Self-supervised Disentanglement of Modality-specific and Shared Factors Improves Multimodal Generative Models

Imant Daunhawer, Thomas M. Sutter, Ričards Marcinkevičs, Julia E. Vogt

Department of Computer Science, ETH Zurich

Abstract. Multimodal generative models learn a joint distribution over multiple modalities and thus have the potential to learn richer representations than unimodal models. However, current approaches are either inefficient in dealing with more than two modalities or fail to capture both modality-specific and shared variations. We introduce a new multimodal generative model that integrates both modality-specific and shared factors and aggregates shared information across any subset of modalities efficiently. Our method partitions the latent space into disjoint subspaces for modality-specific and shared factors and learns to disentangle these in a purely self-supervised manner. Empirically, we show improvements in representation learning and generative performance compared to previous methods and showcase the disentanglement capabilities.

1 Introduction

The promise of multimodal generative models lies in their ability to learn rich representations across diverse domains and to generate missing modalities. As an analogy, humans are able to integrate information across senses to make more informed decisions [33], and exhibit cross-modal transfer of perceptual knowledge [41]; for instance, people can visualize objects given only haptic cues [42]. For machine learning, multimodal learning is of interest in any setting where information is integrated across two or more modalities.

Alternatives to multimodal generative models include unimodal models with late fusion or with coordinated representations, as well as conditional models that translate between pairs of modalities [3]. Yet, both alternatives have disadvantages compared to multimodal approaches. While unimodal models cannot handle missing modalities, conditional models only learn a mapping between sources, and neither integrate representations from different modalities into a joint representation. In contrast, multimodal generative models approximate the joint distribution and thus implicitly provide the marginal and conditional distributions. However, learning a joint distribution remains the more challenging task and there still exists a gap in the generative performance compared to unimodal and conditional models.

We bridge this gap by proposing a new self-supervised multimodal generative model that disentangles modality-specific and shared factors. We argue that this

disentanglement is crucial for multimodal learning, because it simplifies the aggregation of representations across modalities. For conditional generation, this decomposition allows sampling from modality-specific priors without affecting the shared representation computed across multiple modalities. Further, decomposed representations have been found to be more interpretable [6, 17] and more amenable for certain downstream tasks [26].

The main contribution of this work is the development of a new multimodal generative model that learns to disentangle modality-specific and shared factors in a self-supervised manner. We term this new method disentangling multimodal variational autoencoder (DMVAE). It extends the class of multimodal variational autoencoders by modeling modality-specific in addition to shared factors and by disentangling these groups of factors using a self-supervised contrastive objective. In two representative toy experiments, we demonstrate the following advantages compared to previous multimodal generative models:

- Effective disentanglement of modality-specific and shared factors. This allows sampling from modality-specific priors without changing the joint representation computed from multiple modalities.
- Improvements in representation learning over state-of-the-art multimodal generative models. For any subset of modalities, our model aggregates shared information effectively and efficiently.
- Improvements in generative performance over previous work. In a fair comparison, we demonstrate that modeling modality-specific in addition to shared factors significantly improves the conditional generation of missing modalities. For unconditional generation, we demonstrate the effectiveness of using ex-post density estimation [8] to further improve joint generation across all methods, including trained models from previous work.

2 Related Work

Broadly, our work can be categorized as an extension of the class of multimodal generative models that handle more than two modalities (including missing ones) efficiently. Among this class, we present the first method that partitions the latent space into modality-specific and shared subspaces and disentangles these in a self-supervised fashion.

Multimodal generative models. Current approaches are mainly based on encoder-decoder architectures which learn the mapping between modalities based on reconstructions or adversarial objectives (for a comprehensive review, see [3]). Among this class, methods can be distinguished by the type of mapping they use to translate between inputs and outputs and by how they handle missing modalities. Early approaches [35, 16] try to learn all possible mappings, which in the case of missing modalities results in 2^M encoders for M modalities. A more efficient alternative is proposed by [40] who introduce the multimodal variational autoencoder (MVAE) which uses a joint posterior that is proportional to a product of experts (PoE) [14]. Their method handles missing modalities

efficiently, because it has a closed form solution for the aggregation of marginal Gaussian posteriors. However, their derivation of the joint posterior is based on the assumption that all modalities share a common set of factors—an assumption that is often violated in practice, because modalities exhibit a high degree of modality-specific variation. Our model also uses a joint latent space with a product of experts aggregation layer, and thus shares the same theoretical advantages, but it considers modality-specific factors in addition to shared factors. The limitations of the MVAE were shown empirically in [31], where it is stated that the MVAE lacks the abilities of latent factorization and joint generation. With latent factorization the authors refer to the decomposition into modality-specific and shared factors, and by joint generation they mean the semantic coherence of unconditionally generated samples across modalities. They attribute these problems to the joint posterior used by the MVAE and demonstrate empirically that using a mixture of experts, instead of a product, improves generative performance. In contrast, we argue that the product of experts is not a problem per se, but that it is an ill-defined aggregation operation in the presence of modality-specific factors. We resolve this model misspecification by modeling modality-specific factors in addition to shared factors. Compared to the mixture of experts multimodal variational autoencoder (MMVAE) [31], our model has the advantage that it can sample from a modality-specific prior without affecting the shared representation which can still be aggregated efficiently across modalities through the PoE. Especially with more than two modalities, the aggregation of representations, as it is done in our model, shows its benefits compared to the MMVAE (see Section 4.2).

