MULTI-VIEW INFORMATION BOTTLENECK WITHOUT VARIATIONAL APPROXIMATION

Qi Zhang[1], *Shujian Yu*[1,2*†], *Jingmin Xin*[1], *Badong Chen*[1*‡]

[1]Xi'an Jiaotong University, Xi'an 710049, Shaanxi, China
[2]UiT - The Arctic University of Norway, 9037 Tromsø, Norway

ABSTRACT

By "intelligently" fuse the complementary information across different views, multi-view learning is able to improve the performance of classification task. In this work, we extend the information bottleneck principle to supervised multi-view learning scenario and use the recently proposed matrix-based Rényi's α-order entropy functional to optimize the resulting objective directly, without the necessity of variational approximation or adversarial training. Empirical results in both synthetic and real-world datasets suggest that our method enjoys improved robustness to noise and redundant information in each view, especially given limited training samples. Code is available at https://github.com/archy666/MEIB.

***Index Terms*—** Information bottleneck, multi-view learning, matrix-based Rényi's α-order entropy functional

1. INTRODUCTION

Multi-view learning becomes popular in many real-world applications, such as multi-omics data including genomics, epigenomics, proteomics, metabolomics [1]. Traditional methods of multi-view learning include canonical correlation analysis (CCA) [2] and its nonlinear extensions like kernel canonical correlation analysis (KCCA) [3]. However, both CCA and KCCA are restricted to only two views.

Due to the remarkable success achieved by deep neural networks (DNNs), there is a recent trend to leverage the power of DNNs to improve the performances of multi-view learning [4]. Andrew *et al.* [5] proposed deep canonical correlation analysis (DCCA) to perform complex nonlinear transformation on multi-view data. Wang *et al.* [6] proposed deep canonically correlated autoencoders (DCCAE) to further improve the performance of DCCA by adding an autoencoder reconstruction error regularization.

Recently, the notable Information Bottleneck (IB) principle [7] has been extended to multi-view learning problem to compress redundant or task-irrelevant information in the input views and only preserve the most task-relevant features [8, 9]. However, parameterizing the IB principle with DNNs is not a trivial task. A notorious reason is that the mutual information estimation in high-dimensional space is widely acknowledged as intractable or infeasible. To this end, existing deep multi-view IB approaches (e.g., [1, 10, 11, 12, 13]) usually use the variational approximation or adversarial training to maximize a lower bound of the original IB objective. However, the tightness of those derived lower bounds is hard to guarantee in practice, especially if there are only limited training data.

In this work, instead of evaluating a lower bound of mutual information terms $I(X; Z)$ or $I(Y; Z)$ in deep multi-view information bottleneck, we demonstrate that it is feasible to directly optimize the IB objective without any variational approximation or adversarial training. We term our methodology **M**ulti-view matrix-based **E**ntropy **I**nformation **B**ottleneck (MEIB) and make the following contributions:

- We introduce the matrix-based Rényi's α-order entropy functional [14] to estimate $I(X; Z)$ directly, which makes our MEIB enjoys a simple and tractable objective but also provides a deterministic compression of input variables from different views.

- Empirical results suggest that our MEIB outperforms other competing methods in terms of robustness to noise and redundant information, especially given limited training samples in each view.

- Our studies also raise a few new insights to the design and implementation of deep multi-view IB methods in the future, as will be discussed in the last section.

2. PRELIMINARIES

2.1. Information Bottleneck Principle

Let us denote the input random variable as X and desired output variable as Y (e.g., class labels). The information bottleneck (IB) approach [7] considers extracting a compressed representation Z from X that is relevant for predicting Y. Formally, this objective is formulated as finding Z such that

---
*To whom correspondence should be addressed (yusj9011@gmail.com; chenbd@mail.xjtu.edu.cn).

the mutual information $I(Y;Z)$ is maximized, while keeping mutual information $I(X;Z)$ below a threshold α:

$$Z^* = \arg\max_{Z \in \Delta} I(Y;Z) \quad \text{s.t.} \quad I(X;Z) \leq \alpha, \qquad (1)$$

where Δ is the set of random variables Z that obeys a Markov chain $Y - X - Z$, which means that Z is conditional independent of Y given X, and any information that Z has about Y is from X.

