are all Gaussian which means \(p(z) = \mathcal{N}(\mu_0, \Sigma_0)\), \(\tilde{q}_{\phi_l}(z|x^l) = \mathcal{N}(\mu_l, \Sigma_l)\), \(\tilde{q}_{\phi_r}(z|x^r) = \mathcal{N}(\mu_r, \Sigma_r)\), and \(\tilde{q}_{\phi_d}(z|x^d) = \mathcal{N}(\mu_d, \Sigma_d)\), there is an analytical solution of the product distribution acquired earlier [13]

\[
q_\phi(z|x) = \mathcal{N}(\mu, \Sigma) \tag{10}
\]

which means \(q_\phi(z|x)\) is Gaussian with

\[
\mu = (\Sigma_0^{-1}\mu_0 + \Sigma_l^{-1}\mu_l + \Sigma_r^{-1}\mu_r + \Sigma_d^{-1}\mu_d)\Sigma
\]

\[
\Sigma = (\Sigma_0^{-1} + \Sigma_l^{-1} + \Sigma_r^{-1} + \Sigma_d^{-1})^{-1} \tag{11}
\]
whose proof can be found in Appendix A. In practice, we assume that \( p(z) \) is spherical Gaussian prior with \( \mu_0 = \theta \) and \( \Sigma_0 = I \).

BVAE supports single-view input and generates three views at the same time due to the product property of its encoder. In order to model the information intra- and inter-view the three views, in addition to the ELBO defined by (6), we define another three ELBOs

\[
\mathcal{L}_{\text{ELBO}_\text{LAE}} = \mathbb{E}_{q_0}[\lambda_z \log p_0(x^p|z)] + \mathbb{E}_{q_0}[\lambda_r \log p_0(x^r|z)] + \mathbb{E}_{q_0}[\lambda_d \log p_0(x^d|z)] - \beta KL[q_0(z|x^p)||p(z)]
\]

\[
\mathcal{L}_{\text{ELBO}_\text{AE}} = \mathbb{E}_{q_0}[\lambda_z \log p_0(x^p|z)] + \mathbb{E}_{q_0}[\lambda_r \log p_0(x^r|z)] + \mathbb{E}_{q_0}[\lambda_d \log p_0(x^d|z)] - \beta KL[q_0(z|x^r)||p(z)]
\]

\[
\mathcal{L}_{\text{ELBO}_\text{LR}} = \mathbb{E}_{q_0}[\lambda_z \log p_0(x^p|z)] + \mathbb{E}_{q_0}[\lambda_r \log p_0(x^r|z)] + \mathbb{E}_{q_0}[\lambda_d \log p_0(x^d|z)] - \beta KL[q_0(z|x^d)||p(z)]
\]

where \( \mathcal{L}_{\text{ELBO}_\text{LAE}}, \mathcal{L}_{\text{ELBO}_\text{AE}}, \) and \( \mathcal{L}_{\text{ELBO}_\text{LR}} \) mean only using the left, right, and difference single view as the input to the BVAE, respectively. The \( \mathcal{L}_{\text{AE}} \) terms with the same subscript are reconstruction errors intraview and the \( \mathcal{L}_{\text{AE}} \) terms with different subscripts are reconstruction errors interview. \( D_{\text{AE}}, D_{\text{KL}}, \) and \( D_{\text{KL-d}} \) are KL divergence across the three views encoder and the prior, respectively. Above all, the total ELBO of BVAE can be written as (actually, \( \mathcal{L}_{\text{ELBO}_\text{LAE}} \) is the lower bound on the joint loglikelihood, and another three ELBOs are useful for training individual encoders [41])

\[
\mathcal{L}_{\text{ELBO}} = \mathcal{L}_{\text{ELBO}_\text{LAE}} + \mathcal{L}_{\text{ELBO}_\text{AE}} + \mathcal{L}_{\text{ELBO}_\text{LR}} + \mathcal{L}_{\text{ELBO}_\text{D}}.
\]

E. Multi-Label Learning With Emotion-Specific Neural Representations

In order to further model the relationship between multi-view neural representations and emotion labels, we use a label-aware module [42] to get a set of emotion-specific neural representations from \( z \) by utilizing the distinct discriminative properties of each label. First, we apply classifiers that can be implemented by fully connected layers to the multi-view neural representations \( z \in \mathbb{R}^D \)

\[
p_{\omega_1}(y|z) = \text{sigmoid}[f(x; \omega_1)]
\]

where \( \omega_1 \in \mathbb{R}^{C \times D} \) represents the parameters of these classifiers and \( C \) is the number of emotion labels. Each classifier \( \omega_1^c \in \mathbb{R}^D \) extracts the information corresponding to emotion class \( c \) and predicts the probability of emotion class \( c \) from the brain activity. Then, the emotion-specific neural representations can be obtained via

\[
Z = \text{repeat}(z) \odot \omega_1 \in \mathbb{R}^{C \times D}
\]

in which \( \text{repeat}(z) \) indicates the operation of copying \( z \in \mathbb{R}^D \) for \( C \) times and \( \odot \) denotes the Hadamard product. In this way, \( Z \in \mathbb{R}^D \) captures the information related to the emotion class \( c \).

To further model the dependency of emotion labels, we make use of a masked self-attention operation [43] across the emotion-specific neural representations. In this module, emotion-specific neural representations \( Z \) generate a set of query, key, and value matrices \( (Q, K, V) \) by three linear transformations, whose parameters are \( (\omega_q, \omega_k, \omega_h) \).
Then, a $C \times C$ attention matrix $A$ is obtained as follows:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d'}} + \text{Mask}\right) \tag{18}$$

where $d'$ is dim of the query and key, $\text{Mask} \in \mathbb{R}^{C \times C}$ is the label graph adjacency matrix [20], in which $\text{Mask}(j, k)$ is the frequency of emotion $j$ co-occurrence with emotion $k$ in the training ground truth, $N_{j,k}$ is the number of training samples that include both emotion $j$ and $k$, and $N_j$ is the number of training samples, in which emotion $j$ occurs. By this, we apply an emotion labelwise correlation map $A$ onto the emotion-specific neural representations. Then, we can get the refined emotion-specific neural representations $Z$ by

$$Z = AV \tag{20}$$

Then, we use another fully connected layer to the refined emotion-specific neural representations

$$p_{\omega_2}(y|z) = \text{sigmoid}[g(Z; \omega_k)] \tag{21}$$

where $\omega_2 = \{\omega_q, \omega_k, \omega_h, \omega_y\}$. The final output of the ML-BVAE is

$$p(y|z) = p_{\omega_1}(y|z) + p_{\omega_2}(y|z) \tag{22}$$

in which $\omega = \{\omega_1, \omega_2\}$.

