Multi-View Representation Learning via Total Correlation Objective

Anonymous Author(s)
Affiliation
Address
email

Abstract

Multi-View Representation Learning (MVRL) aims to discover a shared representation of observations from different views with the complex underlying correlation. In this paper, we propose a variational approach which casts MVRL as maximizing the amount of total correlation reduced by the representation, aiming to learn a shared latent representation that is informative yet succinct to capture the correlation among multiple views. To this end, we introduce a tractable surrogate objective function under the proposed framework, which allows our method to fuse and calibrate the observations in the representation space. From the information theoretic perspective, we show that our framework subsumes existing multi-view generative models. Lastly, we show that our approach straightforwardly extends to the Partial MVRL (PMVRL) setting, where the observations are missing without any regular pattern. We demonstrate the effectiveness of our approach in the multi-view translation and classification tasks, outperforming strong baseline methods.

1 Introduction

Multi-View Representation Learning (MVRL) aims to learn a shared representation of multiple observations from different types of views, which should be informative enough to represent the observations. In MVRL, it is important to encourage the representation to be comprehensive enough to capture the correlation across views without losing view-specific information so that the learned representation can be readily applied to rich set of downstream tasks. For example, in sensor fusion for the multi-sensor system [51] and clinical diagnosis based on patients' various types of medical records [47, 49], one would like to aggregate all the information from various observations in order to uncover common factors that correlate them. Although using views as many as possible seems to be always beneficial for the task of learning a good representation, it can make the problem itself harder depending on the complexity of correlations across all views.

MVRL becomes even more challenging when the model does not always have access to complete observations from all views for every data instance during training, which we call Partial Multi-View Representation Learning (PMVRL). PMVRL is closer to the practical setting since it is unrealistic to expect that observations from all different kinds of views are always available. For example, it is unlikely for all the sensors in a sensor-based system to have the same frequency to update their measurements or for all kinds of medical records to be available for any patient.

Despite recent advances in MVRL, most of the existing methods have their own limits, which are (1) no support on more than two views [3, 18, 41], (2) no support on training and testing with missing observations [1, 3, 14, 18, 41], or (3) reliance on label information in the data [14, 50].

In this paper, we address the problem of MVRL with a principled approach grounded in information theory. Specifically, we formulate the representation learning task as maximizing the reduction in

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Total Correlation [13, 38, 39], and derive a novel objective function that not only offers tractable optimization but also introduces multiple types of variational information bottlenecks with interpretable roles. We then show that our method naturally extends to the PMVRL setting via inverse variance weighting, a classical approach used in sensor fusion. We demonstrate the validity and effectiveness of our method in the multi-view translation and classification downstream tasks. In particular, we show that our method achieves state-of-the-art performance in the classification task even without label information when training our model.

2 Approach

Let $\vec{o} = \{o_v\}_{v=1}^{V}$ be the observation of a data instance composed of V different views, which is sampled from an unknown joint distribution $p_D(\vec{o})$, where we emphasize that this is a data distribution using the subscript D. Given these observations, MVRL is the task of learning a *complete* representation z across the views $\vec{o}$. We define a representation as being complete if it can describe all factors of variations in $\vec{o}$. While MVRL considers complete observations for training instances, PMVRL assumes otherwise, i.e. some views being missing in the training data. This poses a unique challenge of (1) learning to produce the complete representation with partial views (2) whose availability varies per data in training and testing.

We first relate informativeness of the representation measured in terms of Total Correlation (TC) and the goal of MVRL (Section 2.1). We then show that the TC-based MVRL objective function encompasses existing multimodal generative models and analyze its limits based on a straightforward variational lower bound (Section 2.2). To resolve those limits, we derive an alternative variational lower bound which suits better for MVRL (Section 2.3). Lastly, we finalize our formulation by proposing the representation aggregation model that naturally extends to PVMRL (Section 2.4).

2.1 Multi-View Representation Learning with Total Correlation

A complete representation Z should be informative enough to explain the correlation among V different views. Total correlation (TC) [42], defined as the Kullback-Leibler divergence of the joint distribution from the factored marginals, measures the amount of information shared among a finite set of random variables. In our MVRL context, TC is defined as

$$TC(\vec{O}) \triangleq D_{KL}\left[p_D(\vec{o}) \parallel \prod_{v=1}^{V} p_D(o_v)\right]. \tag{1}$$

We aim to find the encoder $p_\theta(z|\vec{o})$ such that the knowledge of z would reduce TC as much as possible. This can be formulated by maximizing the objective

$$TC_\theta(\vec{O}; Z) \triangleq TC(\vec{O}) - TC_\theta(\vec{O} \mid Z), \tag{2}$$

where the conditional TC in the last term is given by

$$TC_\theta(\vec{O}|Z) \triangleq \mathbb{E}_{p_\theta(z)}\left[D_{KL}\left[p_\theta(\vec{o}|z) \parallel \prod_{v=1}^{V} p_\theta(o_v|z)\right]\right], \tag{3}$$

which is the expected Kullback-Leibler divergence of the joint conditional from the factored conditionals. The parameterized distributions in the above formula involve the encoder $p_\theta(z|\vec{o})$ as follows: $p_\theta(z) = \int p_\theta(z|\vec{o})\, p_D(\vec{o})\, d\vec{o}$, $p_\theta(\vec{o}|z) = p_\theta(z|\vec{o})\, p_D(\vec{o}) / p_\theta(z)$, and $p_\theta(o_v|z) = \int p_\theta(\vec{o}|z)\, d\vec{o}_{\setminus v}$.