In practice, one is hard to solve the above constrained optimization problem of Eq. (1), and Z is found by maximizing the following IB Lagrangian:

$$\mathcal{L}_{IB} = I(Y;Z) - \beta I(X;Z), \qquad (2)$$

where $\beta > 0$ is a Lagrange multiplier that controls the trade-off between the **sufficiency** (the performance on the task, as quantified by $I(Y;Z)$) and the **minimality** (the complexity of the representation, as measured by $I(X;Z)$). In this sense, the IB principle also provides a natural approximation of *minimal sufficient statistic* [15].

2.2. Information Bottleneck for Multi-view Learning

Multi-view learning comes down to the problem of machine learning from data represented by multiple distinct feature sets [16].

The main challenge of applying the IB principle is that the mutual information terms are computationally intractable. In order to solve this problem, some researches have access to variational approximation to estimate mutual information such as [1, 10, 11, 12, 13]. The main idea of variational approximation is to develop a variational bound on the sufficiency and redundancy trade-off. Specifically, the intractable distributions will be replaced by some known distributions like Gaussian distribution to calculate a lower bound of the original target. Then the optimal solution of the original objective is obtained by maximization of the lower bound. IB principle has also been applied for unsupervised multi-view learning. For example, [17] compress redundant information with a collaborative multi-view IB network. Meanwhile, [12] assumes that each view contains the same task relevant information, and therefore suggests maximizing the mutual information between latent representations extracted from each view to compress redundant information. These studies also maximize some sample-based differentiable mutual information lower bound instead of directly optimizing objective. As mentioned above, these methods have the drawback that it is difficult to guarantee the tightness of this approximation.

3. METHODOLOGY

The architecture of our MEIB is illustrated in Fig. 1. Specifically, we consider a general supervised multi-view learning scenario in which we aim to learn a joint yet robust representation z given k different views of features $(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k)$
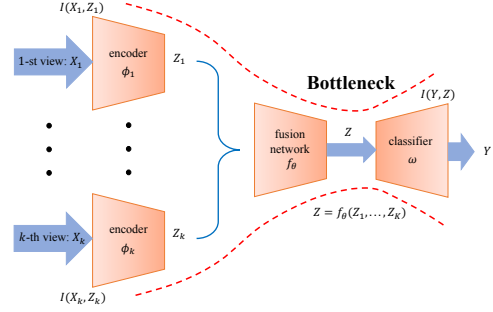

Fig. 1. An illustration of our proposed Multi-view matrix-Entropy based Information Bottleneck (MEIB). MEIB learns a robust joint representation $Z = f_\theta(Z_1, \cdots, Z_k)$ by striking a trade-off between $\max I(Y;Z)$ and $\min \sum_{i=1}^{k} I(X_i;Z_i)$.

and their labels y. To this end, our MEIB consists of k separate encoders $(\phi_1, \phi_2, \cdots, \phi_k)$, a feature fusion network f_θ, and a (nonlinear) classifier w. Each encoder maps view-specific data $\mathbf{x}_i$ ($1 \leq i \leq k$) to a latent representation $\mathbf{z}_i$ to remove noise and redundant information contained in $\mathbf{x}_i$. The latent representations $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_k$ are then fused by a network f_θ to obtain a joint and compact representation $\mathbf{z} = f_\theta(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_k)$.

Note that, in multi-view unsupervised clustering, a common strategy to fuse view-specific latent representations is to take the form of $\mathbf{z} = \sum_{i=1}^{k} w_i \mathbf{z}_i$, s.t., $w_1 + w_2 + \cdots + w_k = 1$ and $w_i \geq 0$ [18]. Here, instead of using the naïve weighted linear combination, our fusion network f_θ is more general and absorbs the linear combiation as a special case.

Therefore, the overall objective of our MEIB can be expressed as:

$$\arg\max_{\phi_1, \cdots, \phi_k, \theta, w} I(Y;Z) - \sum_{i=1}^{k} \beta_i I(X_i;Z_i), \qquad (3)$$
$$\text{s.t.} \quad Z = f_\theta(Z_1, \cdots, Z_k),$$

where β_i refers to regularization parameter for the i-th view.