Domain Adaption/Translation. The research areas of domain adaption and domain translation are in many regards closely related to multimodal generative models. Approaches that have explored many-to-many mappings between different domains have been based on adversarial methods [24, 7], shared autoencoders [36] and cycle-consistency losses [2]. Translation methods have shown remarkable progress on image-to-image style transfer and the conceptual manipulation of images, however, their focus lies on learning conditional mappings, while our method models the joint distribution directly. Further, through the PoE our method aggregates shared representations across any subset of modalities and therefore handles missing modalities efficiently.

Disentanglement. Our goal is not the unsupervised disentanglement of all generative factors, which was shown to be theoretically impossible with a factorizing prior and claimed to be impossible in general [25]. Instead, we are concerned with the disentanglement of modality-specific and shared sets of factors. In the multi-view and multimodal case, there is theoretical evidence for the identifiability of shared factors [9, 19, 27, 37]. Further, the self-supervised disentanglement of shared factors has been previously explored based on grouping information [4], temporal dependencies [23], partly labeled data [18, 38, 39], and spatial information [5]. We take a first step towards disentanglement given multimodal data with modality-specific factors and an implicit, unknown grouping.

3 Method

In this section, we introduce multimodal generative models and derive the variational approximations and information-theoretic objectives that our method optimizes. All proofs are provided in the appendix.

We consider a generative process with a partition into modality-specific and modality-invariant (i.e., shared) latent factors (Figure 1). A multimodal sample $\mathbf{x} = (x_1, \ldots, x_M)$ with data from M modalities is assumed to be generated from a set of shared factors c and a set of modality-specific factors s_m. Consequently, samples from different modalities are assumed to be conditionally independent given c. In the following, we denote the set of all modality-specific factors of a multimodal sample as $\mathbf{s} = (s_1, \ldots s_M)$.

Given a dataset $\{\mathbf{x}^{(i)}\}_{i=1}^{N}$ of multimodal samples, our goal is to learn a generative model $p_\theta(\mathbf{x} \,|c, \mathbf{s})$ with a neural network parameterized by θ. From the above assumptions on the data generating process, it follows the joint distribution

$$p(\mathbf{x}, \mathbf{s}, c) = p(c) \prod_{m=1}^{M} p(s_m) p(x_m \mid c, s_m) \tag{1}$$

which allows to consider only the observed modalities for the computation of the marginal likelihood.

The computation of the exact likelihood is intractable, therefore, we resort to amortized variational inference and instead maximize the evidence lower bound

$$\mathcal{L}_{\text{VAE}}(\mathbf{x}, c) := \sum_{m=1}^{M} \mathbb{E}_{q_\phi(s_m \mid x_m)} \Big[\log p_{\theta_m}(x_m \mid c, s_m) \Big] - D_{\text{KL}} \left(q_\phi(s_m | x_m) \,||\, p(s_m) \right)$$

which is composed of M log-likelihood terms and KL-divergences between approximate posteriors $q_\phi(s_m \mid \mathbf{x})$ and priors $p(s_m)$. Above objective describes M modality-specific VAEs, each of which takes as input an additional context vector c that encodes shared information (described in Section 3.2). We use neural networks for each encoder $q_{\phi_m}(s_m|x_m)$ as well as for each decoder $p_{\theta_m}(x_m|c, s_m)$ and denote the network parameters by the respective subscripts for decoder parameters θ and encoder parameters ϕ. Further, we follow the convention of using an isotropic Gaussian prior and Gaussian variational posteriors parameterized by the estimated means and variances that are the outputs of the encoder.

For each modality-specific VAE, it is possible to control the degree of disentanglement of arbitrary factors with a weight on the respective KL-divergence term, like in the β-VAE [13]. However, there exist theoretical limitations on the feasibility of unsupervised disentanglement of arbitrary factors [25]. In contrast, we focus on the disentanglement of modality-specific and shared factors, for which we use two additional objectives that are introduced in Subsection 3.2.