Intuitively, minimization of Eq. (3) (*i.e.*, maximization of Eq. (2)) suits well for MVRL, since (1) any complete representation Z would minimize Eq. (3) (*e.g.*, $Z = \vec{o}$), and (2) it accords with the theoretical result that complete representation z should factorize the generative distribution [44, 50].

Further decomposition of (2) reveals that it encourages Z to encode the correlation across views [13], a desirable property for MVRL:

$$TC_\theta(\vec{O}; Z) = \sum_{v=1}^{V} I_\theta(O_v; Z) - I_\theta(\vec{O}; Z). \tag{4}$$

The first term in Eq. (4), Mutual Information (MI) between each observation and the representation, enforces the representation to be informative for every observation. On the other hand, the second

term takes the role of Information Bottleneck (IB), which encourages the encoder to learn minimal sufficient representation [37]. Consequently, simultaneous optimization of both terms naturally allows the representation Z to capture correlations among views while being minimally sufficient. Note that when $V = 2$, Eq. (4) coincides with Interaction Information [4, 19, 25, 36], which quantifies the amount of information shared among two views and their joint representation.

2.2 Limitations of VAE models for multi-view data

Unfortunately, a direct optimization of MIs in Eq. (4) is intractable [2, 12]. A straightforward approach would be employing approximate distributions $q_\phi^v\left(o_v|z\right) \approx p_\theta\left(o_v|z\right)$ and $r\left(z\right) \approx p_\theta\left(z\right)$, and optimize a variational lower bound of Eq. (4) as follows (see Section A.1 in the supplementary material for full derivation):

$$TC_\theta(\vec{O};Z) \geq \sum_{v=1}^{V}\left[H\left(O_v\right) + \mathbb{E}_{p_\theta(z|\vec{o})p_D(\vec{o})}\left[\ln q_\phi^v\left(o_v|z\right)\right]\right] - \underbrace{\mathbb{E}_{p_D(\vec{o})}\left[D_{KL}\left[p_\theta(z|\vec{o})\|r(z)\right]\right]}_{\text{VIB}}, \quad (5)$$

where the entropy terms can be dropped from optimization since they are determined by the true data distribution $p_D\left(\vec{o}\right)$. Without the entropy terms, we note that Eq. (5) is essentially identical to evidence lower bound (ELBO) of the variational auto-encoder (VAE) models for multi-view data [31, 34, 35, 46] by switching the notations, for p for encoder and q for decoder. This notation switch follows the convention in [2, 12].

Although this lower bound contains the variational information bottleneck term that encourages the representation to be minimally sufficient for generalizing well even with small training data [2], it has the following fundamental limitations:

1. **Missing views**: The encoder $p_\theta\left(z|\vec{o}\right)$ requires complete observations $\vec{o}$ not only in training but also in testing phases, while MVRL requires the model to handle incomplete observations $\tilde{o} \subseteq \vec{o}$ in test time. Furthermore, PMVRL requires to handle incomplete observations even in training.
2. **Unbalanced representation**: If a subset of views is overwhelmingly informative enough to reconstruct the others, the encoder $p_\theta\left(z|\vec{o}\right)$ may learn to rely on those views while ignoring the rest, yielding a degenerate solution of Eq. (4). This is problematic in MVRL since such views may not be available at test time.

In order to overcome these challenges, prior methods impose special structures (i.e. inductive hypotheses) for $p_\theta\left(z|\vec{o}\right)$, such as Product of Experts (PoE) [17, 46], Mixture of Experts (MoE) [31], or Mixture of Product of Experts (MoPoE) [35]. In our work, we present a more principled approach to this problem by deriving an alternative lower bound, described in the next section.

2.3 Conditional Information Bottleneck

To resolve the first issue raised in the previous section, we start from Eq. (4) and reformulating it as follows:

$$\begin{aligned} TC_\theta(\vec{O};Z) &= \sum_{v=1}^{V}\left[\frac{V-1}{V}I_\theta\left(O_v;Z\right) + \frac{1}{V}I_\theta\left(O_v;Z\right) - \frac{1}{V}I_\theta\left(\vec{O};Z\right)\right] \\ &= \frac{1}{V}\sum_{v=1}^{V}\left[\left(V-1\right)I_\theta\left(O_v;Z\right) - I_\theta(\vec{O}_{\setminus v};Z|O_v)\right], \end{aligned} \quad (6)$$

where the last equality is due to the chain rule of MI (see Section A.5 in the supplementary material for details). Compared to Eq. (4), Eq. (6) introduces *conditional* IBs for each view in the second term, each of which penalizes the conditional MI between the latent representation and $V-1$ other views given each view. As a result, the latent representation is regularized to be *inferrable* from any view and thus cannot be dominated by a few one, which naturally encourages the encoder p_θ to avoid the imbalance in the learned representation.