For the first term $\max I(Y;Z)$, it can be replaced with the risk associated with Z to the prediction performance on Y according to the cross-entropy loss $\text{CE}(Y;\hat{Y})$[1] [22, 23]. Therefore, Eq. (3) is equivalent to:

$$\arg\min_{\phi_1, \cdots, \phi_k, \theta, w} \text{CE}(Y, \hat{Y}) + \sum_{i=1}^{k} \beta_i I(X_i;Z_i), \qquad (4)$$
$$\text{s.t.} \quad Z = f_\theta(Z_1, \cdots, Z_k).$$

In this sense, the main challenge in optimizing Eq. (3) or Eq. (4) is that the exact computation of the compression term $I(X_i;Z_i)$ is almost impossible or intractable due to the high dimensionality of the data.

In this work, we address this issue by simply making use of the matrix-based Rényi's α-order entropy functional

---

[1]The same strategy has also been used in the variational information bottleneck (VIB) [19], the nonlinear information bottleneck (NIB) [20], and deep deterministic information bottleneck (DIB) [21].

to estimate $I(X; Z)$ in each view. Specifically, given N pairs of samples from a mini-batch of size N in the i-th view, i.e., $\{\mathbf{x}^m, \mathbf{z}^m\}_{m=1}^N$, each $\mathbf{x}^m$ be an input sample and each $\mathbf{z}^m$ denotes its latent representation by encoder ϕ^2. We can view both $\mathbf{x}$ and $\mathbf{z}$ as random vectors. According to [14], the entropy of $\mathbf{x}$ can be defined over the eigenspectrum of a (normalized) Gram matrix $K_{\mathbf{x}} \in \mathbb{R}^{N \times N}$ ($K_{\mathbf{x}}(m, n) = \kappa(\mathbf{x}^m, \mathbf{x}^n)$ and κ is a Gaussian kernel) as:

$$H_\alpha(A_{\mathbf{x}}) = \frac{1}{1-\alpha} \log_2 (\text{tr}(A_{\mathbf{x}}^\alpha)) = \frac{1}{1-\alpha} \log_2 \left(\sum_{m=1}^N \lambda_m(A_{\mathbf{x}})^\alpha \right), \tag{5}$$

where $\alpha \in (0, 1) \cup (1, \infty)$. $A_{\mathbf{x}}$ is the normalized version of $K_{\mathbf{x}}$, i.e., $A_{\mathbf{x}} = K_{\mathbf{x}}/\text{tr}(K_{\mathbf{x}})$. $\lambda_m(A_{\mathbf{x}})$ denotes the m-th eigenvalue of $A_{\mathbf{x}}$.

The entropy of $\mathbf{z}$ can be measured similarly on the eigenspectrum of another (normalized) Gram matrix $A_{\mathbf{z}} \in \mathbb{R}^{N \times N}$.

Further, the joint entropy for $\mathbf{x}$ and $\mathbf{z}$ can be defined as:

$$H_\alpha(A_{\mathbf{x}}, A_{\mathbf{z}}) = H_\alpha \left(\frac{A_{\mathbf{x}} \circ A_{\mathbf{z}}}{\text{tr}(A_{\mathbf{x}} \circ A_{\mathbf{z}})} \right), \tag{6}$$

where $\circ$ denotes Hadamard (or element-wise) product.

Given Eqs. (5) and (6), the matrix-based Rényi's α-order mutual information $I_\alpha(X; Z)$ in analogy of Shannon's mutual information is given by:

$$I_\alpha(X; Z) = H_\alpha(A_{\mathbf{x}}) + H_\alpha(A_{\mathbf{z}}) - H_\alpha(A_{\mathbf{x}}, A_{\mathbf{z}}). \tag{7}$$

The differentiability of matrix-based Rényi's α-order entropy functional has been derived in [24]. In practice, the automatic eigenvalue decomposition is embedded in mainstream deep learning APIs like TensorFlow and PyTorch.

Compared with variational approximation or adversarial training that requires an additional auxiliary network (e.g., the mutual information neural estimator or MINE [25]) to approximate a lower bound of mutual information values, Eq. (7) is mathematically well defined and computational efficient, especially for large networks. Moreover, there are only two hyper-parameters here: the order α and the kernel width σ. Throughout this work, we set $\alpha = 1.01$ to approximate the Shannon entropy. For σ, we evaluate the k ($k = 10$) nearest distances of each sample and take the mean. We choose σ as the average of mean values for all samples.

4. EXPERIMENTS

In this section, we compare the performance of our MEIB with several popular multi-view learning methods on both synthetic and real-word datasets. The selected competitors include linear CCA [2], DCCA [5], DCCAE [6], DNN, and deep IB [10]. Here, DNN refers to a network shown in Fig. 1 trained by cross-entropy loss. We repeat the experiment 5 times to obtain average classification errors.