### F. Loss Function

In the supervised scenario, to more effectively address the sample imbalance problem, we substitute the log-likelihood in the classification error term with asymmetric focal loss [44], [45]

$$\log p_{\omega_1}(y|z) \approx \frac{1}{C} \sum_{c=1}^{C} \frac{1}{y_c + \gamma - y_c} \log \left(\frac{p_{\omega_1}^c}{p_{\omega_1}}\right), \; y_c = 1$$

$$\log \left(1 - p_{\omega_1}^c\right), \; y_c = 0 \tag{23}$$

where $y_c$ is the label that indicates whether brain activity $x$ has emotional state $c$ and $p_{\omega_1}^c$ is the $c$th component of $p_{\omega_1}(y|z)$. $\gamma^+$ and $\gamma^-$ are hyperparameters that balance the loss of positive and negative labels. Above all, the lower bound of the marginal likelihood can be

$$\mathcal{L}(\phi, \theta, \omega) = \lambda_1 \mathcal{ELBO}(\phi, \theta) + \lambda_2 \mathbb{E}_{q_\phi}(\mathbb{E}_{q_\theta}(\mathbb{E}_{q_X}(\log p_{\omega_1}(y|z)))) \tag{24}$$

where $\lambda_1$ and $\lambda_2$ are hyperparameters that balance the loss between the generative and the discriminative component. After training, given a test brain activity pattern $x$, we can predict the probability of multiple emotional states by

$$p(y|x) = p_{\omega^*}(y|z)q_{\phi^*}(z|x) \tag{25}$$

where $\omega^*$ and $\phi^*$ are optimal parameters of the multi-label classification network and the encoder of BVAE, respectively.

### IV. Experiments

#### A. Datasets

The properties of the two visually evoked emotional datasets used in our experiment are summarized in Table I. In the following, we will introduce them in detail.

### Table I

| Dataset     | #Subjects | #Instances | #Voxels | #ROIs | #Emotions |
|-------------|-----------|------------|---------|-------|-----------|
| MEMO27     | 5         | 2196       | 115070  | 360   | 27        |
| MEMO80     | 8         | 5400       | 83695   | 148   | 80        |

1) MEMO27: MEMO27 is a fine-grained multi-label emotion decoding dataset based on a publicly available fMRI dataset, which contains the blood-oxygen-level dependent (BOLD) responses of five subjects. fMRI data were collected using a 3T Siemens scanner with a multiband gradient echo-planar imaging (EPI) sequence (TR, 2000 ms; TE, 43 ms; flip angle, 80°; FOV, 192 mm × 192 mm; voxel size, 2 mm × 2 mm × 2 mm; number of slices, 76; multiband factor, 4) [6]. Subjects were presented with 2196 videos whose durations ranged from 0.15 to 90 s [6]. These videos can provoke a variety of emotions. Each video was voted by multiple raters across 34 emotion categories and the voting ratio of each emotion category was used as the rating of that emotion category. The fMRI data were preprocessed and averaged with each video stimulus [6] which means the brain activity of one voxel is a scalar for a video stimulus. We use the HCP360 atlas defined in [38] to segment the cerebral cortex of each subject in MEMO27 into 360 ROIs (180 ROIs per hemisphere), as illustrated in [6].

As for the emotion ratings, we first choose 27 of 34 emotion categories according to [4]. Then, we set a threshold of 0.1 to construct an emotion label matrix from emotion ratings. The average number of emotion labels is 4.64 and the corresponding label density is 0.172.

2) MEMO80: MEMO80 is a more fine-grained multi-label emotion decoding dataset, which contains the BOLD responses of eight subjects. fMRI data were collected using a 3T Siemens scanner with a multiband gradient EPI sequence (TR, 2000 ms; TE, 30 ms; flip angle, 62°; FOV, 192 mm × 192 mm; voxel size, 2 mm × 2 mm × 2 mm; number of slices, 72; multiband factor, 3) [5]. Subjects were presented with emotion-inducing audiovisual movies over a period of three hours. For each of the one-second movie scenes, the movie was annotated with respect to 80 emotions by multiple raters. The fMRI data and emotion ratings were preprocessed, as illustrated in [5]. According to [5], each sampling moment of fMRI can be regarded as a sample which leads to 5400 samples in total for one subject. We use Destrieux atlas defined in [34] to segment the cerebral cortex of each subject in MEMO80 into 148 ROIs (74 ROIs per hemisphere), as illustrated in [5].

As for the emotion ratings, we exploit all 80 emotion categories for completeness and diversity. Then, we set a threshold of 0.5 to construct an emotion label matrix from emotion ratings. The average number of emotion labels is 17.9 and the corresponding label density is 0.224.