Since the conditional MIs in Eq. (6) involve $p_\theta(z|o_v) = \int p_\theta(z|\vec{o})p_D(\vec{o}_{\setminus v}|o_v)d\vec{o}_{\setminus v}$ which requires to compute intractable integration, we use the variational upper bound of those terms as follows (see

Section A.2 in the supplementary material for the full derivation and analysis):

$$TC_\theta(\vec{O};Z) \geq \frac{V-1}{V}\sum_{v=1}^{V}\left[H\left(O_v\right)+\mathbb{E}_{p_\theta(z|\vec{o})p_D(\vec{o})}\left[\ln q_\phi^v\left(o_v|z\right)\right]\right] - \frac{1}{V}\sum_{v=1}^{V}\underbrace{\mathbb{E}_{p_D(\vec{o})}\left[D_{KL}\left[p_\theta(z|\vec{o})\|r_\psi^v(z|o_v)\right]\right]}_{\text{Conditional VIB}}. \tag{7}$$

This lower bound provides a number of benefits over Eq. (5) in handling challenges mentioned in the previous section. First, it introduces the view-specific encoders $r_\psi^v\left(z|o_v\right)$ for each view, allowing us to extract the representation z even when some views are missing in the observation. Second, it uses *forward* KL divergence $D_{KL}[p_\theta\left(z|\vec{o}\right)\|r_\psi^v\left(z|o_v\right)]$ to calibrate each encoder $r_\psi^v\left(z|o_v\right)$ to the joint encoder $p_\theta\left(z|\vec{o}\right)$ in the conditional VIB term, encouraging $r_\psi^v\left(z|o_v\right)$ to cover all the supports or modes of $p_\theta\left(z|\vec{o}\right)$. This property is critically important in (P)MVRL, where one needs to infer the complete representation from the partially available views without being overly confident. We remark that mmJSD [34] adopts *reverse* KL divergence, which is not ideal as we later demonstrate in the experiments.

Finally, although Eq. (7) has aforementioned desirable properties for MVRL, we sometimes observed that the model is prone to overfitting when the size of training data is limited. In order to prevent overfitting, we found that VIB in Eq. (5) is an effective regularization as it favors the minimal sufficient encoding of the representation. Therefore, we formulate our objective function as a convex combination of Eq. (5) and Eq. (7) so that we regularize the training via VIB (see Section A.3 in the supplementary file for full derivation):

$$TC_\theta(\vec{O};Z) \geq \frac{V-\alpha}{V}\sum_{v=1}^{V}\left[H\left(O_v\right)+\mathbb{E}_{p_\theta(z|\vec{o})p_D(\vec{o})}\left[\ln q_\phi^v\left(o_v|z\right)\right]\right] - \frac{\alpha}{V}\sum_{v=1}^{V}\underbrace{\mathbb{E}_{p_D(\vec{o})}\left[D_{KL}\left[p_\theta(z|\vec{o})\|r_\psi^v(z|o_v)\right]\right]}_{\text{Conditional VIB}} - (1-\alpha)\underbrace{\mathbb{E}_{p_D(\vec{o})}\left[D_{KL}\left[p_\theta(z|\vec{o})\|r(z)\right]\right]}_{\text{VIB}}, \tag{8}$$

where α is the hyperparmeter that trades off learning minimal sufficient representation in favor of calibrating r_ψ^v. For simplicity, we model the encoder, decoder, and approximate marginal distributions using the parameterized Gaussians with the diagonal covariance matrix, i.e. $r_\psi^v\left(z|o_v\right)=N\left(\mu_v,\sigma_v^2 I\right)$, $q_\phi^v\left(o_v|z\right)=N\left(\hat{\mu}_v,I\right)$, and $r\left(z\right)=N(0,\mathrm{I})$, respectively.

While the view-specific encoders r_ψ^v allow us to extract the representation from any available view individually, combining these representations (fusion) still remains as a problem. In addition, the use of joint-view encoder $p(z|\vec{o})$ makes our method limited to complete observations $\vec{o}$ during training, which needs to be addressed for PMVRL. In the next section, we show that these issues can be effectively resolved by a simple model design for $p_\theta\left(z|\vec{o}\right)$.

2.4 Models for Joint Encoder

We review the models adopted by prior methods that make the joint representation encoder $p_\theta\left(z|\vec{o}\right)$ amenable to missing views and discuss their strengths and weaknesses.

PoE Product-of-Experts (PoE) [17] combines multiple probability distributions by their product. MVAE [46] models the joint representation encoder as a PoE, treating view-specific encoders as experts. The PoE can produce a sharper distribution as we increase the number of input views, thus an effective method for aggregating information across any subset of view-specific encoders. Assuming each of view-specific encoders as Gaussian distributions such that $r_\psi^v\left(z|o_v\right)=N\left(\mu_v,\sigma_v^2 I\right)$, the PoE joint encoder is obtained by

$$p_\theta\left(z|\vec{o}\right)\triangleq N\left(\mu_p,\sigma_p^2\mathbf{I}\right), \quad \text{where} \quad \mu_p\triangleq\frac{\sum_{v=1}^{V}\mu_v/\sigma_v^2}{\sum_{v=1}^{V}1/\sigma_v^2} \quad \text{and} \quad \sigma_p^2\triangleq\frac{1}{\sum_{v=1}^{V}1/\sigma_v^2}. \tag{9}$$

Eq. (9) is also the formula of Inverse-Variance Weighted (IVW) method [6, 7], a classical method in statistics for aggregating multiple random variables, such as sensor fusion.

Unfortunately, a naive application of the PoE to the ELBO formulation (Eq.(5)) may fail to optimize the individual encoders, which is important in learning the balanced representation. MVAE [46], as an example, randomly samples subsets of views among 2^V combinations and jointly optimizes their ELBOs in order to ensure that all the view-specific encoders are optimized under PoE. However, such treatment may result in a precision miscalibration of view-specific encoders [31].