3.1 Multimodal inference network

A key aspect in the design of multimodal models should be the capability to handle missing modalities efficiently [3]. In our case, only the shared representation depends on all modalities and should ideally be able to cope with any

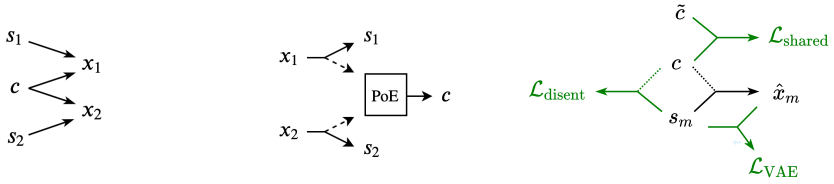


Fig. 1: Graphical model and network architecture for the special case of two modalities. *Left:* A sample x_m from modality m is assumed to be generated by modality-specific factors s_m and modality-invariant factors c. *Center:* Inference network that aggregates shared factors through a product of experts (PoE) layer. Dashed lines represent simulated missing modalities as used during training. *Right:* Decoder network (black) for modality m and loss terms (green). Dotted lines denote paths that are not being backpropagated through. Shared factors are learned by a contrastive objective which takes as input representations c and $\tilde{c}$ computed from different subsets of modalities. Modality-specific factors are inferred by regularizing out shared information from the latent space of the VAE. All loss terms are defined in Subsection 3.2.

combination of missing inputs, which would require 2^M inference networks in a naive implementation. A more efficient alternative is offered in [40], where a product of experts (PoE) [14] is used to handle missing modalities. Under the assumption of shared factors, previous work [40] has shown that the posterior $p(c \mid \mathbf{x})$ is proportional to a product of unimodal posteriors

$$p(c \mid \mathbf{x}) \propto \frac{1}{p(c)^{M-1}} \prod_{m=1}^{M} p(c \mid x_m) \tag{2}$$

which—for the special case of Gaussian posteriors—has an efficient closed-form solution (see Appendix A.3). We also assume Gaussian unimodal posteriors $q_{\psi_m}(c|x_m)$ where ψ denotes the encoder parameters, part of which can be shared with the encoder parameters ϕ_m of a unimodal VAE. The choice of Gaussian posteriors allows us to employ the PoE as an aggregation layer for shared factors. This allows the model to use M unimodal inference networks to handle all 2^M combinations of missing modalities for the inference of shared factors.

While the PoE is a well defined aggregation operation for shared factors, it is not suitable for modality-specific factors, because it averages over representations from different modalities.[1] Therefore, we partition the latent space into $M + 1$ independent subspaces, one that is specific for each modality (denoted by s_m) and one that has shared content between all modalities (denoted by c), as illustrated in Figure 1. The PoE is only used for the shared representation, so modality-specific information is not forced through the aggregation layer.

[1] This problem has also been observed in [21] where it is described as "averaging over inseparable individual beliefs".

In theory, a partitioned latent space provides the possibility to encode both modality-specific and shared information in separate subspaces; in practice, however, objective $\mathcal{L}_{\text{VAE}}$ does not specify what information (modality-specific or shared) should be encoded in which subspace. For example, the first log-likelihood term $\log p_\theta(x_1 \mid c, s_1)$ can be maximized if *all* information from input x_1 flows through the modality-specific encoder $q_\phi(s_1 \mid x_1)$ and none through the shared encoder. Thus, we posit that the model requires an additional objective for disentangling modality-specific and modality-invariant information. Next, we formalize our notion of disentanglement and introduce suitable contrastive objectives.

3.2 Disentanglement of c and $\mathbf{s}$

We take an information-theoretic perspective on disentanglement and representation learning. Consider multimodal data to be a random variable X and let $h_1(X)$ and $h_2(X)$ be two functions, each of which maps the data to a lower-dimensional encoding. Consider the objective

$$\max_{h_1, h_2 \in \mathcal{H}} I(X; h_1(X)) + I(X; h_2(X)) - I(h_1(X); h_2(X)) \tag{3}$$

where I denotes the mutual information between two random variables and $\mathcal{H}$ is the set of functions that we optimize over, for instance, the parameters of a neural network. Objective (3) is maximized by an encoding that is maximally informative about the data while being maximally independent between $h_1(X)$ and $h_2(X)$. In our case, these two functions should encode modality-specific and shared factors respectively. The proposed model learns such a representation by using suitable estimators for the individual information terms.

