Unsupervised layer-wise feature extraction algorithm for surface electromyography based on information theory

Mingqiang Li¹*, Ziwen Liu²†, Siqi Tang¹, Jianjun Ge¹ and Feng Zhang¹

¹Information Science Academy, China Electronics Technology Group Corporation, Beijing, China
²School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China

Feature extraction is a key task in the processing of surface electromyography (SEMG) signals. Currently, most of the approaches tend to extract features with deep learning methods, and show great performance. And with the development of deep learning, in which supervised learning is limited by the excessive expense incurred due to the reliance on labels. Therefore, unsupervised methods are gaining more and more attention. In this study, to better understand the different attribute information in the signal data, we propose an information-based method to learn disentangled feature representation of SEMG signals in an unsupervised manner, named Layer-wise Feature Extraction Algorithm (LFEA). Furthermore, due to the difference in the level of attribute abstraction, we specifically designed the layer-wise network structure. In TC score and MIG metric, our method shows the best performance in disentanglement, which is 6.2 lower and 0.11 higher than the second place, respectively. And LFEA also get at least 5.8% accuracy lead than other models in classifying motions. All experiments demonstrate the effectiveness of LFEA.

KEYWORDS
information theory, feature extraction, unsupervised learning, information bottleneck, disentangled representation, surface electromyography

Introduction

Feature engineering is an important component of pattern recognition and signal processing. Learning good representations from observed data can help reveal the underlying structures. In recent decades, feature extraction methods (He et al., 2016; Howard et al., 2017; Hassani and Khasahmadi, 2020; Zhontar et al., 2021) have drawn considerable attention. Due to the high cost of obtaining labels, supervised learning methods suffer from data volume limitations. Unsupervised learning methods
therefore becomes critical for feature extraction. Most of these are based on probabilistic models, such as maximum likelihood estimation (Myung, 2003), maximum a posteriori probability estimation (Richard and Lippmann, 1991), and mutual information (MI) (Thomas and Joy, 2006). Methods such as principal component analysis (PCA) (Abdi and Williams, 2010), linear discriminant analysis (Izenman, 2013), isometric feature mapping (Tenenbaum et al., 2000), and Laplacian eigenmaps (Belkin and Niyogi, 2003) are widely used owing to their good performance, high efficiency, flexibility, and simplicity. Other algorithms are based on reconstruction errors or generative criteria, such as autoencoders (Bengio et al., 2013) and generative adversarial networks (GANs) (Goodfellow et al., 2014). Occasionally, the reconstruction error criterion also has a probabilistic interpretation.

In recent years, deep learning has become a dominant method of representation learning, particularly in the supervised case. A neural network simulates the mechanism of hierarchical information processing in the brain and is optimized using the back propagation (BP) algorithm (LeCun et al., 1988). Because several feature engineering tasks are unsupervised, that is, no label information is available in the real situation and collecting considerable labeled data is expensive, methods to discover the feature representation in an unsupervised case have been significantly developed in recent years. MI maximization (Bell and Sejnowski, 1995) and minimization criteria (Matsuda and Yamaguchi, 2003) are powerful tools for capturing salient features of data and disentangling these features. In particular, variational autoencoder (VAE) (Kingma and Welling, 2013) and generative adversarial networks (GANs) (Goodfellow et al., 2014). Occasionally, the reconstruction error criterion also has a probabilistic interpretation.

Disentangled representation

The disentanglement problem has played a significant role, particularly because of its better interpretability and controllability. The VAE variants construct representations in which each dimension is independent and corresponds to a dedicated attribute. β-VAE (Higgins et al., 2016) adds a hyperparameter to control the trade-off between compression and expression. An analysis of β-VAE by Burgess et al. (2018) is provided, and the capacity term is proposed to obtain a better balance of the reconstruction error. Penalizing the total correlation term to reinforce the independence among representation dimensions was proposed in Factor VAE (Kim and Mnih, 2018) and β-TCVAE (Chen et al., 2018). FHVAE (Hsu et al., 2017) and DSVAE (Yingzhen and Mandt, 2018) constructed a new model architecture and factorized the latent variables into static and dynamic parts. Cheng et al. (2020b) described a GAN model using MI. Similar to our study, Gonzalez-Garcia et al. (2018) proposed a model to disentangle the attributes of paired data into shared and exclusive representations.