MoE The Mixture-of-Expert (MoE) takes an arithmetic mean of probability distributions. Assuming Gaussian experts for view-specific encoders, MMVAE [31] and mmJSD [34] adopt MoE as the model for the joint representation encoder, which results in a mixture of Gaussians. In MMVAE, the MoE is trained by pair-wise optimization in such a way that the latent representation from a view-specific encoder can reconstruct the observation in other views as well as its own view. However, this does not necessarily imply that the latent representation successfully aggregates the information across views. mmJSD addresses this issue by adopting a common learnable prior across views.

We remark that, in Eq. (8), modeling $p_\theta\left(z|\vec{o}\right)$ as MoE of $r_\psi^v(z|o_v)$ and setting $\alpha = 0$ yields the ELBO objective version of MMVAE. Assuming the same model for $p_\theta\left(z|\vec{o}\right)$ and $\alpha = 0$, and modeling $r(z)$ as the PoE of $r_\psi^v(z|o_v)$ yields the objective function of mmJSD. For tractable optimization of the KL term involving MoE, mmJSD derives a lower bound on the ELBO by decomposing it into multiple KL terms. This bound can be also obtained from Eq. (8) by setting $\alpha = 1$ and using the *reverse* KL for the conditional VIB terms. However, we empirically show that minimization of the reverse KL is not helpful with learning complete representation in Section 4.

MoPoE The Mixture-of-Product-of-Experts (MoPoE) is a mixture of 2^V combination of PoE experts, which is used as the model for joint encoder in MoPoE-VAE [35]. Since the MoPoE joint encoder takes into account all possible combinations of views, it naturally learns to aggregate information across any given views while optimizing every view-specific encoder. Similar to mmJSD, MoPoE-VAE derives a lower bound on the objective to decompose the KL term into 2^V KL terms with analytic solutions. However, this would render the method intractable for tasks with many views.

Multi-View Total Correlation Auto-Encoder Since our model uses conditional VIBs that explicitly calibrate all the representations encoded by view-specific encoders, we can safely choose PoE as the model for the joint representation encoder without suffering from the precision miscalibration.

We call our resulting model the Multi-View Total Correlation Auto-Encoder (MVTCAE), depicted in Figure 4 in the supplementary material. When training with partial observations, MVTCAE simply treats the covariance matrix of $q_\phi^v\left(o_v|z\right)$ to be $\infty\mathbf{I}$ for any missing observation $o_v \notin \tilde{o}$, so that it does not contribute to the reconstruction loss for missing views, similar to [50].

3 Related Work

Information-Theoretic Representation Learning In representation learning, using regularization based on information theory has appeared in many important tasks, such as learning minimal sufficient / invariant / and disentangled representations. Information Bottleneck (IB) [37] was introduced as a regularization method to obtain minimal sufficient encoding by constraining the amount of information captured by the latent variable from the observed variable.Deep Variational IB (VIB) [2] extended IB by parameterizing it with a neural network, which results in a simple yet effective method for regularization in training neural networks [30]. Furthermore, using VIB in unsupervised learning has been revealed to have close relationships among VIB, VAE [22] and β-VAE [16]. A number of follow-up works [27, 32] propose encouraging the encoders to learn representation invariant to any attribute given in advance. Similarly, a modified version of VIB was introduced [11] for learning a domain-invariant representation across two data domains. While IB and VIB are concerned with computing MI, two similar but distinct generalizations of MI have been applied to learning disentangled representation, which are TC [42] and Interaction Information (II) [4, 25, 36]. The TC quantifies the dependency among all dimensions of the single latent variable, which motivated many works that learn disentangled representation in a single domain [5, 10, 12, 20, 21]. II was used for disentangling shared representation from domain-specific representation in cross-domains [19].

Multi-View Representation Learning Canonical Correlation Analysis (CCA) [18] and its variants [1, 3, 41] are classical approaches for unsupervised cross-view representation learning. CCA projects two different views into one common latent space in a way that those two views are maximally correlated in the latent space. KCCA [1] uses kernels and DCCA [3] uses neural networks

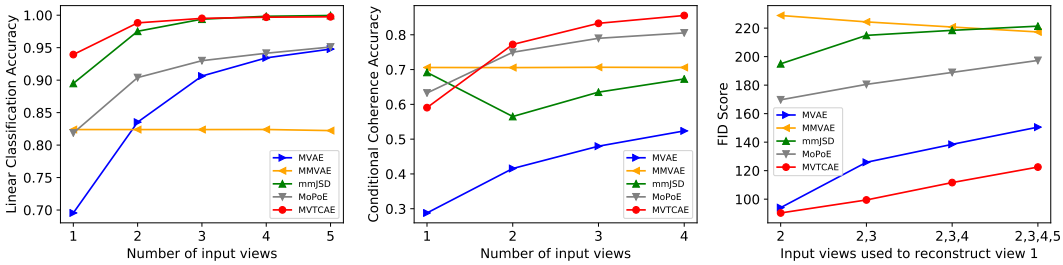


Figure 1: Performance evaluation on the Poly MNIST dataset.

to learn the common representation. Similarly, DCCAE [41] trains autoencoders to obtain common representations. More recently, a number of notable MVRL methods have been proposed to support more than 2 views. DMF-MVC [52] extracts a common representation of multiple views through deep matrix factorization. MDcR [48] maps each view to a lower-dimensional space and applies the kernel matching to regularize the dependence across multiple views. ITML [9] learns a Mahalanobis distance function by Bregman optimization, whereas LMNN [43] learns a Mahalanobis distance metric to optimize the k-nearest neighbors classifier using labeled data. CPM-Nets [50] gives a formal definition of complete representation in PMVRL and proposes to learn it without encoders.