The objective optimized by a VAE can be viewed as a lower bound on the mutual information between data and encoding (e.g., see [1, 15]). However, on itself a VAE does not suffice to learn a disentangled encoding, because of theoretical limitations on disentanglement in an unsupervised setting [25]. So in addition, we equip the VAE with two contrastive objectives: one that learns an encoding of information shared between modalities, maximizing a lower bound on $I(\mathbf{x}; c)$, and one that infers modality-specific factors by regularizing out shared information from the latent space of a modality-specific VAE. The overall objective that is being maximized is defined as

$$\mathcal{L} = \mathcal{L}_{\text{VAE}} + \gamma \mathcal{L}_{\text{shared}} - \delta \mathcal{L}_{\text{disent}} \tag{4}$$

where $\mathcal{L}_{\text{VAE}}$ is the ELBO optimized by the VAEs, $\mathcal{L}_{\text{shared}}$ learns an encoding of shared factors, $\mathcal{L}_{\text{disent}}$ disentangles shared and modality-specific information, and the hyperparameters γ and δ can be used to control these terms respectively. The proposed objective estimates shared factors directly, while modality-specific factors are inferred indirectly by regularizing out shared information from the encoding of a modality-specific VAE. Further, as in the β-VAE [13], the reconstruction loss and KL-divergence contained in $\mathcal{L}_{\text{VAE}}$ can be traded off to

control the quality of reconstructions against the quality of generated samples. Figure 1 shows a schematic of the network including all loss terms that are being optimized. In the following, we define the contrastive objectives used for the approximation of the respective mutual information terms.

To learn shared factors, we use a contrastive objective [32, 10] that maximizes a lower bound on the mutual information $I(\mathbf{x}; c)$ (see Appendix A for the derivation). We estimate the mutual information with the sample-based InfoNCE estimator [29] adapted to a multimodal setting. The objective is defined as

$$\mathcal{L}_{\text{shared}} := -\mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\log\frac{e^{f(\mathbf{x}^{(i)},\tilde{\mathbf{x}}^{(i)})}}{\frac{1}{K}\sum_{j=1}^{K}e^{f(\mathbf{x}^{(i)},\tilde{\mathbf{x}}^{(j)})}}\right] \tag{5}$$

where the expectation goes over K independent samples $\{\mathbf{x}^{(i)}, \tilde{\mathbf{x}}^{(i)}\}_{i=1}^{K}$ from $p(\mathbf{x}, \tilde{\mathbf{x}})$ where $\tilde{\mathbf{x}}$ is a subset of modalities $\tilde{\mathbf{x}} \subset \mathbf{x}$ and f is a critic that maps to a real-valued score. In particular, we use an inner product critic $f_\phi(\mathbf{x}, \tilde{\mathbf{x}}) = \langle c, \tilde{c}\rangle$ where c and $\tilde{c}$ are the representations computed from a full multimodal sample and a subset of modalities respectively. Intuitively, the objective contrasts between a positive pair coming from the same multimodal sample and $K-1$ negative pairs from randomly paired samples [e.g., 11]. By using a large number of negative samples, the bound becomes tighter [29], therefore we use a relatively large batch size of $K = 1024$ such that for every positive, we have 1023 negative samples by permuting the batch. In Appendix A we prove that the contrastive objective is a lower bound on $I(\mathbf{x}; c)$ and we further discuss the approximation as well as our choice of critic.

To regularize out shared information from the encoding of a modality-specific VAE, we use a discriminator that minimizes the total correlation $TC(c, s_m)$, a measure of statistical dependence between a group of variables. In the case of two variables, the total correlation is equivalent to the mutual information. We approximate the total correlation using the density-ratio trick [28, 34] and refer to the approximation by $\mathcal{L}_{\text{disent}}$ (see Appendix A). This procedure is very similar to the one used by [20] with the important difference that we do not estimate the total correlation between all elements in a single latent representation, but between partitions c, s_m of the latent space, of which c is shared between modalities. In theory, one can use a single discriminator to minimize $TC(c, \mathbf{s})$ jointly, however, we found that in practice one has more control over the disentanglement by using individual terms $\mathcal{L}_{disent} = \delta_m \sum_m \mathcal{L}_{disent}(c, s_m)$ weighted by separate disentanglement coefficients δ_m, instead of a global δ.

4 Experiments

In this section, we compare our method to previous multimodal generative models both qualitatively and quantitatively. In the first experiment, we use a bimodal dataset that has been used in previous studies and compare our method to the MVAE [40] and MMVAE [31], the current state-of-the-art multimodal generative models. In the second experiment, we go beyond two modalities and

construct a dataset with 5 simplified modalities that allows us to analyze the aggregation of representations across multiple modalities, which, to the best of our knowledge, has not been done previously.