Information theory

Shannon’s MI theory (Shannon, 2001) is a powerful tool for characterizing good representation. However, one major problem encountered in the practical application of information theory is computational difficulties in high-dimensional spaces. Numerous feasible computation methods have been proposed, such as Monte Carlo sampling, population coding, and the mutual information neural estimator (Belghazi et al., 2018). In addition, the information bottleneck (IB) principle
Random variables or factors. Given two random variables $X$ and $Z$, the MI is defined as follows:

$$I(X; Z) = \mathbb{E}_{p(x,z)} \left[ \log \frac{p(x,z)}{p(x)p(z)} \right]$$  \hfill (1)

Regarding the data processing flow as a Markov chain $X \rightarrow Z \rightarrow Y$, the information bottleneck (IB) principle desires that the useful information in the input $X$ can pass through the ‘bottleneck’ while the noise and irrelevant information are filtered out. The IB principle is expressed as follow:

$$\min \ R_{IB} = I(X; Z) - \beta I(Z; Y)$$  \hfill (2)

where, $\beta$ is the tradeoff parameter between the complexity of the representation and the amount of relevant essential information.

Framework

The diagram of our proposed Layer-wise Feature Extraction Algorithm (LFEA) is illustrated in Figure 1. Our algorithm aims to learn a representation that satisfies three main properties: “Compression,” “Expression” and “Disentanglement.” To this end, three key information process modules are introduced, including the information compression module (ICM), information expression module (IEM), and information separation module (ISM) in each layer.

In the ICM, input $s^{i-1}$ of layer $i$ is compressed into $h^i$ ($s^0 = X$). In the IEM, $z^i$ as part of $h^i$ is constrained to represent the original input $X$. In the ISM section, $s^i$ and $z^i$ are irrelevant. The parameters of the ICM and IEM in layer $i$ are denoted as $\phi^i$ and $\theta^i$. The data information flow can be expressed as follows:

$$h^i \sim q_{\phi^i}(h^i|s^{i-1})$$  \hfill (3)

$$h^i = (z^i, s^i)$$  \hfill (4)

$$\tilde{X} \sim p_{\theta^i}(\tilde{X}|z^i)$$  \hfill (5)

where, $s^0 = X$, and $q_{\phi^i}$ and $p_{\theta^i}$ are the condition distributions with $\phi^i$ and $\theta^i$ for $h^i$ and $\tilde{X}$. In following sections, we describe these three modules in detail.

Information compression module

According to (3), $h^i$ is the hidden representation of $s^{i-1}$. To ensure information ‘compression,’ the optimal representation of $s^{i-1}$ should forget redundant information altogether, that is, $h^i$ represents $s^{i-1}$ with the lowest bits. Formally, the objective in the $i$-th layer to be minimized is as follows:

$$\min_{L_{ICM}} \ L_{\phi^i}(s^{i-1}; h^i)$$  \hfill (6)
FIGURE 1
The diagram of Layer-wise Feature Extraction Algorithm (LFEA). LFEA contains three core modules: Information Compression Module (ICM), Information Expression Module (IEM) and Information Separation Module (ISM), to ensure compression, expression and disentanglement of representation, respectively.

Due to intractability of mutual information, optimizing $L_{ICM}$ with gradient methods directly is not feasible. We therefore derived the upper bound of $L_{ICM}$ with the variational inference method and get decomposition as follows:

$$I_{\phi_i}(S_i^{l-1}; H^i) = E_{q_{\phi_i}(h^i)} \left[ \log \frac{q_{\phi_i}(h^i \mid s_i^{l-1}) p(h^i)}{q_{\phi_i}(h^i \mid s_i^{l-1}) p(h^i)} \right]$$

$$= L_{upper}^{KL} - D_{\text{KL}}(q_{\phi_i}(h^i \mid s_i^{l-1}) || p(h^i)),$$  

where, $p(h^i)$ is the prior, and $L_{upper}^{KL}$ is the upper bound of $L_{ICM}$ defined as follows:

$$L_{upper}^{KL} = E_{q_{\phi_i}(h^{l-1})} \left[ D_{\text{KL}}(q_{\phi_i}(h^i \mid s_i^{l-1}) || p(h^i)) \right].$$