---

1available at https://doi.org/10.6084/m9.figshare.11988351.v1

2Restricted by copyright, MEMO80 is available once the corresponding author of [5] gives permission.
B. Compared Methods

The performance of ML-BVAE is compared with seven methods, including a linear model, three single-view multi-label learning methods and two multi-view multi-label learning methods: 1) LR [6]: a traditional fMRI decoding method. In order to be consistent with the settings of [6], we trained a linear model for each emotion category with all cortical voxels. Then, we binarized the predicted ratings using the same threshold as the label matrix construction. 2) Benchmark: a multilayer perceptron (MLP) where the number of parameters is no less than ML-BVAE. The input is the concatenation of all view features. 3) CA2E [46]: the first DNNs-based multi-label learning model which can learn deep latent space by mapping the feature and label jointly. The hyperparameter $\alpha$ is set to 2.0 and the dimension $l$ of the latent space is set to 128 and 64 for MEMO27 and MEMO80, respectively. 4) CA2E [47]: a multi-label learning method based on GCNs. We adopt MLP used in a benchmark for feature extraction. 5) SIMM [32]: a multi-view multi-label learning method, which leverages shared subspace exploitation and view-specific information extraction. The hyperparameters $\alpha$ and $\beta$ are both set to 0.1. 6) GARDIS [48]: a multi-view multi-label learning method based on graph and spectral clustering. The hyperparameters $\{\alpha, \gamma, \eta, k\}$ are set to $\{0.9, 0.1, 0.1, 8\}$, respectively.

C. Hyperparameters Setting

For the $\ell_{ELBO}$ in (15), $\{\lambda_1, \lambda_r, \lambda_d\}$ are all set to 1 and $\beta$ is slowly annealed to 1 according to $\beta(i, j) = \frac{j + (i - 1)M + 1}{\alpha M}$ (26), where $i$ and $j$ are the epoch and batch index, respectively, and $M$ is the batch size which is set to 100. $\alpha$ is the annealing ratio, which is set to 100. $\{\gamma^+, \gamma^-\}$ in asymmetric focal loss are set to 0 and 1, respectively, according to the original reference. $\lambda_1/\lambda_2$ in (24) is searched in $\{0.001, 0.01, 0.1, 1, 2, 5, 10, 15, 20\}$. $N_{ROIF}$ is searched in $\{8, 27, 64, 125, 216, 343\}$. The dimensionality of neural representations $z$ is set to 64 in both datasets. We train the model using the Adam optimizer, with weight decay of 0.1, $\beta_1 = 0.9$, $\beta_2 = 0.9999$, and a learning rate of $10^{-4}$.

D. Evaluation Metrics

We employ six widely used metrics for measuring multi-label emotion decoding performance, including one-error (OneE), ranking loss (RL), micro F1 (miF1), macro F1 (maF1), example-based average precision (e-AP) [49], and mean average precision (mAP) [42]. For OneE and RL, the smaller the values, the better the performance. For the other four metrics, the larger the values the better the performance. In these metrics, OneE, RL, and e-AP are example-based metrics that can evaluate the model performance on each test example separately and then return the mean value across the test set, while miF1, maF1, and mAP are label-based metrics, which are able to evaluate the model performance on each emotion category separately and then return the mean value across all emotion categories. Among these, miF1, maF1, and mAP are the primary metrics considering their comprehensiveness.

E. Experimental Results

1) Multi-Label Emotion Decoding Performance: For MEMO27, tenfold cross-validation is performed for each subject where the mean results of each metric are recorded for all compared approaches. For MEMO80, we follow the setting in [5] leading to 3600 samples in the training dataset and 1800 samples in the test dataset, which can ensure that all the stimuli used in the test dataset have not been seen during training.

Detailed experimental results are reported in Tables II and III. Compared with other methods, ML-BVAE shows obvious superiority in both datasets. It outperforms all the compared methods in all six metrics regarding a total of five subjects in MEMO27 and achieves the best performance in most cases in MEMO80. It is also noticeable

| Subject | Method  | OneE  | RL   | miF1  | maF1  | e-AP   | mAP   |
|---------|---------|-------|------|-------|-------|--------|-------|
|         | LR [6]  | 0.546 | 0.506 | 0.197 | 0.155 | 0.309  | 0.565 |
|         | Benchmark | 0.315 | 0.200 | 0.449 | 0.336 | 0.609  | 0.438 |
|         | CA2E [46] | 0.642 | 0.255 | 0.428 | 0.345 | 0.519  | 0.364 |
|         | ML-GCN [47] | 0.307 | 0.019 | 0.645 | 0.331 | 0.606  | 0.421 |
|         | SIMM [32] | 0.360 | 0.217 | 0.396 | 0.239 | 0.573  | 0.399 |
|         | GARDIS [48] | 0.315 | 0.230 | 0.229 | 0.120 | 0.530  | 0.394 |
|         | ML-BVAE | 0.308 | 0.183 | 0.564 | 0.338 | 0.529  | 0.451 |
|         |          |       |      |       |       |        |       |
|         | LR [6]  | 0.552 | 0.493 | 0.192 | 0.141 | 0.317  | 0.251 |
|         | Benchmark | 0.320 | 0.182 | 0.457 | 0.330 | 0.619  | 0.447 |
|         | CA2E [46] | 0.427 | 0.249 | 0.427 | 0.327 | 0.527  | 0.383 |
|         | ML-GCN [47] | 0.309 | 0.181 | 0.306 | 0.324 | 0.631  | 0.441 |
|         | SIMM [32] | 0.363 | 0.205 | 0.381 | 0.230 | 0.578  | 0.414 |
|         | GARDIS [48] | 0.327 | 0.221 | 0.208 | 0.110 | 0.572  | 0.391 |
|         | ML-BVAE | 0.384 | 0.377 | 0.515 | 0.310 | 0.534  | 0.466 |

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
TABLE III
MULTI-LABEL EMOTION DECODING PERFORMANCE OF EACH COMPARED APPROACH IN TERMS OF ONE-ERROR, RANKING LOSS, MICRO F1, MACRO F1, EXAMPLE-BASED AVERAGE PRECISION, AND MEAN AVERAGE PRECISION IN MEMO80. ↑ (↓) INDICATES THE LARGER (SMALLER) THE VALUE, THE BETTER THE PERFORMANCE.