For the quantitative evaluation, we employ metrics that were used in previous studies. Mainly, we focus on generative coherence [31], which takes a classifier (pretrained on the original data) to classify generated samples and computes the accuracy of predicted labels compared to a ground truth. For unconditional samples, coherence measures how often the generated samples match across all modalities. To measure the quality of generated images, we compute Frchet Inception Distances (FIDs) [12]. It is important to note that a generative model can have perfect coherence yet very bad sample quality (e.g., blurry images of the correct class, but without any diversity). Analogously, a model can achieve very good FID without producing coherent samples. Therefore, we also propose to compute class-specific conditional FIDs for which the set of input images is restricted to a specific class and the set of conditionally generated images is compared to images of that class only. Hence, class-specific conditional FID provides a measure of both coherence and sample quality. Finally, we evaluate the quality of the learned representations by training a linear classifier on the outputs of the encoders.

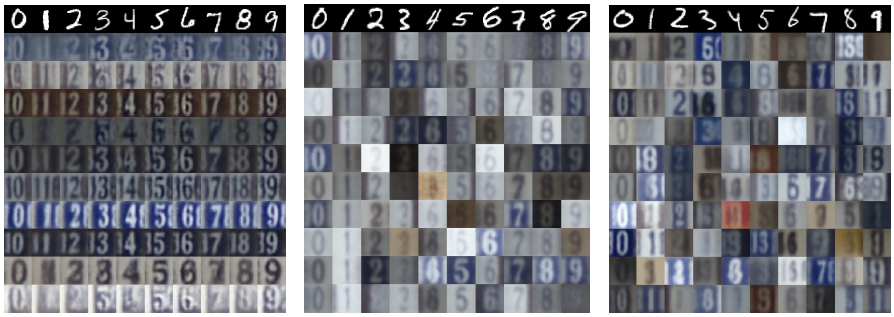
4.1 MNIST-SVHN

A popular dataset for the evaluation of multimodal generative models is the MNIST-SVHN dataset [38, 31], which consists of digit images from two different domains, hand-written digits from MNIST [22] and street-view house numbers from SVHN [30]. The images are paired by corresponding digit labels, and similar to [31] we use 20 random pairings for each sample in either dataset. The pairing is done for the training and test sets separately and results in a training set of 1,121,360 and test set of 200,000 image pairs. The dataset is convenient for the evaluation of multimodal generative models, because it offers a clear separation between shared semantics (digit labels) and perceptual variations across modalities. This distinctive separation is required for the quantitative evaluation via generative coherence and class-specific conditional FID.

For a fair comparison to previous work, we employ the same architectures, likelihood distributions, and training regimes across all models. The setup is adopted from the official implementation of the MMVAE.[2] For our model we use a 20 dimensional latent space of which 10 dimensions are shared between modalities and 10 dimensions are modality-specific.[3] This does not increase the total number of parameters compared to the MMVAE or MVAE where a 20 dimensional latent space is used respectively. All implementation details are listed in Appendix C.

[2] https://github.com/iffsid/mmvae.

[3] The size of latent dimensions for modality-specific and shared representations is a hyperparameter of our model. Empirically, we found the effect of changing the dimensionality to be minor, as long as neither latent space is too small.



(a) DMVAE (ours) (b) MMVAE (c) MVAE

Fig. 2: Comparison of conditionally generated SVHN samples given the respective MNIST digit in the first row. Across a column, we sample from the modality-specific prior (our model) or from the posterior (other models). Only our model keeps consistent styles across rows, as it disentangles modality-specific and shared factors (without supervision).

Qualitative results. Figure 2 illustrates the conditional generation of SVHN given MNIST. Only our method is capable of keeping consistent styles across rows, because our model allows to draw samples from the modality-specific prior without changing the shared representation computed from the input. For both MVAE and MMVAE, we sample from the posterior to generate diverse images along one column.[4] One can already observe that our model and the MMVAE are both capable of generating images with coherent digit labels, while the MVAE struggles to produce matching digits, as already observed in [31]. The results are similar for the conditional generation of MNIST given SVHN (see Appendix B), demonstrating that our method is effective in disentangling modality-specific and shared factors in a self-supervised manner.

Quantitative results. Since the setup of this experiment is equivalent to the one used by [31] to evaluate the MMVAE, we report the quantitative results from their paper. However, we decided to implement the MVAE ourselves, because we found that the results reported in [31] were too pessimistic.

Table 1 presents linear latent classification accuracies as well as conditional and unconditional coherence results. Across all metrics, our model achieves significant improvements over previous methods. Most strikingly, joint coherence improves from 42.1% to 85.9% as a result of ex-post density estimation. As previously noted, it can be misleading to look only at latent classification and coherence, because these metrics do not capture the diversity of generated samples. Therefore, in Table 2 we also report FIDs for all models. In terms of FIDs,

[4] We further observed that without sampling from the posterior (i.e., reparameterization) both the MVAE and MMVAE tend to generate samples with very little diversity, even if diverse input images are used.

Table 1: Results on MNIST/SVHN, where x_1 corresponds to MNIST and x_2 to SVHN. Numbers denote median values over 5 runs (standard deviations in parentheses). For MMVAE, numbers are based on the original work and standard deviations were computed with the publicly available code. For latent classification, we use linear classifiers and for the DMVAE only the shared representation is used (concatenation further improves the results).