Information expression module

With the ICM guaranteeing the information compression, LFEA also need to ensure the expressiveness of the representation to the data. We therefore propose the information expression module (IEM). To ensure sufficient information to reconstruct the original data $X$, we maximize the MI between and $Z$ in $i$-th layer, that is,

$$\max L_{IEM} \triangleq I_{\phi_i, \theta_i}(Z^i; X)$$

For $L_{IEM}$, we can obtain a lower bound using the variational approximation method as follows:

$$L_{IEM} \geq L_{lower}^{IEM} - D_{\text{KL}}(p(x) \parallel p_{\theta_i}(x)),$$  

where, $p_{\theta_i}(x)$

$$L_{lower}^{IEM} = E_{p(x)} \left[ E_{q_{\phi_i}(z^i \mid x)} \log p_{\theta_i}(z^i \mid x) \right].$$

Information separation module

To achieve disentanglement of representations (Independent of each block $z^1, z^2, \ldots, z^n$ in $Z$), we further introduce the information separation module (ISM) in each layer. In $i$-th layer, the principle of ISM is to ensure that there is no intersection information between $z^i$ and $s^i$, that is,

$$\max L_{ISM} \triangleq I_{\phi_i}(z^i; s^i)$$

$$= D_{\text{KL}}(q_{\phi_i}(h^i) || q_{\phi_i}(z^i) q_{\phi_i}(s^i)).$$

In practice, the products of $q_{\phi_i}(z^i)$ and $q_{\phi_i}(z^i)$ are not analytical in nature. We introduce discriminator $D(.)$ (see Figure 2) to distinguish samples from the joint distribution and the product of the marginal distribution, that is,

$$L_{ISM} \approx L_{IEM} = E_{q_{\phi_i}(h^i)} \log \frac{D(.)}{1 - D(.)}.$$  

(13)
FIGURE 2
Discriminator $D(.)$. To compute and optimize $L_{ISM}$, we need an additional discriminator as shown in Eq. (13).

FIGURE 3
Movements in NinaPro DB2. (A) Isometric, isotomic hand configurations. (B) Basic movements of the wrist. (C) Grasps and functional movements. (D) Single and multiple fingers force measurement patterns. (E) Rest position. Available from: http://ninapro.hevs.ch/node/123.

TABLE 1 Subject attribute information of NinaPro DB2 dataset.

| Subject | Hand    | Laterality       | Gender   | Age | Height (cm) | Weight (kg) |
|---------|---------|------------------|----------|-----|-------------|-------------|
| 1       | Intact  | Right Handed     | Male     | 29  | 187         | 75          |
| 2       | Intact  | Right Handed     | Male     | 29  | 183         | 75          |
| 3       | Intact  | Right Handed     | Male     | 31  | 174         | 69          |
| 4       | Intact  | Left Handed      | Female   | 30  | 154         | 50          |
| 5       | Intact  | Right Handed     | Male     | 25  | 175         | 70          |
TABLE 2  Detail parameters for LFEA.

| Parameter       | Value |
|-----------------|-------|
| Number of layers| 4     |
| Size of $z^i$   | 5     |
| $\lambda$       | 0.1   |
| $\beta$         | 0.2   |

TABLE 3  Results of TC score.

| Method        | TC score | MIG  |
|---------------|----------|------|
| LFEA (Ours)   | 12.3     | 0.72 |
| VAE           | 23.6     | 0.54 |
| $\beta$-VAE  | 25.8     | 0.61 |
| PCA           | 18.5     | 0.49 |

We compare our method the classic methods including VAE, $\beta$-VAE and PCA. Our HFEA method is much better than others. The bold indicates the best results.

Algorithm optimization

As presented above, our model contains three modules: ICM, IEM, and ISM. However, during optimization, the back-propagation algorithm is computationally intensive and potentially problematic when training deep networks, so we propose a layer-wise training step. After training one layer of the network, we fix the parameters of the trained layers and only train the next layer in the next step. Finally, we can obtain the final model after training all the layers. Such optimization design allows for training parameters at the bottom layers without back-propagation from the top layers, avoiding the problems that often occur with deep network optimization, like vanishing and exploding gradient.

Numerical results

Dataset

In our experiments, we used the NinaPro* DB2 dataset and DB5 dataset. Atzori et al. (2014), Gijsberts et al. (2014) as the benchmark to perform numerical comparisons. NinaPro is a standard dataset for the gesture recognition of sparse multichannel SEMG signals. The SEMG signals in DB2 were obtained from 40 subjects and included 49 types of hand movements (see Figure 3). Detailed attribute information of the five subjects in NinaPro DB2 is shown in Table 1. The original SEMG signal was processed through sliding windows, and the size of the sample data used in the experiment was (200,12). Figure 4 shows 20 processed data points.