| Subject | Method | One-error | RL1 | miF1 | maF1 | e-AP | mAP |
|---------|--------|-----------|-----|------|------|------|-----|
| Subject1 | LR [5] | 0.536 | 0.393 | 0.282 | 0.281 | 0.382 | 0.359 |
| | Benchmark | 0.446 | 0.316 | 0.227 | 0.205 | 0.462 | 0.427 |
| | CAZE [46] | 0.566 | 0.389 | 0.338 | 0.348 | 0.394 | 0.344 |
| | ML-GCN [47] | 0.460 | 0.303 | 0.242 | 0.206 | 0.472 | 0.435 |
| | SSDM [32] | 0.482 | 0.316 | 0.263 | 0.242 | 0.455 | 0.454 |
| | GARDES [48] | 0.490 | 0.352 | 0.133 | 0.134 | 0.437 | 0.419 |
| | ML-BVAE | 0.410 | 0.295 | 0.184 | 0.181 | 0.402 | 0.374 |

TABLE IV
FRIEDMAN STATISTICS \( F_F \) IN TERMS OF EACH METRIC AND THE CRITICAL VALUE AT 0.05 SIGNIFICANCE LEVEL. (# COMPARED ALGORITHMS \( k = 6 \), # SUBJECTS \( N = 13 \) FOR MEMO27 AND MEMO80.)

| One-error | RL1 | miF1 | maF1 | e-AP | mAP | Critical value |
|-----------|-----|------|------|------|-----|--------------|
| 75.306 | 243.328 | 40.616 | 60.651 | 106.300 | 118.719 | 2.368 |

Among the compared approaches, LR shows the least superiority for multi-label emotion decoding tasks, especially in MEMO27 due to its relatively low representation learning ability and neglect of label correlations. ML-GCN and Benchmark are two relatively strong baselines, which can rank second or third in most cases. CA2E shows some superiority with regard to miF1 and maF1. Surprisingly, SSDM and GARDES, as two multi-view multi-label learning methods, perform worse than other single-view baselines and far worse than ML-BVAE. Of them, GARDES achieves the worst results perhaps because it only exploits deep learning. This shows that the emotion decoding task needs to learn more expressive neural representations, which are consistent with the original intention of our hybrid model design.

Furthermore, we adopt the Friedman test [50] for statistical tests in order to discuss the relative performance among the compared methods. If there are \( k \) algorithms and \( N \) datasets (the number of subjects in our experiment), we make use of the average ranks of algorithms \( R_j = \frac{1}{N} \sum_i r_{ij} \) for the Friedman test, in which \( r_{ij} \) is the ranks of the \( j \)th algorithm on the \( i \)th dataset. If the null hypothesis is that all the algorithms have the equivalent performance, the Friedman statistic \( F_F \) which will satisfy the F-distribution with \( k-1 \) and \( \left( k-1 \right) \left( N-1 \right) \) degrees of freedom can be written as

\[
F_F = \frac{(N - 1) \chi^2_F}{N(k - 1) - \chi^2_F} \quad \text{(27)}
\]

in which

\[
\chi^2_F = \frac{12N}{k(k+1)} \sum_{j=1}^k R_j^2 - \frac{k(k+1)^2}{4} \quad \text{(28)}
\]

Table IV shows the Friedman statistics \( F_F \) and the corresponding critical value in regard to each metric [# comparing algorithms \( k = 6 \) (except LR), # subjects \( N = 13 \)]. With respect to each metric, the null hypothesis of equivalent performance among the compared methods can be rejected at the 0.05 significance level.

Then, we perform the strict posthoc Bonferroni–Dunn test [50], which is used to account for the relative performance between ML-BVAE (control algorithm) and other compared approaches. The critical difference (CD) value of the rank difference between the two algorithms is

\[
CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \quad \text{(29)}
\]

that ML-BVAE suppresses all the compared algorithms as far as the primary metrics miF1, maF1, and mAP among the 39 configurations (13 subjects \( \times 3 \) metrics). Specifically, for MEMO27 our method has a relative improvement of 12.3% and 18.8% on miF1 and maF1, respectively (five subjects averaged), compared with the second-place method and a relative improvement of 19.0% and 17.1% can be reached for MEMO80, which is a significant improvement in the neural decoding area.
Fig. 5. Comparisons of ML-BVAE (control algorithm) against other comparing algorithms with the Bonferroni–Dunn test in 13 subjects of MEMO27 and MEMO80. Algorithms not connected with ML-BVAE in the CD diagram are considered to have significantly different performance from the control algorithm ($CD = 1.890$ at 0.05 significance level). (a) OneE. (b) RL. (c) miF1. (d) maF1. (e) e-AP. (f) mAP.

in which $q_α = 2.576$ at 0.05 significance level. Therefore, ML-BVAE can be considered as having significantly different performance than a compared algorithm if their average rank difference is larger than CD ($CD = 1.890$ in our experimental setting). Fig. 5 reports the CD diagrams on each metric, where the average rank of each compared method is marked along the axis (the smaller the better). Algorithms not connected with ML-BVAE in the CD diagram are considered to have a significantly different performance from the control algorithm. We can observe that: 1) ML-BVAE achieves the best average rank with respect to all metrics. 2) As far as the primary metrics, ML-BVAE is significantly better than the compared methods other than CA2E in terms of miF1 and maF1 and achieves the significantly best results in mAP compared with other methods, rather than ML-GCN. 3) Although some strong baselines, such as ML-GCN, are not significantly different from ML-BVAE in terms of a few metrics which are of little importance, our method still has obvious superiority when all metrics are taken into consideration and achieves consistently better average ranks. These experimental results convincingly illustrate the significance of the superiority of our ML-BVAE approach.

2) Average Precision Across All Categories: We also report the average precision\(^5\) across all emotion categories with respect to MEMO27 and MEMO80 in Tables V and VI. The results for each specific emotion label can be found in the bar graphs of Appendix B. We can observe that ML-BVAE outperforms other compared methods in most emotion categories in both datasets. To be specific, ML-BVAE ranks first and second in 92.6% (25/27) and 7.4% (2/27), respectively, in MEMO27, and 83.8% (67/80) and 13.8% (11/80), respectively, in MEMO80. To further analyze the decoding performance of ML-BVAE in each emotion category, we show the average precision corresponding to each emotional state in order of sample size for each category in Figs. 6 and 7. Based on these results, we can make the following observations: 1) the experimental results show a strong consistency on multiple subjects in each dataset, which indicates that the decoding performance of ML-BVAE is stable to some extent in the face of individual differences in fMRI data.