---

^{2}For simplicity, we remove the subscript i of the view index in the remaining of this section.

4.1. Synthetic data

The method for synthesizing data is the same as in [10]. We sample s points per class from $N(0.5e, I)$ or $N(-0.5e, I)$ to form latent representation Z. To distort the classification, we concatenate the extra features ext_Z to the latent representation and synthesize new features D, i.e., $D = [Z, ext\_Z]$. Here the extra features for view v_1 contain $2s/3$ samples from $N(e, I)$ and $4s/3$ samples from $N(-e, I)$. The extra features for view v_2 contain $s/3$ samples from $N(e, I)$ and $5s/3$ samples from $N(-e, I)$. Then the nonlinear transformation of D is accompanied by adding noise, which is sampled from $N(0, t * I)$, where t denotes the noise level. Thus, the generation method of each view can be expressed as $X_i = f(D) + noise$, $i = \{1, 2\}$. Of all the synthetic datasets, 80% are used for training and 20% testing.

4.1.1. Robustness to Noise

To evaluate the performances of all competing methods on their robustness to noise, we change the noise level t which is calculated as $t = a \times \max(abs(v_i))$, where a is set to be $[0.2\,0.4\,0.6\,0.8\,1.0\,1.2]$. The dimension of extra features and latent representation are 5 and 20, respectively. The sample size s per class is set to be 500. So view v_1 can be expressed as $v_1 = tanh(tanh(Z; ext\_Z)) + 0.1$ and view v_2 can be expressed as $sigmoid(Z; ext\_Z) - 0.5$. β_1 and β_2 are tuned in $[1e-5, 5e-5, 1e-4, 5e-4, 1e-3, 5e-3, 1e-2]$. For a fair comparison, in DNN, deep IB and our MEIB, the encoder for each view is a MLP with fully connected layers of the form $512 - 512 - 512$ for view v_1 and a single hidden layer 512 for view v_2. The fusion layer to classification layer is the form of $256 - 10$ and all of the activate function is $ReLU$. The experi-
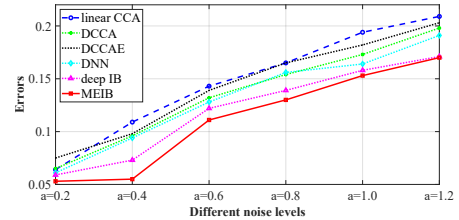

Fig. 2. "Classification error v.s. noise level" for all methods.

mental results of noise robustness are shown in Fig. 2. We can see that the performance of all competing methods decreases with the increase of noise level a. However, our MEIB and deep IB are consistently better than others, which indicates the advantage of bottleneck regularization, i.e., $I(X; Z)$, in suppressing noise.

4.1.2. Robustness to Redundant Dimensions

We further test the robustness of all competing methods to redundant dimensions. To this end, we change the extra features' dimension from 5 to 55, with an interval of 10. All

4320

other settings are the same as in Section 4.1.1, except that the noise factor a is fixed to be 1. The experimental results are shown in Fig. 3 (left). Again, our MEIB and deep IB are superior to other methods. We also plot the ℓ_2 norm of the learned weights (by MEIB) associated with input dimensions 1-20 (correspond to informative features) and dimensions 21-75 (correspond to additional redundant features). As can be seen in Fig. 3 (right), our MEIB is able to identify redundant features dimensions and put small network weights therein.
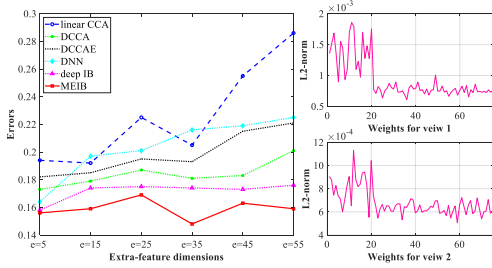

Fig. 3. Robustness to redundant dimensions

4.2. Real-world dataset

We evaluate our proposed method on 3 real-world multi-view datasets. **MNIST** has 2 views. The first view is obtained by rotating raw images with an angle of θ following a uniform distribution in the range $[-\pi/4, \pi/4]$; the second view is obtained by adding uniformly distributed background noise. **BBC**[3] consists of $2,225$ documents from BBC news in five topical areas. The first view has $4,659$ features and the second view has $4,633$ features. **Wisconsin X-Ray MicroBeam (XRMB)** [26] consists of 2 views. The first view has 273 features and the second view has 112 features. The performance of different methods is shown in Table 1, which further indicates the superiority of our MEIB.