DB1 consists of 11 subjects and the data set of each subject contains three types of gestures, which are Exercise A, Exercise B, and Exercise C. Exercise A includes 12 basic movements of fingers (see Figure 5). Exercise B includes 17 movements. Exercise C includes 23 grasping and functional movements.

We preprocessed the dataset with the digital filter to cutoff frequency and sliding window to split signal, which follows He et al. (2018).

Model setting

In the following experiments, we used four layers model. The loss function is as follows:

$$
\min L \triangleq L_{ICM}^{upper} - \lambda L_{IEM}^{lower} + \beta L_{ISM}^{lower}
$$
Results

First, we used total correlation (TC) as the quantitative metric for the quality of the disentanglement of the representation. TC is defined as follows:

\[
TC(z^1, z^2, z^3, z^4) = E_{p(z^1, z^2, z^3, z^4)} \left[ \log \frac{p(z^1, z^2, z^3, z^4)}{p(z^1) p(z^2) p(z^3) p(z^4)} \right].
\]

The TC was estimated using a three-like algorithm (Cheng et al., 2020a). A low TC score indicated that the representation had less variance. MIG metric (Chen et al., 2018) is another disentanglement metric; the higher the value, the more disentangled representation is. We compared the quality of disentanglement among PCA, β-VAE, VAE, and HFEA. Table 3 shows the comparison results on TC score and MIG.
### TABLE 4 Classification results on NinaPro DB2 dataset.

| Methods          | Windowing | Train/Test | Accuracy   |
|------------------|-----------|------------|------------|
| LFEA + SVM (Ours)| 200 ms    | 2/1        | 75.2 ± 2.3%|
| CNN              | 200 ms    | 2/1        | 65.7 ± 5.9%|
| LSTM + MLP       | 200 ms    | 1/1        | 75.4 ± 8.2%|
| Random forest    | 200 ms    | 2/1        | 75.0 ± 5.1%|
| KNN              | 200 ms    | 2/1        | 61.1 ± 3.4%|
| SVM              | 200 ms    | 2/1        | 67.2 ± 5.2%|

The bold indicates better result.

In TC score and MIG metric, HFEA has the best performance, which is 6.2 lower and 0.11 higher than the second place, respectively.

Furthermore, in Figure 6, we visualize the distribution of $z_1$, $z_2$, $z_3$, and $z_4$, respectively in a two-dimensional space based on t-distributed stochastic neighbor embedding. We can find that the variance of representation decreases with deeper layers, which indicates that the deeper networks learn more robust representations.

Classification results on NinaPro DB2 dataset is described in Table 4. Our method is based on LFEA and SVM and the feature $Z$ used in SVM is computed by LFEA.

$$Z = (z_1, z_2, z_3, z_4)$$

The bold indicates better result.

The methods used for comparison include LSTM + CNN (He et al., 2018), k-nearest neighbor (KNN), support vector machine (SVM), random forest, and convolutional neural network (CNN) (Atzori et al., 2016). In all experiments, our method was second best in all methods and only 0.2% lower than the best. What is more, our method showed more stable results (2.3% fluctuations) than others.

Discrimination results for Exercise A, Exercise B, and Exercise C in DB1 and DB2 is shown in Figures 7, 8, respectively. For each exercise, we compare feature combinations from layer 1–4. Detail feature combinations is described in Table 5. Tables 6–8 list the classification accuracy with different feature combinations for DB1, respectively.

Discrimination value in Tables 6–8 measures the representation capability of feature in each layer. The higher the value, the better the feature representation ability. In Exercise A, C4 obtains the highest discrimination value, which means feature $z_3$ plays the most important role in Exercise A. Similarly, feature $z_2$ makes little difference in Exercise A.

### Conclusion

In this manuscript, we propose an Unsupervised Layer-wise Feature Extraction Algorithm (LFEA) to perform the sEMG signal processing and downstream classification tasks. The model contains three core modules: Information Compression Module (ICM), Information Expression
Figure 8: Feature discrimination results for DB2.