\(^5\)This average precision used in this section refers to the intermediate result when calculating the label-based metric mAP.

2) The number of samples is a very important factor that affects the decoding performance. MEMO27 suffers from the long tail distribution of labels leading to lower decoding performance in the categories with a smaller number of samples in general. 3) The intensity of emotional stimuli is another factor that has an effect on emotion decoding. For certain strong emotions, such as sexual desire and romance in MEMO27 and sexiness in MEMO80, they still have high decoding accuracy despite small samples. 4) The label quality also exerts influence on the decoding accuracy. For example, in MEMO80, lethargic and fever have the lowest consistency in emotion ratings among annotators [5], in which fever has very low decoding accuracy, while lethargic has not a significantly bad performance due to large sample size. We can also find in Fig. 8 that confusion in MEMO27 and admiration in MEMO80 are both noisy labels because they have no obvious co-occurrence with other labels, which leads to low accuracy. 5) Although the number of decoding categories of MEMO80 is much greater than that of MEMO27 leading to a harder decoding problem, it still achieves considerable decoding accuracy in terms of mAP. This may be due to the higher quality of fMRI data thanks to extra auditory stimulation in the movies of MEMO80.

3) Masked Self-Attention Visualization: We present the masked self-attention matrix $A$ in (18), which is adopted in the multi-label classification network for modeling the emotion label dependency in Fig. 8. The attention matrix can be seen...
### TABLE V

| Method | 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 | Ω | mAP |
|--------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LR [6] | 16.7 | 18.8 | 45.5 | 58.7 | 3.1 | 33.3 | 35.7 | 6.2 | 17.9 | 8.9 | 15.8 | 8.6 | 42.4 | 12.2 | 23.7 | 20.6 | 38.3 | 32.4 | 36.6 | 26.5 | 17.5 | 7.8 | 18.7 | 14.2 | 17.7 | 50.7 | 35.8 | 24.6 |
| Benchmark | 26.1 | 51.6 | 66.7 | 76.1 | 24.1 | 51.8 | 59.4 | 23.3 | 26.3 | 32.1 | 31.1 | 27.9 | 57.4 | 32.0 | 51.0 | 31.8 | 57.8 | 51.5 | 59.2 | 25.3 | 28.7 | 22.2 | 43.5 | 31.4 | 26.3 | 70.0 | 56.4 | 43.4 |
| CA2E [46] | 20.3 | 34.0 | 40.4 | 49.8 | 17.9 | 21.8 | 23.4 | 26.5 | 21.2 | 42.7 | 23.6 | 37.4 | 24.0 | 44.1 | 39.7 | 51.3 | 41.4 | 22.0 | 0.1 | 31.2 | 26.0 | 21.3 | 57.5 | 45.8 | 34.1 |
| ML-GCN [47] | 25.7 | 51.7 | 66.4 | 76.5 | 19.9 | 51.9 | 58.7 | 21.4 | 25.5 | 29.8 | 34.2 | 25.0 | 36.6 | 29.2 | 50.6 | 30.1 | 57.2 | 30.4 | 61.1 | 52.7 | 27.7 | 19.5 | 36.9 | 32.4 | 25.3 | 65.6 | 56.8 | 42.2 |
| SIM2 [32] | 25.0 | 42.1 | 61.2 | 73.8 | 19.1 | 50.3 | 56.4 | 21.1 | 24.2 | 30.2 | 32.0 | 15.2 | 54.1 | 30.8 | 47.7 | 47.6 | 53.6 | 46.9 | 59.7 | 47.3 | 27.5 | 22.3 | 33.1 | 30.4 | 26.1 | 65.3 | 54.2 | 39.8 |
| GARDIS [48] | 19.3 | 43.4 | 62.9 | 73.8 | 11.7 | 47.4 | 54.9 | 15.8 | 20.4 | 28.1 | 25.4 | 42.2 | 53.5 | 25.3 | 48.2 | 22.8 | 52.5 | 44.7 | 61.0 | 48.3 | 26.1 | 1.4 | 31.8 | 21.6 | 19.0 | 67.2 | 55.1 | 37.9 |