Multi-modal Generative Models (VAEs) MVAE [46] and its variants [31, 34, 35] are multi-modal generative models that learn shared representation by maximizing the log-likelihood of joint modalities via latent variables. We compared these methods to ours in Section 2.2 and 2.4.

4 Experiments

4.1 Multi-View Representation Learning

To verify that our method successfully learns complete representation capturing both common factors and view-specific factors, we evaluate our method on the following two datasets used as evaluation benchmarks in the MVRL literature.

4.1.1 Multi-View Classification / Translation on PolyMNIST

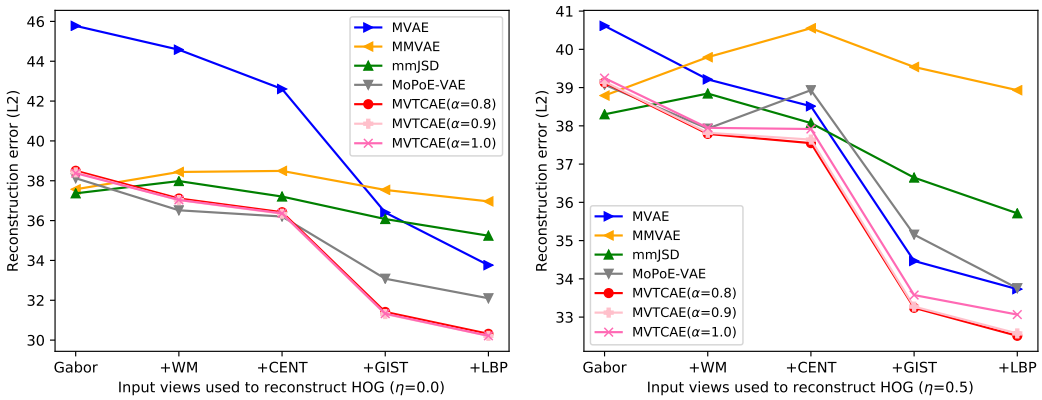
We employ PolyMNIST dataset [35] composed of tuples with 5 different MNIST images, which have different backgrounds and writing style but share the same digit label. The background of each view is randomly cropped from one image which is not used by other views. Thus, the digit identity is the common factor of variation while the background and writing style are view-specific factors. There are 60K tuples of training samples and 10K of test samples. Although the digit ID should be observable in any view, it is hard to identify in some images depending on the background and the writing style. Therefore, aggregating information across views is essential for predicting the label.

Evaluation protocol To evaluate the learned representation, we follow the protocol in [35]. Specifically, after training all the models in an unsupervised manner using complete observations, we evaluate the learned representation with three different metrics, which are linear classification accuracy and conditional coherence accuracy. We also evaluate the quality of conditional generation with FID score [15]. We compare our method with various state-of-the-art multimodal generative models which are MVAE [46], MMVAE [31], mmJSD [34], and MoPoE-VAE [35].

To apply our method to classification, we fix the encoders and train a linear classifier to predict labels using the joint representation extracted from $p_\theta(z|\vec{o})$ feeding complete observations in training set. We then classify the representations of all subsets and compute the average classification accuracy over all subsets with the same subset size.

To measure the conditional coherence accuracy, we extract the representation of every subset of views using p_θ and generate views that are absent in the subset using q_ϕ. Those generated views are fed into the pretrained CNN-based classifier and see if the prediction from the classifier matches the label of the given subset. The results are averaged over all subsets with the same size.

Finally, we evaluate the sample quality of the first view images generated from different combination of input views ({2}, {2,3}, {2,3,4}, {2,3,4,5}) in terms of FID score. Since FID compares statistics



(a) Result of training from complete observations (b) Result of training from partial observations

Figure 2: The multi-view translation performance of training with the complete and partial observations, measuring the reconstruction error of the HOG feature by incrementally adding features.

of two sets (one is the set of samples generated by the models and the other is the first view images in the training set in our context), it takes into account not only the quality of generated samples but also the diversity of them. Thus, unlike two previous evaluation metrics, marking a lower (better) FID score requires the model to learn the representation that captures not only the common factors of variation but also view-specific factors to express diversity within the view.

Results The left plot in Figure 1 shows the result of linear classification. Using the PoE as a joint representation encoder same as MVAE, our method clearly outperforms all the baseline methods in the classification task, reaching 94% and 99% accuracy even when only one view and two views are given respectively. Comparing to MVAE, this result implies that conditional VIBs in our method calibrate every view-specific encoders according to the information shared across views so that each of them successfully captures the digit identity, resolving the issue of unbalanced representation. On the other hand, while all the methods show monotonic improvement on its performance as the number of given views increases, the accuracy of MMVAE does not show any noticeable change. It is remarkable that our method even outperforms mmJSD and MoPoE-VAE with simpler aggregation model for the joint encoder, although the two prior methods use both PoE and MoE.

The middle plot in Figure 1 summarizes the result of conditional coherence. Our method outperforms all the baseline methods except when only one view is given. mmJSD and MoPoE fail to leverage additional input views, showing the performance staying flat or even degrading, due to the incapability of aggregating information inherent in MoE as we discussed in Section 2.4. In contrast, our method monotonically improves the coherence accuracy with more input views, implying that the view-specific encoders are well calibrated by conditional VIBs so that the joint representation aggregated by PoE produces accurate encoding of the digit identity.