##### Table 5 Feature combinations.

| Feature combinations | Accuracy |
|----------------------|----------|
| C1 \((z_1, z_2, z_3, z_4)\) | 0.79     |
| C2 \((z_2, z_3, z_4)\) | 0.72     |
| C3 \((z_1, z_3, z_4)\) | 0.74     |
| C4 \((z_1, z_2, z_4)\) | 0.53     |
| C5 \((z_1, z_2, z_3)\) | 0.61     |

The bold values mean the lowest and highest discrimination values.

##### Table 6 Classification results with different feature combinations for Exercise A.

| Feature Combinations | Accuracy | Discrimination (C1-Accuracy) |
|----------------------|----------|-------------------------------|
| C1                   | 0.79     | 0                             |
| C2                   | 0.72     | 0.07                          |
| C3                   | 0.74     | 0.05                          |
| C4                   | 0.53     | 0.26                          |
| C5                   | 0.61     | 0.18                          |

The bold values mean the lowest and highest discrimination values.

Module (IEM) and Information Separation Module (ISM), that ensure that the learning representation is compact, informative and disentangled. We further use a layer-wise optimization procedure to reduce the computation cost and avoid some optimization problem, like vanishing and exploding gradient. Experimentally, we also verify that the untangling effect and downstream classification tasks give better results.

In the future, we hope to combine the advantages of supervised and unsupervised to build a semi-supervised learning framework that can be adapted to more scenarios.

Data availability statement

The original contributions presented in this study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

ML and ZL contributed to the conception and design of the study. FZ organized the database. JG performed the statistical analysis. ML and ST wrote the first draft of the
Funding

This study was supported by the National Key R&D Program of China (2021YFA1000401) and the National Natural Science Foundation of China (U19B2040).

Conflict of interest

ML, ST, JG, and FZ were employed by Information Science Academy, China Electronics Technology Group Corporation.

References

Abdi, H., and Williams, L. J. (2010). Principal component analysis. Wiley Interdiscip. Rev. 2, 433–459. doi: 10.1002/wiit.101

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. (2016). Deep variational information bottleneck. arXiv [preprint]. arXiv:1612.00410.

Atzori, M., Cognolato, M., and Müller, H. (2016). Deep learning with convolutional neural networks applied to electromyography data: A resource for the classification of movements for prosthetic hand prostheses. Front. Neurorobot. 10.9. doi: 10.3389/fnins.2016.00009

Atzori, M., Gijsberts, A., Castellini, C., Caputo, B., Hager, A. G. M., Elias, S., et al. (2014). Electromyography data for non-invasive naturally-controlled robotic hand prostheses. Sci Data 1:140053. doi: 10.1038/sdata.2014.53

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., et al. (2018). “Mutual information neural estimation,” in International Conference on Machine Learning, 531–540. (Stockholm: Stockholm Sweden).

Belkin, M., and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. Neuro Comput. 15, 1373–1396. doi: 10.1162/089976603321780317

Bell, A. J., and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. Neural Comput. 7, 1129–1159. doi: 10.1162/neco.1995.7.6.1129

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1798–1828. doi: 10.1109/TPAMI.2013.50

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., and Lerchner, A. (2018). Understanding disentangling β in vae. arXiv [Preprint]. arXiv:1804.03599.

Chen, R. T., Li, X., Grosse, R. B., and Duvvada, D. K. (2018). Isolating sources of disentanglement in variational autoencoders. Adv. Neural Inf. Proc. Syst. [Preprint]. arXiv:1802.04842.

Cheng, P., Min, M. R., Shen, D., Malon, C., Zhang, Y., Li, Y., et al. (2020b). Improving disentangled text representation learning with information-theoretic guidance. arXiv [preprint]. arXiv:2006.00693. doi: 10.18653/v1/2020.acl-main.673

Cheng, P., Hao, W., and Carin, L. (2020a). Estimating Total Correlation with Mutual Information Bounds. arXiv [Preprint]. arXiv:2011.04794.

Gijsberts, A., Atzori, M., Castellini, C., Müller, H., and Caputo, B. (2014). Measuring movement classification performance with the movement error rate. IEEE Trans. Neural Syst. Rehabil. Eng. 22, 735–744. doi: 10.1109/TNSRE.2014.2303594

Gonzalez-Garcia, A., Van De Weijer, J., and Bengio, Y. (2018). Image-to-image translation for cross-domain disentanglement. Adv. Neural Inf. Proc. Syst. 31, 1287–1298

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In Advances in neural information processing systems (Berlin: Springer), 2672–2680.