**Remarks:**
- **ML-BVAE:** multi-view multi-label fine-grained emotion decoding from human brain activity.
- **AP:** anchor posterior.
- **MAP:** maximum a posteriori.
- **ML-BV AE:** multi-view multi-label fine-grained emotion decoding from human brain activity.
- **MEMO27:** five subjects averaged.
- **MEMO80:** eight subjects averaged.
- **ADMIRATION:** admiration.
- **ADORATION:** adoration.
- **ESTHETIC APPRECIATION:** aesthetic appreciation.
- **AMUSEMENT:** amusement.
- **ANGER:** anger.
- **ANXIETY:** anxiety.
- **AWKWARDNESS:** awkwardness.
- **BOREDOM:** boredom.
- **CALMNESS:** calmness.
- **CONFUSION:** confusion.
- **CRAVING:** craving.
- **DISGUST:** disgust.
- **EMPATHIC PAIN:** empathic pain.
- **ENTRANCEMENT:** entrancement.
- **EXCITEMENT:** excitement.
- **FEAR:** fear.
- **HORROR:** horror.
- **INTEREST:** interest.
- **JOY:** joy.
- **NOSTALGIA:** nostalgia.
- **RELIEF:** relief.
- **ROMANCE:** romance.
- **SADNESS:** sadness.
- **SATISFACTION:** satisfaction.
- **SEXUAL DESIRE:** sexual desire.
- **SURPRISE:** surprise.
- **THROB:** throb.
- **THREATENING:** threatening.
- **TENSION:** tension.
- **UNREST:** unrest.
- **EXUBERANCE:** exuberance.
- **APPRECIATION OF BEAUTY:** appreciation of beauty.
- **FEVER:** fever.
- **PLEASURE:** pleasure.
- **POSITIVE-EXPECTATION:** positive expectation.
- **THERAPY:** therapy.
- **CONFUSION:** confusion.
- **ANGER:** anger.
- **ANXIETY:** anxiety.
- **UNREST:** unrest.
- **EXUBERANCE:** exuberance.
- **APPRECIATION OF BEAUTY:** appreciation of beauty.
- **FEVER:** fever.
- **PLEASURE:** pleasure.
- **POSITIVE-EXPECTATION:** positive expectation.
- **THERAPY:** therapy.
- **CONFUSION:** confusion.
- **ANGER:** anger.
- **ANXIETY:** anxiety.
- **UNREST:** unrest.
- **EXUBERANCE:** exuberance.
- **APPRECIATION OF BEAUTY:** appreciation of beauty.
- **FEVER:** fever.
- **PLEASURE:** pleasure.
- **POSITIVE-EXPECTATION:** positive expectation.
- **THERAPY:** therapy.
- **CONFUSION:** confusion.
- **ANGER:** anger.
- **ANXIETY:** anxiety.
- **UNREST:** unrest.
- **EXUBERANCE:** exuberance.
- **APPRECIATION OF BEAUTY:** appreciation of beauty.
- **FEVER:** fever.
- **PLEASURE:** pleasure.
- **POSITIVE-EXPECTATION:** positive expectation.
- **THERAPY:** therapy.
- **CONFUSION:** confusion.
- **ANGER:** anger.
- **ANXIETY:** anxiety.
- **UNREST:** unrest.
- **EXUBERANCE:** exuberance.
- **APPRECIATION OF BEAUTY:** appreciation of beauty.
- **FEVER:** fever.
- **PLEASURE:** pleasure.
- **POSITIVE-EXPECTATION:** positive expectation.
- **THERAPY:** therapy.
- **CONFUSION:** confusion.
- **ANGER:** anger.
- **ANXIETY:** anxiety.
- **UNREST:** unrest.
- **EXUBERANCE:** exuberance.
- **APPRECIATION OF BEAUTY:** appreciation of beauty.
- **FEVER:** fever.
- **PLEASURE:** pleasure.
- **POSITIVE-EXPECTATION:** positive expectation.
Fig. 7. Bar graph of the number of samples corresponding to each emotion category and line graphs of each subject’s emotional decoding AP in MEMO80.

Fig. 8. Visualization of the masked self-attention matrix in MEMO27 and MEMO80. The indexes of emotion labels in (a) and (b) are the same as in Figs. V and VI.

Fig. 9. Visualization of a few rows of the masked self-attention matrix. Each element can be regarded as the probability of co-occurrence with the emotion corresponding to the row. (a) Adoration in MEMO27. (b) Anger in MEMO27. (c) Love in MEMO80. (d) Nervous in MEMO80.

and joy; while if we feel angry we may also express disgust, fear, horror, and sadness. Similar results can be observed in Fig. 9(c) and (d) regarding MEMO80.

4) Effectiveness of Bihemisphere Discrepancy: To further prove that the bihemisphere discrepancy can provide additional information for emotion decoding, which leads to better decoding performance, we compare the two views in which the difference view is not included and all three views as the input on all subjects in MEMO27 and MEMO80. Fig. 10 reports the corresponding results on the primary metrics miF1, maF1, and mAP. We can observe that, compared with the experiments without the difference view, the decoding performance on all subjects in both datasets has been improved. It is worth noting that the improvement of miF1 and maF1 in MEMO27 is particularly significant when bihemisphere discrepancy is considered.

Apart from this, Fig. 10 also includes the performance when our model is trained on only the difference view. For a fair comparison, we realize this by setting all three inputs of BVAE to the difference view such that the expressive neural representations can also be learned. It is observed that when only the difference view is input, the mAP of the model is lower than the result of the input of the left and right hemisphere views (red bar) but miF1 and maF1 are a little bit higher except maF1 in MEMO27 and individual subjects in MEMO80. The model can decode emotional states with relative accuracy when only the difference view is input, which indicates that the discrepancy of the left and right hemispheres indeed provides useful information for emotion decoding. Furthermore, a model with all three views as input still has the
Fig. 11. Performance of ML-BV AE changes as the hyperparameters $N_{ROI}$ and $\lambda_1/\lambda_2$ varies in terms of $\text{miF1} + \text{maF1}$. (a) $N_{ROI}$ in MEMO27. (b) $\lambda_1/\lambda_2$ in MEMO27. (c) $\lambda_1/\lambda_2$ in MEMO80.

Table VII

| Dataset | Metric | w/o ROI-P | w/o BVAE | w/o ML-C | w/o E | w/o M | ML-BVAE |
|---------|--------|-----------|----------|----------|-------|-------|---------|
| MEMO27  | One2R  | 0.354     | 0.314    | 0.312    | 0.309 | 0.310 | 0.302   |
|         | RL2    | 0.221     | 0.165    | 0.186    | 0.190 | 0.187 | 0.186   |
|         | miF1   | 0.462     | 0.380(0.125) | 0.461(0.044) | 0.463 | 0.455 | 0.595   |
|         | maF1   | 0.313     | 0.230(0.146) | 0.291(0.107) | 0.302 | 0.281 | 0.398   |
|         | A-AP*  | 0.569     | 0.413    | 0.412    | 0.411 | 0.413 | 0.619   |
|         | M-AP*  | 0.386     | 0.47(0.031) | 0.456(0.018) | 0.433 | 0.431 | 0.444   |

MEMO80  | One2R  | 0.451     | 0.493    | 0.443    | 0.481 | 0.442 | 0.624   |
|         | RL2    | 0.311     | 0.307    | 0.305    | 0.329 | 0.306 | 0.298   |
|         | miF1   | 0.341     | 0.200(0.177) | 0.333(0.048) | 0.374 | 0.361 | 0.378   |
|         | maF1   | 0.355     | 0.146(0.185) | 0.314(0.055) | 0.364 | 0.351 | 0.349   |
|         | A-AP*  | 0.474     | 0.472    | 0.479    | 0.456 | 0.479 | 0.488   |
|         | M-AP*  | 0.469     | 0.441(0.030) | 0.446(0.011) | 0.455 | 0.462 | 0.471   |

5) **Ablation Study**: In order to explore the role of the three key components of our method separately, we design three variant models of ML-BV AE for ablation study: 1) w/o ROI-P: use whole brain voxel signal instead of ROI pooling features as input to our model; 2) w/o BVAE: first, use MLP for single-view features, respectively, and then concatenate the output of each MLP as multi-view neural representations; and 3) w/o ML-C: directly connect the output of the encoder of BVAE to a linear classifier for prediction without multi-label classification network.