The right plot in Figure 1 presents the result of FID scores. Our method achieves the best performance in any combination of input views. In the supplementary material section D.1, we compare the samples generated by methods that qualitatively show MVTCAE is much better at (1) identifying the shared factors in the given views and (2) expressing the variation of factors in the target views.

4.1.2 Multi-View Translation on Caltech-101

We also evaluate our method on the multi-view dataset used in [24] where six visual features are extracted from images in Caltech-101 dataset. In this dataset, each image is associated with six features, which are of Gabor filter [29], Wavelet Moments (WM) [26], CENTRIST [45], Histogram of Oriented Gradients (HOG) [8], GIST [29], and Local Binary Pattern (LBP) [28]. We treat each feature as a view, varying from 40 up to 1984 dimensions. We remark that each feature can be considered as a lossy compression of the original image, where the extracted information from one view may not be necessarily inferrable for other views, thus making it nontrivial to learn the correlation across views.

Evaluation protocol Given the partial observations $\tilde{o}$ at test time, we extract the joint representation z by Eq. (9), and reconstruct the missing view $o_v \notin \tilde{o}$ using the output of the decoder $q_\theta^v(o_v|z)$ dedicated to view v. We then evaluate the L2 reconstruction error. We investigate the impact of missing views on the completeness of the representation. We train our model with complete

observations, and apply the model to reconstruct the HOG feature[1] missing in the test time using partial observations $\tilde{o} \in \vec{o}_{\setminus \text{HOG}}$. The performance is averaged over the results of 10 independent runs. We compare to the same baseline methods employed in Section 4.1.

Results Figure 2a shows the impact of missing views at test time. We observe that our method effectively decreases the reconstruction error as the input views accumulates in the observation, only with slightly higher error in reconstruction with one view. Interestingly, we notice that the reconstruction error reduces significantly at some point (*i.e.*, when adding GIST feature to the observation). It implies that some views contain more information than the others, and our method is able to learn to utilize this property. On the other hand, MoPoE-VAE is less effective in utilizing GIST or LBP features, while mmJSD and MMVAE hardly show monotonic performance improvement. Remarkably, MVAE performs worst, although the joint encoder model is same as ours, i.e. PoE. As an ablation study, we explore the impact of weighting parameter α, which trades off between VIB and conditional VIB (Eq. (8)) in Section C.1 in the supplementary material.

4.2 Partial Multi-View Representation Learning

To verify that our method can be effectively generalized to the partial observation setting, we conduct experiments on two tasks: multi-view *translation* and *classification*. In both tasks, we aim to show that the representation learned by our method is complete enough to infer the missing views given the incomplete observations (translation task) and useful for downstream tasks (classification task). To simulate the partial observations in both tasks, we follow the protocol of Zhang et al. [50] and generate random view-missing patterns with missing rate $\eta = \sum_{v=1}^{V} U_v/(V \times S)$ $(0 \leq \eta < 1)$, where S is the number of entire samples and U_v is the number of samples missing in the v-th view.

4.2.1 Multi-View Translation on Caltech-101

In this experiment, we investigate if our model can infer complete representations from partial observations. To quantify the completeness of the representation, we employ the multi-view translation task. The goal is to reconstruct a missing view using the representation extracted from other view(s). In this case, the reconstruction error serves as a direct measurement of the completeness.

Evaluation protocol Following the same procedures of Section 4.1.2, we repeat the same experiments but using the model trained with incomplete observation ($\eta = 0.5$). This experiment can demonstrate the robustness of our model to missing views in both inference and training time.

Results Figure 2b summarizes the experimental results using the incomplete training data ($\eta = 0.5$). We observe that our method exhibits similar trends with complete observation training data, exhibited by monotonic performance improvement in the number of available views, while other methods get negatively affected by additional views when trained with incomplete data (MMVAE, mmJSD, MoPoE-VAE). Interestingly, we observe that MVAE performs better when trained with incomplete data. This is because the missing views in the training data serve as a regularization similar to Dropout [33] or sub-sampled training paradigm [46]. Although MVAE shows monotonic improvement as ours, its performance is turned out to be not comparable to ours.

We also note that our method achieves the best performance with $\alpha = 0.8$ when trained with incomplete data, while the best is achieved at $\alpha = 0.9$ when trained with complete data. We suspect that this is because optimizing conditional VIB with partial observation becomes more difficult than with complete observation, and the model tends to solve it via memorization. In this case, the VIB can be useful to improve the generalization.

To summarize, our method achieves the best performance for intermediate $0 < \alpha < 1$, which implies that simultaneous optimization of VIB and conditional VIBs in Eq. (8) is effective to solve MVRL and PMVRL. Qualitative results can be found in Section D.2 in the supplementary material.

4.2.2 Multi-View Classification on Six Datasets [50]

The previous section suggests that our method is able to learn to calibrate representations across views, even when the training data is composed of incomplete observations. In this section, we evaluate the effectiveness of our approach as an unsupervised pre-training of the multi-view representation, investigating if the learned representation is useful for downstream tasks such as classification.

[1] We choose the HOG since it has the largest dimension thus the reconstruction is most nontrivial.