Method	**Latent accuracy (in %)**			**Coherence (in %)**		
	x_1	x_2	Aggregated	Joint	$x_1 \rightarrow x_2$	$x_2 \rightarrow x_1$
MVAE	79.8 (± 3.8)	65.1 (± 4.6)	80.2 (± 3.6)	38.0 (± 1.8)	31.8 (± 1.4)	57.1 (± 3.4)
MMVAE	91.3 (± 0.4)	68.0 (± 0.6)	N/A	42.1 (± 1.9)	86.4 (± 0.5)	69.1 (± 2.5)
DMVAE	**95.0** (± 0.6)	**79.9** (± 1.4)	**92.9** (± 1.8)	**85.9** (± 1.0)	**91.6** (± 0.8)	**76.4** (± 0.4)

Table 2: Comparison of generative quality on MNIST/SVHN, where x_1 corresponds to MNIST and x_2 to SVHN. Numbers represent median FIDs (lower is better) computed across 5 runs with standard deviations in parentheses. For the MMVAE, we computed FIDs based on the publicly available code.

Method	**Unconditional FID**		**Conditional FID**		**Class-Conditional FID**	
	x_1	x_2	$x_1 \rightarrow x_2$	$x_2 \rightarrow x_1$	$x_1 \rightarrow x_2$	$x_2 \rightarrow x_1$
MVAE	21.2 (± 1.1)	68.2 (± 1.9)	**65.0** (± 2.2)	**19.3** (± 0.4)	**83.8** (± 1.8)	53.6 (± 1.9)
MMVAE	36.6 (± 3.1)	98.9 (± 1.5)	97.0 (± 0.6)	28.6 (± 1.1)	125.3 (± 0.8)	52.6 (± 4.8)
DMVAE	**15.7** (± 0.7)	**57.3** (± 3.6)	67.6 (± 4.0)	**18.7** (± 0.9)	91.9 (± 4.4)	**23.3** (± 1.0)

our model shows the best overall performance, with an exception in the conditional generation of SVHN given MNIST, for which the MVAE has slightly lower FIDs. However, looking at the results as a whole, DMVAE demonstrates a notable improvement compared to state-of-the-art multimodal generative models. Ablations across individual loss terms are provided in Appendix B.

Ex-post density estimation [8], which we employ for sampling from the shared space of the DMVAE, proves to be very effective for improving certain metrics (Table 3). In particular, it can be used as an additional step after training, to improve the joint coherence and, partially, unconditional FIDs of already trained models. Note that ex-post density estimation does not influence any other metrics reported in Tables 1 and 2 (i.e., latent classification, conditional coherence, and conditional FID).

4.2 Paired MNIST

To investigate how well the aggregation of shared representations works for more than two modalities, we create a modified version of the MNIST dataset, which consists of M-tuples of images that depict the same digit. We view each image in the tuple $(x_1, \ldots, x_M)$ as coming from a different modality $x_m \sim X_m$, even

Table 3: Comparison of sampling from the prior vs. using ex-post density estimation with a Gaussian mixture model (GMM) with 100 components and full covariance matrix. After training, the GMM is fitted on the embeddings computed from the training data. For FIDs, the first number refers to MNIST, the second to SVHN, respectively. Overall, ex-post density estimation improves most metrics for both MVAE and MMVAE.

	MVAE		**MMVAE**		**DMVAE**	
Sampling	FIDs	Coherence	FIDs	Coherence	FIDs	Coherence
Prior	21.2 / 68.2	38.0	36.6 / 98.9	42.1	N/A	N/A
GMM	**13.4** / 73.7	68.5	28.7 / 119.7	80.3	15.7 / **57.3**	**85.9**

though each instance is drawn from MNIST. Further we perturb each image with a high degree of Gaussian noise, which makes it difficult to infer digit labels from a single image (for an example, see Appendix B), and train the models as denoising variational autoencoders. We use comparable architectures, likelihoods, and training regimes across all methods. All implementation details are provided in Appendix C.

The dataset is generated by repeatedly pairing M images with the same label. We vary $M = 2, ...5$ to investigate how the methods perform with an increasing number of modalities. This pairing is done separately for training and test data and results in 60,000 and 10,000 image M-tuples for the training and test sets respectively. The resulting dataset offers a simple benchmark that requires no modality-specific weights for the likelihood terms, has a clear characterization of shared and modality-specific factors, and allows visual inspection of the results.[5]

The goal of this experiment is to test whether models are able to integrate shared information across multiple modalities and if the aggregated representation improves with more modalities. To the best of our knowledge, experiments evaluating the aggregation with more than two modalities have not been performed before. Unlike the previous experiment, paired MNIST allows measuring how well models generate a missing modality given two or more inputs. To quantify this, we measure the average coherence over leave-one-out mappings $\{x_i\}_{i \neq j} \rightarrow x_j$. Further, we compute the average class-specific conditional FID over leave-one-out mappings, which combines both coherence and generative quality in a single metric.