Furthermore, we design another two variant models for a more fine-grained ablation study. For verifying the importance of modeling the information intraview and interview for neural representations learning, we consider a variant model w/o E: Remove the three extra ELBOs in (12)–(14) from the loss function. For illustrating the necessity of prior knowledge injection in the label correlation learning, we remove the mask from (18) leading to the variant model w/o M.

Results in the two datasets are shown in Table VII and we have averaged the results of all subjects in each dataset. Results show that all three key components in ML-BV AE are critical to more accurate emotion decoding from brain activity. Furthermore, the extra three ELBOs are helpful to neural representation learning and the prior knowledge mask facilitates label correlation learning, which can both contribute to multi-label emotion decoding.

Besides, it is necessary to analyze whether the high performance is attributed to the discriminative component with huge parameter space, in order to exclude the posterior collapse phenomenon. We mark the model performance degradation caused by the lack of BVAE (the generative component) and ML-C (the discriminative component) in Table VII in terms of the comprehensive metrics $\text{miF1}$, $\text{maF1}$, and $\text{mAP}$. We can observe that, compared with w/o ML-C, w/o BVAE has a greater impact on model performance. In other words, the high performance should be attributed more to the expressive neural representations learning of BVAE.

6) **Parameter Sensitivity**: The two most important hyperparameters in our experiment are the number of features per ROI $N_{ROI}$ and the tradeoff parameter $\lambda_1/\lambda_2$ in (24). We design experiments by varying one parameter while fixing another parameter. We utilize the criterion $\text{miF1} + \text{maF1}$ for parameter selection.

Fig. 11(a) gives an illustrative example of how the performance of ML-BV AE changes when $N_{ROI}$ changes in MEMO27. We find that both global average pooling of the voxels in ROI and utilizing all voxels are not suitable. In general, the performance of ML-BV AE increases first and then decreases with the increase of this parameter. Therefore, we select the parameter, which can reach the optimal performance in terms of this criterion for each subject.

Fig. 11(b) and (c) shows the performance of ML-BV AE changes when the tradeoff parameter $\lambda_1/\lambda_2$ varies. In MEMO27, generally speaking, the model performance increases first and then falls down as the parameter increases. It is noticeable that the performance can be significantly degraded when the parameter becomes too large. Similar conclusions can be drawn in MEMO80.

V. CONCLUSION

We have proposed a hybrid model ML-BV AE for fine-grained multi-view multi-label emotion decoding from visually

6Since the number of voxels for each ROI in the MEMO80 is small, we select eight by default.
Fig. 12. Comparisons of AP and mAP in % of ML-BVAE and compared methods in MEMO27 (five subjects averaged). The best performance is marked with a black outline.

Fig. 13. Comparisons of AP and mAP in % of ML-BVAE and compared methods in MEMO80 (eight subjects averaged). The best performance is marked with a black outline.

evoked brain activity. The proposed method can be divided into three key components. First, we employed ROI pooling for fMRI dimensionality reduction which can alleviate overfitting. Second, BVAE was used for multi-view neural representation learning and we also took the discrepancy of bihemisphere into consideration. At last, a multi-label classification network was implemented for emotion-specific representation learning and modeling the dependency of emotion labels. Our method can extract expressive neural representations for accurate multi-label emotion decoding up to 80 fine-grained emotion categories. We leveraged two fine-grained multi-label emotion decoding benchmark datasets, and comprehensive experiments on them have confirmed the superiority of the proposed method. In the future, decoding emotional states from noisy labels may be a promising research direction as there are some emotion categories whose ratings have low consistency among annotators.

APPENDIX A

PROOF OF THE DISTRIBUTION OF A PRODUCT OF GAUSSIAN EXPERTS

Proposition 1: Give a finite number $N$ of multidimensional Gaussian distributions $p_i(x)$ with mean $\mu_i$ and covariance $\Sigma_i$, in which $i = 1, \ldots, N$, the product $\prod_{i=1}^{N} p_i(x)$ is
Gaussian with mean \( (\sum_{i=1}^{N} \Sigma_{i}^{-1} \mu_{i}) (\sum_{i=1}^{N} \Sigma_{i}^{-1})^{-1} \) and covariance \( (\sum_{i=1}^{N} \Sigma_{i}^{-1})^{-1} \).

**Proof:** The canonical form of the probability density of a Gaussian distribution is \( K \exp[-\eta^T x - (1/2)^T \Lambda x] \), where \( K \) is a normalizing constant, \( \Lambda = \Sigma^{-1} \) and \( \eta = \Sigma^{-1} \mu \). Then, we can get the probability density of the product of \( N \) Gaussian distributions \( \prod_{i=1}^{N} P_{i} \propto \exp(\sum_{i=1}^{N} \eta_{i}^T x - (1/2)^T \Lambda_{i} x) \). Therefore, this product itself has the Gaussian distribution with \( \eta = \sum_{i=1}^{N} \eta_{i} \) and \( \Lambda = \sum_{i=1}^{N} \Lambda_{i} \). Converting back from canonical form, the product Gaussian has a mean of \( \mu = (\sum_{i=1}^{N} \Sigma_{i}^{-1} \mu_{i}) (\sum_{i=1}^{N} \Sigma_{i}^{-1})^{-1} \) and covariance \( (\sum_{i=1}^{N} \Sigma_{i}^{-1})^{-1} \).