Hassani, K., and Khashahami, A. H. (2020). Contrastive multi-view representation learning on graphs. arXiv [Preprint]. arXiv:2006.05582.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (Washington, DC: IEEE Computer Society), 770–777. doi: 10.1109/CVPR.2016.90

He, Y., Fukuda, O., Bu, N., Okumura, H., and Yamaguchi, N. (2018). “Surface emg pattern recognition using long short-term memory combined with multilayer perception,” in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). (Shanghai: IEEE), 5636–5639. doi: 10.1109/EMBC.2018.8513595

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., et al. (2016). “Beta-VAE: Learning basic visual concepts with a constrained variational framework,” in Proceedings of the international conference on learning representations.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. Neural Comput. 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv [Preprint]. arXiv:1704.04861.

Hsu, W. N., Zhang, Y., and Glass, J. (2017). Unsupervised learning of disentangled and interpretable representations from sequential data. Adv. Neural Inf. Proc. Syst. [Preprint]. arXiv:1709.07902.

Izenman, A. J. (2013). “Linear discriminant analysis,” in Modern multivariate statistical techniques. (New York, NY: Springer), 237–280. doi: 10.1002/9780470698896_8

Jeon, I., Lee, W., Pyeon, M., and Kim, G. (2021). “Ib-gan: Disentangled representation learning with information bottleneck generative adversarial networks,” in Proceedings of the AAAI Conference on Computer Vision and Pattern Recognition, 7926–7934.

Kim, H., and Mnih, A. (2018). “Disentangling by factorising,” in International Conference on Machine Learning, 2649–2658.

Kingsma, D. P., and Welling, M. (2013). Auto-encoding variational bayes. arXiv [Preprint]. arXiv:1312.6114v1.

LeCun, Y., Bordes, A., Bengio, Y., and Bengio, Y., Courville, A., and Vincent, P. (2013). Representation From Information Theory Perspective. arXiv [Preprint]. arXiv:2101.08408.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.
Matsuda, Y., and Yamaguchi, K. (2003). "The InfoMin criterion: An information theoretic unifying objective function for topographic mappings," in *Artificial Neural Networks and Neural Information Processing—ICANN/ICONIP 2003*, (Berlin: Springer), 401–408. doi: 10.1007/3-540-44989-2_48

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *J. Math. Psychol.* 47, 90–100. doi: 10.1016/S0022-2496(02)00028-7

Richard, M. D., and Lippmann, R. P. (1991). Neural network classifiers estimate Bayesian a posteriori probabilities. *Neural Comput.* 3, 461–483. doi: 10.1162/neco.1991.3.4.461

Shannon, C. E. (2001). A mathematical theory of communication. *GetMobile* 5, 3–55. doi: 10.1145/584091.584093

Shwartz-Ziv, R., and Tishby, N. (2017). Opening the black box of deep neural networks via information. *arXiv*. [Preprint]. arXiv:1703.00810.

Tenenbaum, J. B., Silva, V. D., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 2319–2323. doi: 10.1126/science.290.5500.2319

Thomas, M. T. C. A. J., and Joy, A. T. (2006). *Elements of information theory*. Hoboken, NJ: Wiley-Interscience.

Tishby, N., and Zaslavsky, N. (2015). "Deep learning and the information bottleneck principle," in *2015 IEEE Information Theory Workshop (ITW)*, (Jeju Island: IEEE), 1–5. doi: 10.1109/ITW.2015.7133169

Tishby, N., Pereira, F. C., and Bialek, W. (2000). The information bottleneck method. *arXiv*. [Preprint]. physics/0004057.

Xing, K., Ding, Z., Jiang, S., Ma, X., Yang, K., Yang, C., et al. (2018). "Hand gesture recognition based on deep learning method," in *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*, (Jeju Island: IEEE), 542–546. doi: 10.1109/DSC.2018.00087

Yingzhen, L., and Mandt, S. (2018). "Disentangled sequential autoencoder," in *International Conference on Machine Learning*, 5670–5679.

Zbontar, J., Jing, L., Misra, I., Lecun, Y., and Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. *Int. Conference Mach. Learn.* 139, 12310–12320.

Zhai, X., Jiels, B., Chan, R. H., and Tin, C. (2017). Self-recalibrating surface EMG pattern recognition for neuroprosthesis control based on convolutional neural network. *Front. Neurosci.* 11:379. doi: 10.3389/fnins.2017.00379