Figure 3 presents the results for an increasing number of input modalities. The left subplot shows that for the MVAE and DMVAE leave-one-out coherence consistently improves with additional modalities, supporting our hypothesis that the PoE is effective in aggregating shared information. Notably, the MMVAE fails to take advantage of more than two modalities, as it does not have a shared representation that aggregates information. The right subplot shows that the DMVAE outperforms the other methods in class-specific conditional FIDs, demonstrating

[5] Note that the weights of likelihood terms have been observed to be important hyperparameters in both [40] and [31].

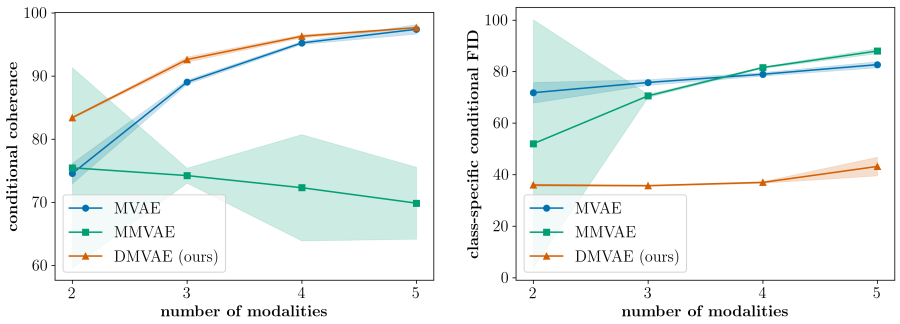


Fig. 3: Results on paired MNIST with varying number of "modalities". Markers denote median values, error-bars standard deviations, computed across 5 runs. *Left:* Leave-one-out conditional coherence (higher is better). *Right:* Class-specific conditional FIDs (lower is better).

that it can achieve both high sample quality and strong coherence. We provide further metrics and ablations for this experiment in Appendix B.

5 Conclusion

We have introduced DMVAE, a novel multimodal generative model that learns a joint distribution over multiple modalities and disentangles modality-specific and shared factors completely self-supervised. The disentanglement allows sampling from modality-specific priors and thus facilitates the aggregation of shared information across modalities. We have demonstrated significant improvements in representation learning and generative performance compared to previous methods. Further, we have found that ex-post density estimation, that was used to sample from the shared latent space of the DMVAE, improves certain metrics dramatically when applied to trained models from existing work. This suggests that the latent space learned by multimodal generative models is more expressive than previously expected, which offers exciting opportunities for future work. Moreover, the DMVAE is currently limited to disentangling modality-specific and shared factors and one could extend it to more complex settings, such as graphs of latent factors.

Acknowledgements

Thanks to Mario Wieser for discussions on learning invariant subspaces, to Yuge Shi for providing code, and to Francesco Locatello for sharing his views on disentanglement in a multimodal setting. ID is supported by the SNSF grant #200021_188466.

Bibliography

[1] Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. In: International Conference on Learning Representations (2017)

[2] Almahairi, A., Rajeswar, S., Sordoni, A., Bachman, P., Courville, A.C.: Augmented CycleGAN: Learning Many-to-Many Mappings from Unpaired Data. In: International Conference on Machine Learning (2018)

[3] Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal Machine Learning: A survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(2), 423–443 (2019)

[4] Bouchacourt, D., Tomioka, R., Nowozin, S.: Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In: AAAI Conference on Artificial Intelligence (2018)

[5] Chartsias, A., Joyce, T., Papanastasiou, G., Semple, S., Williams, M., Newby, D.E., Dharmakumar, R., Tsaftaris, S.A.: Disentangled representation learning in cardiac image analysis. Medical Image Analysis **58**, 101535 (2019)

[6] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In: Advances in Neural Information Processing Systems (2016)

[7] Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In: Conference on Computer Vision and Pattern Recognition (2018)

[8] Ghosh, P., Sajjadi, M.S.M., Vergari, A., Black, M., Scholkopf, B.: From variational to deterministic autoencoders. In: International Conference on Learning Representations (2020)

[9] Gresele, L., Rubenstein, P.K., Mehrjou, A., Locatello, F., Schölkopf, B.: The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ICA. In: Conference on Uncertainty in Artificial Intelligence (2019)

[10] Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: International Conference on Artificial Intelligence and Statistics (2010)