**APPENDIX B**

**BAR GRAPHS OF COMPARISONS OF AP AND MAP IN % OF ML-BLVAE AND COMPARED METHODS**

See Figs. 12 and 13.

**ACKNOWLEDGMENT**

The authors would like to thank Tomoyasu Horikawa and Naoko Koide-Majima for providing the necessary data for their research.

**REFERENCES**

[1] J. Li, Z. Zhang, and H. He, “Hierarchical convolutional neural networks for EEG-based emotion recognition,” Cogn. Comput., vol. 10, no. 2, pp. 368–380, 2018.

[2] C. Du et al., “Semi-supervised deep generative modelling of incomplete multi-modality emotional data,” in Proc. ACM Multimedia Conf. Multimedia Conf., Seoul, South Korea, Oct. 2018, pp. 108–116.

[3] Y. Li et al., “A novel bi-hemispheric discrepancy model for eeg emotion recognition,” IEEE Trans. Cognit. Develop. Syst., vol. 13, no. 2, pp. 354–367, 2020.

[4] A. S. Cowen and D. Keltner, “Self-report captures 27 distinct categories of emotion bridged by continuous gradients,” Proc. Nat. Acad. Sci. USA, vol. 114, no. 38, pp. E7900–E7909, Sep. 2017.

[5] N. Koide-Majima, T. Nakai, and S. Nishimoto, “Distinct dimensions of emotion in the human brain and their representation on the cortical surface,” NeuroImage, vol. 222, Nov. 2020, Art. no. 117258.

[6] T. Horikawa, A. S. Cowen, D. Keltner, and Y. Kamitani, “The neural representation of visually evoked emotion is high-dimensional, categorical, and distributed across transmodal brain regions,” iScience, vol. 23, no. 5, May 2020, Art. no. 101060.

[7] P. A. Krabel and K. S. LaBar, “Decoding the nature of emotion in the brain,” Trends Cognit. Sci., vol. 20, no. 6, pp. 444–455, 2016.

[8] V. Putkinen et al., “Decoding music-evoked emotions in the auditory and motor cortex,” Cerebral Cortex, vol. 31, no. 5, pp. 2549–2560, 2021.

[9] H. Saarimäki et al., “Classification of emotion categories based on functional connectivity patterns of the human brain,” NeuroImage, vol. 247, Feb. 2021, Art. no. 118800.

[10] S. J. Dimond, L. Farrington, and P. Johnson, “Differing emotional response from right and left hemispheres,” Nature, vol. 261, no. 5562, pp. 690–692, Jun. 1976.

[11] R. J. Davidson, “Anterior cerebral asymmetry and the nature of emotion,” Brain Cognit., vol. 20, no. 1, pp. 125–151, Sep. 1992.

[12] D. Huang, S. Chen, C. Liu, L. Zheng, Z. Tian, and D. Jiang, “Differences first in asymmetric brain: A bi-hemispheric discrepancy convolutional neural network for EEG emotion recognition,” Neurocomputing, vol. 448, pp. 140–151, Aug. 2021.

[13] Y. Cao and D. J. Fleet, “Generalized product of experts for automatic and principled fusion of Gaussian process predictions,” 2014.

[14] L. B. Baucom, D. H. Weddell, J. Wang, D. N. Wang, and S. V. Shinkerava, “Decoding the neural representation of affective states,” NeuroImage, vol. 59, no. 1, pp. 718–727, 2012.
[39] I. Higgins et al., “Beta-VAE: Learning basic visual concepts with a constrained variational framework,” in Proc. 5th Int. Conf. Learn. Represent. (ICLR), Toulon, France, Apr. 2017.

[40] S. R. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio, “Generating sentences from a continuous space,” in Proc. 20th SIGNLL Conf. Comput. Natural Lang. Learn., Berlin, Germany, 2016, pp. 10–21. [Online]. Available: https://www.aclweb.org/anthology/K16-1002

[41] M. Wu and N. D. Goodman, “Multimodal generative models for scalable weakly-supervised learning,” in Proc. Adv. Neural Inf. Process. Syst., Montréal, QC, Canada, 2018, pp. 5580–5590.

[42] Z. Chen, X.-S. Wei, P. Wang, and Y. Guo, “Learning graph convolutional networks for multi-label recognition and applications,” IEEE Trans. Pattern Anal. Mach. Intell., early access, Mar. 3, 2021, doi: 10.1109/TPAMI.2021.3063496.

[43] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.

[44] E. Ben-Baruch et al., “Asymmetric loss for multi-label classification,” 2020, arXiv:2009.14119.

[45] S. Liu, L. Zhang, X. Yang, H. Su, and J. Zhu, “Query2Label: A simple transformer way to multi-label classification,” 2021, arXiv:2107.10834.

[46] C.-K. Yeh, W.-C. Wu, W.-J. Ko, and Y.-C.-F. Wang, “Learning deep latent space for multi-label classification,” in Proc. AAAI Conf. Artif. Intell., vol. 31, no. 1, Feb. 2017, pp. 2838–2844.

[47] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, “Multi-label image recognition with graph convolutional networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 5177–5186.

[48] Z. Chen, X. Wu, Q. Chen, Y. Hu, and M. Zhang, “Multi-view partial multi-label learning with graph-based disambiguation,” in Proc. 44th AAAI Conf. Artif. Intell. (AAAI), 2020, pp. 3553–3560.

[49] M.-L. Zhang and Z.-H. Zhou, “A review on multi-label learning algorithms,” IEEE Trans. Knowl. Data Eng., vol. 26, no. 8, pp. 1819–1837, Jun. 2013.

[50] J. Demšar, “Statistical comparisons of classifiers over multiple data sets,” J. Mach. Learn. Res., vol. 7, pp. 1–30, Jan. 2006.