Table 1. Average classification error in real-world datasets

Methods	MNIST	BBC	XRMB
linear CCA	0.423 ± 0.002	0.168 ± 0.002	0.461 ± 0.007
DCCA	0.255 ± 0.002	0.346 ± 0.012	0.379 ± 0.008
DCCAE	0.374 ± 0.004	0.279 ± 0.002	0.383 ± 0.001
DNN	0.195 ± 0.018	0.079 ± 0.011	0.222 ± 0.010
deep IB	0.194 ± 0.022	0.077 ± 0.018	0.217 ± 0.008
MEIB	$\mathbf{0.173\pm0.022}$	$\mathbf{0.076\pm0.023}$	$\mathbf{0.181\pm0.010}$

5. CONCLUSION AND FUTURE WORK

We developed the Multi-view matrix-Entropy based Information Bottleneck (MEIB) for multi-view learning. Using

the recently proposed matrix-based Rényi's α-order entropy functional, MEIB can be optimized directly by SGD or ADAM, without variational approximation or adversarial training. Empirical results show that our MEIB outperforms other competing methods in terms of robustness to noise and redundant information contained in source data.

Our study also raise a few insights (or open problems) concerning the general IB approach and its implementation: **1) When is the matrix-based entropy necessary?** As show in Fig. 4, our advantages become weak with the increase of the number of samples in each view. This makes sense, because the increased sample size makes the variational approximation or the adversarial training with an additional network becomes stable. The lower bound of mutual information values also becomes tighter.
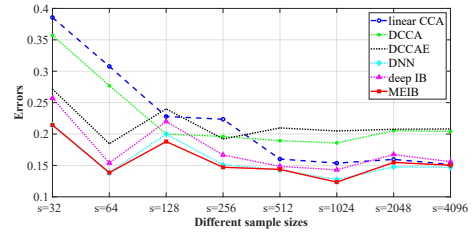

Fig. 4. Classification error (on synthetic data) of all competing methods with respect to sample sizes s in each view.

2) Practical concern for more than 2 views? Extending existing multi-view learning methods to more than 2 views is always a challenging problem [12]. Although our framework can be simply extended to arbitrary number of views by just add a new regularization term $I(X_{\text{new}}; Z_{\text{new}})$, it causes an additional hyper-parameter β_{new} as well. Although we observed that the performance of our MEIB is stable for a suitable range of β_i (Fig. 5 right), we also observed that increasing the number of views does not continually to improve classification performance (Fig. 5 left). This suggests that the IB regularization term $I(X; Z)$ alone is insufficient. One possible way is to take into consideration the "synergistic" information [27] amongst all views in the future.
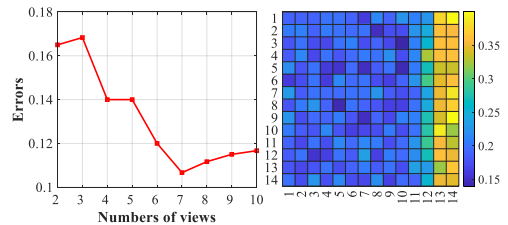

Fig. 5. Left: error with respect to number of views for MEIB; Right: error with respect to different values of β_1 and β_2 (tuned in $[0, 1e-5, 5e-5, 1e-4, 5e-4, 1e-3, 5e-3, 1e-2, 5e-2, 0.01, 0.05, 0.1, 1, 10]$).

---

[3] http://mlg.ucd.ie/datasets/bbc.html

6. REFERENCES

[1] Changhee Lee and Mihaela Schaar, "A variational information bottleneck approach to multi-omics data integration," in *AISTATS*. PMLR, 2021, pp. 1513–1521.

[2] Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan, "Multi-view clustering via canonical correlation analysis," in *ICML*, 2009, pp. 129–136.

[3] Raman Arora and Karen Livescu, "Kernel cca for multi-view learning of acoustic features using articulatory measurements," in *SMLSLP*, 2012.

[4] Xiaoqiang Yan, Shizhe Hu, Yiqiao Mao, Yangdong Ye, and Hui Yu, "Deep multi-view learning methods: a review," *Neurocomputing*, 2021.