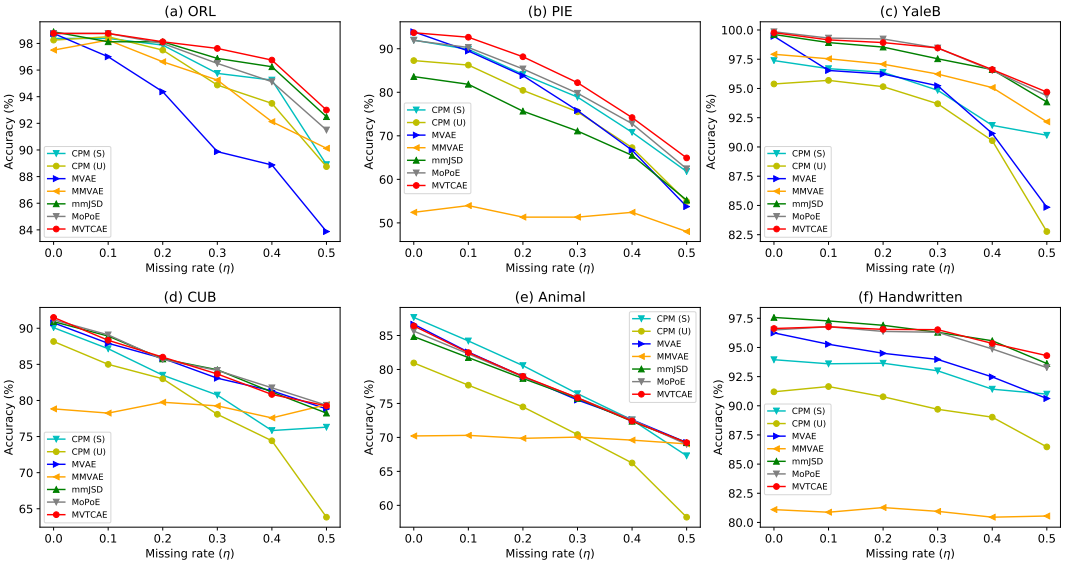


Figure 3: Comparison on classification performance under various view missing rate.

Datasets We evaluate our method on six feature-based image classification datasets used in multi-view learning [50], which are **ORL**, **PIE**, **YaleB**, **CUB**, **Animal**, and **Handwritten**. Each dataset is associated with 2 and up to 6 visual features. Similar to the previous section, we treat each feature as a view of data. For all datasets, we follow the same preprocessing and training/test splits used in [50]. See Section E in the supplementary material for a comprehensive description of the datasets.

Evaluation protocol We follow [50] to evaluate our model trained with partial observations with various missing rates $\eta = \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. Similar to the experiment in Section 4.1, we first train our model in an unsupervised manner using the observable views. Then we fix the encoders, and train the classifier with labels using the learned representation. To isolate the impact of the additional classifier, we employ the simplest classifier, i.e. logistic regression with the learned representation as input. We report the performance by averaging the results from 10 independent runs.

Result Figure 3 summarizes the results under varying η. In addition to the baseline methods compared in previous sections, we include two strong baselines, CPM-Nets(S) and CPM-Nets(U), which address the PMVRL with and without label information in the representation learning stage, respectively. Compared to the supervised baseline (CPM-Nets(S)), our method achieves competitive performance on PIE and Animal datasets, and much better performance in the rest of the datasets, although our model is trained in a purely unsupervised manner. Compared to the unsupervised baseline methods (CPM-Nets(U), MVAE, MMVAE, mmJSD, and MoPoE), our method achieves noticeable improvements in datasets with more than two views (ORL, PIE, YaleB, Handwritten), especially when the missing rate is reasonably high ($0.3 \leq \eta \leq 0.5$). It shows that our method is robust in learning the cross-view correlation under partial observations, and the learned representation is informative enough to be useful in downstream tasks, even though we do not use label information in the data or adopt a sophisticated aggregation model in the joint encoder.

5 Conclusion

We presented an information theoretic model for unsupervised partial multi-view representation learning. Based on Total Correlation (TC), we derived a novel variational lower bound that allows us to train the model that encodes complete latent representation from partial-view observations. Strictly trained in an unsupervised manner, we also demonstrated that the learned representation is highly effective in downstream tasks, such as multi-view classification and multi-view translation. Although we demonstrated that our method can even learn from partial multi-view data, it still has room for improvement such as learning from unaligned view data, and investigating more sophisticated representation aggregation models for the joint encoder, which we leave as future work.

References

[1] S. Akaho. A kernel method for canonical correlation analysis. *International Meeting of Psychometric Society*, 2001.

[2] A. Alemi, I. Fischer, J. Dillon, and K. Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

[3] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255. PMLR, 2013.

[4] A. Bell. The co-information lattice, 921–926. In *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Source Separation (ICA2003), Nara, Japan*, 2003.

[5] R. T. Q. Chen, X. Li, R. Grosse, and D. Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, 2018.

[6] W. G. Cochran. The combination of estimates from different experiments. *Biometrics*, 10(1): 101–129, 1954.

[7] W. G. Cochran and S. P. Carroll. A sampling investigation of the efficiency of weighting inversely as the estimated variance. *Biometrics*, 9(4):447–459, 1953.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[9] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning*, pages 209–216, 2007.

[10] B. Esmaeili, H. Wu, S. Jain, A. Bozkurt, N. Siddharth, B. Paige, D. H. Brooks, J. Dy, and J.-W. van de Meent. Structured disentangled representations. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2525–2534. PMLR, 16–18 Apr 2019.

[11] M. Federici, A. Dutta, P. Forré, N. Kushman, and Z. Akata. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*, 2020.

[12] S. Gao, R. Brekelmans, G. V. Steeg, and A. Galstyan. Auto-encoding total correlation explanation. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1157–1166. PMLR, 16–18 Apr 2019.

[13] S. Gao, R. Brekelmans, G. Ver Steeg, and A. Galstyan. Auto-encoding total correlation explanation. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1157–1166. PMLR, 2019.