[11] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.B.: Momentum contrast for unsupervised visual representation learning. In: Conference on Computer Vision and Pattern Recognition (2020)

[12] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems (2017)

[13] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In: International Conference on Learning Representations (2017)

[14] Hinton, G.E.: Training products of experts by minimizing contrastive divergence. Neural Computation **14**(8), 1771–1800 (2002)
[15] Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. In: International Conference on Learning Representations (2019)
[16] Hsu, W.N., Glass, J.: Disentangling by Partitioning: A Representation Learning Framework for Multimodal Sensory Data. arXiv preprint arXiv:1805.11264 (2018)
[17] Hsu, W.N., Zhang, Y., Glass, J.: Unsupervised learning of disentangled and interpretable representations from sequential data. In: Advances in Neural Information Processing Systems (2017)
[18] Ilse, M., Tomczak, J.M., Louizos, C., Welling, M.: DIVA: Domain Invariant Variational Autoencoders. arXiv preprint arXiv:1905.10427 (2019)
[19] Khemakhem, I., Kingma, D.P., Monti, R.P., Hyvärinen, A.: Variational autoencoders and nonlinear ICA: A unifying framework. In: International Conference on Artificial Intelligence and Statistics (2020)
[20] Kim, H., Mnih, A.: Disentangling by factorising. In: International Conference on Machine Learning (2018)
[21] Kurle, R., Guennemann, S., van der Smagt, P.: Multi-Source Neural Variational Inference. In: AAAI Conference on Artificial Intelligence (2019)
[22] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
[23] Li, Y., Mandt, S.: Disentangled sequential autoencoder. In: International Conference on Machine Learning (2018)
[24] Liu, A.H., Liu, Y.C., Yeh, Y.Y., Wang, Y.C.F.: A unified feature disentangler for multi-domain image translation and manipulation. In: Advances in Neural Information Processing Systems (2018)
[25] Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., Bachem, O.: Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In: International Conference on Machine Learning (2019)
[26] Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B., Bachem, O.: On the fairness of disentangled representations. In: Advances in Neural Information Processing Systems (2019)
[27] Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., Tschannen, M.: Weakly-supervised disentanglement without compromises. In: International Conference on Machine Learning (2020)
[28] Nguyen, X., Wainwright, M.J., Jordan, M.I.: Estimating divergence functionals and the likelihood ratio by convex risk minimization. IEEE Transactions on Information Theory **56**(11), 5847–5861 (2010)
[29] Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
[30] Sermanet, P., Chintala, S., LeCun, Y.: Convolutional neural networks applied to house numbers digit classification. In: International Conference on Pattern Recognition. pp. 3288–3291. IEEE (2012)

[31] Shi, Y., Siddharth, N., Paige, B., Torr, P.: Variational mixture-of-experts autoencoders for multi-modal deep generative models. In: Advances in Neural Information Processing Systems (2019)
[32] Smith, N.A., Eisner, J.: Contrastive estimation: Training log-linear models on unlabeled data. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 354–362 (2005)
[33] Stein, B.E., Stanford, T.R., Rowland, B.A.: The neural basis of multisensory integration in the midbrain: its organization and maturation. Hearing research **258**(1-2), 4–15 (2009)
[34] Sugiyama, M., Suzuki, T., Kanamori, T.: Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation. Annals of the Institute of Statistical Mathematics **64**(5), 1009–1044 (2012)
[35] Suzuki, M., Nakayama, K., Matsuo, Y.: Joint multimodal learning with deep generative models. arXiv preprint arXiv:1611.01891 (2016)
[36] Tian, Y., Engel, J.: Latent translation: Crossing modalities by bridging generative models. arXiv preprint arXiv:1902.08261 (2019)
[37] Träuble, F., Creager, E., Kilbertus, N., Goyal, A., Locatello, F., Schölkopf, B., Bauer, S.: Is independence all you need? On the generalization of representations learned from correlated data. arXiv preprint arXiv:2006.07886 (2020)
[38] Tsai, Y.H.H., Liang, P.P., Zadeh, A., Morency, L.P., Salakhutdinov, R.: Learning Factorized Multimodal Representations. In: International Conference on Learning Representations (2019)
[39] Wieser, M., Parbhoo, S., Wieczorek, A., Roth, V.: Inverse learning of symmetry transformations. In: Advances in Neural Information Processing Systems (2020)
[40] Wu, M., Goodman, N.: Multimodal Generative Models for Scalable Weakly-Supervised Learning. In: Advances in Neural Information Processing Systems (2018)
[41] Yildirim, I.: From perception to conception: learning multisensory representations. Ph.D. thesis, University of Rochester (2014)
[42] Yildirim, I., Jacobs, R.A.: Transfer of object category knowledge across visual and haptic modalities: Experimental and computational studies. Cognition **126**(2), 135–148 (2013)