[5] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu, "Deep canonical correlation analysis," in *ICML*. PMLR, 2013, pp. 1247–1255.

[6] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes, "On deep multi-view representation learning," in *ICML*. PMLR, 2015, pp. 1083–1092.

[7] Naftali Tishby, Fernando C Pereira, and William Bialek, "The information bottleneck method," in *Allerton*, 1999, pp. 368–377.

[8] Chang Xu, Dacheng Tao, and Chao Xu, "Large-margin multi-view information bottleneck," *IEEE TPAMI*, vol. 36, no. 8, pp. 1559–1572, 2014.

[9] Zhengzheng Lou, Yangdong Ye, and Xiaoqiang Yan, "The multi-feature information bottleneck with application to unsupervised image categorization," in *IJCAI*, 2013.

[10] Qi Wang, Claire Boudreau, Qixing Luo, Pang-Ning Tan, and Jiayu Zhou, "Deep multi-view information bottleneck," in *SDM*. SIAM, 2019, pp. 37–45.

[11] Inaki Estella Aguerri and Abdellatif Zaidi, "Distributed variational representation learning," *IEEE TPAMI*, vol. 43, no. 1, pp. 120–138, 2019.

[12] M Federici, A Dutta, P Forré, N Kushmann, and Z Akata, "Learning robust representations via multi-view information bottleneck," in *ICLR*, 2020.

[13] Jingqi Song et al., "Multicolor image classification using the multimodal information bottleneck network (mmib-net) for detecting diabetic retinopathy," *Optics Express*, vol. 29, no. 14, pp. 22732–22748, 2021.

[14] Luis Gonzalo Sanchez Giraldo, Murali Rao, and Jose C Principe, "Measures of entropy from data using infinitely divisible kernels," *IEEE TIT*, vol. 61, no. 1, pp. 535–548, 2014.

[15] Ran Gilad-Bachrach, Amir Navot, and Naftali Tishby, "An information theoretic tradeoff between complexity and accuracy," in *Learning Theory and Kernel Machines*, pp. 595–609. Springer, 2003.

[16] Shiliang Sun, "A survey of multi-view machine learning," *NCAA*, vol. 23, no. 7, pp. 2031–2038, 2013.

[17] Zhibin Wan, Changqing Zhang, Pengfei Zhu, and Qinghua Hu, "Multi-view information-bottleneck representation learning," in *AAAI*, 2021, pp. 10085–10092.

[18] Daniel J Trosten, Sigurd Lokse, Robert Jenssen, and Michael Kampffmeyer, "Reconsidering representation alignment for multi-view clustering," in *CVPR*, 2021, pp. 1255–1265.

[19] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy, "Deep variational information bottleneck," in *ICLR*, 2017.

[20] Artemy Kolchinsky, Brendan D Tracey, and David H Wolpert, "Nonlinear information bottleneck," *Entropy*, vol. 21, no. 12, pp. 1181, 2019.

[21] Xi Yu, Shujian Yu, and José C Príncipe, "Deep deterministic information bottleneck with matrix-based entropy functional," in *ICASSP*. IEEE, 2021, pp. 3160–3164.

[22] Alessandro Achille and Stefano Soatto, "Information dropout: Learning optimal representations through noisy computation," *IEEE TPAMI*, vol. 40, no. 12, pp. 2897–2905, 2018.

[23] Rana Ali Amjad and Bernhard C Geiger, "Learning representations for neural network-based classification using the information bottleneck principle," *IEEE TPAMI*, vol. 42, no. 9, pp. 2225–2239, 2019.

[24] Shujian Yu, Francesco Alesiani, Xi Yu, Robert Jenssen, and Jose Principe, "Measuring dependence with matrix-based entropy functional," in *AAAI*, 2021, pp. 10781–10789.

[25] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm, "Mutual information neural estimation," in *ICML*. PMLR, 2018, pp. 531–540.

[26] Weiran Wang, Raman Arora, Karen Livescu, and Jeff A Bilmes, "Unsupervised learning of acoustic features via deep canonical correlation analysis," in *ICASSP*. IEEE, 2015, pp. 4590–4594.

[27] Paul L Williams and Randall D Beer, "Nonnegative decomposition of multivariate information," *arXiv preprint arXiv:1004.2515*, 2010.