[14] Z. Han, C. Zhang, H. Fu, and J. T. Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=OOsR8BzCnl5`.

[15] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017.

[16] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations*, 2(5):6, 2017.

[17] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[18] H. Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. Springer, 1992.

[19] H. Hwang, G.-H. Kim, S. Hong, and K.-E. Kim. Variational interaction information maximization for cross-domain disentanglement. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[20] Y. Jeong and H. O. Song. Learning discrete and continuous factors of data via alternating disentanglement. In *International Conference on Machine Learning (ICML)*, 2019.

[21] H. Kim and A. Mnih. Disentangling by factorising. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2649–2658, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[22] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In Y. Bengio and Y. LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[23] F.-F. Li, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

[24] Y. Li, F. Nie, H. Huang, and J. Huang. Large-scale multi-view spectral clustering via bipartite graph. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[25] W. McGill. Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory*, 4(4):93–111, 1954.

[26] H. A. Moghaddam, T. T. Khajoie, A. H. Rouhi, and M. S. Tarzjan. Wavelet correlogram: a new approach for image indexing and retrieval. *Pattern Recognition*, 38(12):2506–2518, 2005.

[27] D. Moyer, S. Gao, R. Brekelmans, A. Galstyan, and G. Ver Steeg. Invariant representations without adversarial training. In *Advances in Neural Information Processing Systems*, pages 9084–9093, 2018.

[28] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

[29] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[30] X. B. Peng, A. Kanazawa, S. Toyer, P. Abbeel, and S. Levine. Variational discriminator bottleneck: Improving imitation learning, inverse RL, and GANs by constraining information flow. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=HyxPx3R9tm`.

[31] Y. Shi, N. Siddharth, B. Paige, and P. H. Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in Neural Information Processing Systems*, 2019.

[32] J. Song, P. Kalluri, A. Grover, S. Zhao, and S. Ermon. Learning controllable fair representations. *international conference on artificial intelligence and statistics*, 2018.

[33] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15 (56):1929–1958, 2014. URL `http://jmlr.org/papers/v15/srivastava14a.html`.

[34] T. M. Sutter, I. Daunhawer, and J. E. Vogt. Multimodal generative learning utilizing jensen-shannon-divergence. 2020.

[35] T. M. Sutter, I. Daunhawer, and J. E. Vogt. Generalized multimodal elbo. 2021.

[36] H. K. Ting. On the amount of information. *Theory of Probability & Its Applications*, 7(4): 439–447, 1962.

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

[38] G. Ver Steeg and A. Galstyan. Discovering structure in high-dimensional data through correlation explanation. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 577–585. Curran Associates, Inc., 2014. URL `https://proceedings.neurips.cc/paper/2014/file/4f6ffe13a5d75b2d6a3923922b3922e5-Paper.pdf`.

[39] G. Ver Steeg and A. Galstyan. Maximally informative hierarchical representations of high-dimensional data. In *Artificial Intelligence and Statistics*, pages 1004–1012. PMLR, 2015.

[40] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[41] W. Wang, R. Arora, K. Livescu, and J. Bilmes. On deep multi-view representation learning. In *International conference on machine learning*, pages 1083–1092. PMLR, 2015.

[42] S. Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4(1):66–82, 1960.

[43] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2), 2009.

[44] M. White, Y. Yu, X. Zhang, and D. Schuurmans. Convex multi-view subspace learning. In *Nips*, pages 1682–1690. Lake Tahoe, Nevada, 2012.

[45] J. Wu and J. M. Rehg. Centrist: A visual descriptor for scene categorization. *IEEE transactions on pattern analysis and machine intelligence*, 33(8):1489–1501, 2010.

[46] M. Wu and N. Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in Neural Information Processing Systems*, 2018.

[47] Y. Yuan, G. Xun, K. Jia, and A. Zhang. A multi-view deep learning framework for eeg seizure detection. *IEEE journal of biomedical and health informatics*, 23(1):83–94, 2018.

[48] C. Zhang, H. Fu, Q. Hu, P. Zhu, and X. Cao. Flexible multi-view dimensionality co-reduction. *IEEE Transactions on Image Processing*, 26(2):648–659, 2016.

[49] C. Zhang, E. Adeli, T. Zhou, X. Chen, and D. Shen. Multi-layer multi-view classification for alzheimer's disease diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[50] C. Zhang, Z. Han, y. cui, H. Fu, J. T. Zhou, and Q. Hu. Cpm-nets: Cross partial multi-view networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 559–569. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/11b9842e0a271ff252c1903e7132cd68-Paper.pdf`.

[51] H. Zhang, T. S. Huang, N. M. Nasrabadi, and Y. Zhang. Heterogeneous multi-metric learning for multi-sensor fusion. In *14th International Conference on Information Fusion*, pages 1–8. IEEE, 2011.

[52] H. Zhao, Z. Ding, and Y. Fu. Multi-view clustering via deep matrix factorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Please read abstract and introduction
   (b) Did you describe the limitations of your work? [Yes] Please read conclusion
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Due to the space limit, we discuss it in the supplementary material.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See supplementary.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] we will submit the code as the supplementary file.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] we specify them in the supplementary material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We ran 5 10 runs with different random seeds. Related information can be found in the supplementary file.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] included the supplementary material.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] In appropriate places
   (b) Did you mention the license of the assets? [N/A] Not applicable.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] All the datasets we used are public.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] None of datasets contain identifiable content.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] Not applicable.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] Not applicable.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] Not applicable.