Neural Networks 169 (2024) 417–430

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Discriminative multimodal learning via conditional priors in generative models

Rogelio A. Mancisidor [a,*], Michael Kampffmeyer [b,c], Kjersti Aas [c], Robert Jenssen [b,c]

[a] *Department of Data Science and Analytics, BI Norwegian Business School, Nydalsveien 37, 0484 Oslo, Norway*
[b] *Department of Physics and Technology, Faculty of Science and Technology, UiT The Arctic University of Norway, Hansine Hansens veg 18, 9037 Tromsø, Norway*
[c] *Norwegian Computing Center, P.O. Box 114 Blindern Oslo, Norway*

ARTICLE INFO

Keywords:
Multimodal learning
Generative models
Representation learning
Variational autoencoder

ABSTRACT

Deep generative models with latent variables have been used lately to learn joint representations and generative processes from multi-modal data, which depict an object from different viewpoints. These two learning mechanisms can, however, conflict with each other and representations can fail to embed information on the data modalities. This research studies the realistic scenario in which all modalities and class labels are available for model training, e.g. images or handwriting, but where some modalities and labels required for downstream tasks are missing, e.g. text or annotations. We show, in this scenario, that the variational lower bound limits mutual information between joint representations and missing modalities. We, to counteract these problems, introduce a novel conditional multi-modal discriminative model that uses an informative prior distribution and optimizes a likelihood-free objective function that maximizes mutual information between joint representations and missing modalities. Extensive experimentation demonstrates the benefits of our proposed model, empirical results show that our model achieves state-of-the-art results in representative problems such as downstream classification, acoustic inversion, and image and annotation generation.

1. Introduction

Measurement modalities $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ depict different viewpoints of an object (Fig. 1) and are used in multi-modal learning to learn $\boldsymbol{z}$, a joint representation which captures information from all modalities and that can be used for clustering, active and transfer learning, or, where class labels y are available, downstream classification. According to Shi et al. (2019) multi-modal learning models should satisfy four criteria: latent factorization, coherent joint and cross generation, and synergy.

Deep neural networks (DNNs) and deep generative models (DGMs) with latent representations have been used in multi-modal learning (Andrew et al., 2013; Du et al., 2018, 2019; Shi et al., 2019; Sutter et al., 2020, 2021; Suzuki et al., 2016; Wang et al., 2015a, 2017; Wu & Goodman, 2018). DGMs learn joint latent representations using variational approximations of the posterior distribution, and learn generative models for data modalities by optimizing a variational lower bound on the log-likelihood of the data. These two mechanisms can, however, conflict with each other. Generative models may focus on generating modalities without using the joint latent representation. Therefore, the posterior distribution for the joint representation fails to embed information on the modalities, collapsing into a non-informative prior distribution. This is called posterior collapse (Dieng et al., 2019; Lucas et al., 2019), and harms the performance of downstream tasks based on joint representations, e.g. classification or modality generation. Posterior collapse has been studied in uni-modal frameworks, but less so in multi-modal domains.

There are different applications in which the data come from different sources referred to here as modalities, e.g., tuples of images and annotations or acoustic and articulatory measurements. However, not all observations necessarily come in tuples, because annotating images or measuring articulatory movements can e.g. be costly (Sutter et al., 2020, 2021) and take time to be generated. Hence, we are interested in modeling conditional distributions that are able to learn multi-modal latent representations, which can then be used to generate data from missing modalities. Learning such latent representations is important because we can capture relationships between modalities that are valuable for generative and discriminative downstream tasks. Towards this goal, we introduce a conditional multi-modal discriminative (CMMD) model that works in the aforementioned scenarios,

[*] Corresponding author.
E-mail addresses: rogelio.a.mancisidor@bi.no (R.A. Mancisidor), michael.c.kampffmeyer@uit.no (M. Kampffmeyer), kjersti@nr.no (K. Aas), robert.jenssen@uit.no (R. Jenssen).

https://doi.org/10.1016/j.neunet.2023.10.048
Received 16 August 2022; Received in revised form 15 September 2023; Accepted 30 October 2023
Available online 2 November 2023

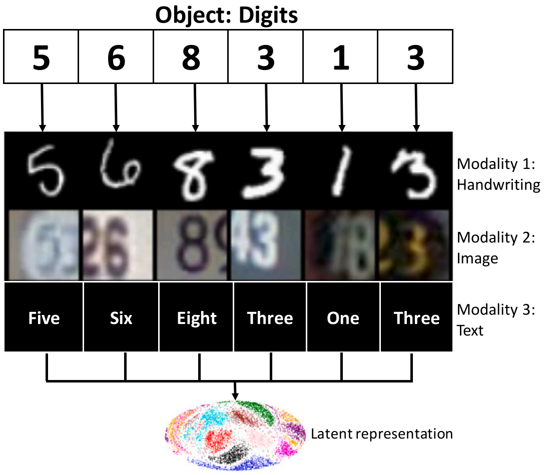


Fig. 1. A graphical scheme of multi-modal learning in which the data modalities depict different viewpoints of an object and a latent representation embeds information from all modalities.

where all modalities and class labels are available for model training, but where some modalities and class labels required for downstream tasks are missing. Missing modalities, in the context of this research, refer to modalities that are costly to obtain for downstream tasks or modalities that we are interested in generating conditional on all other modalities, i.e. cross-modal generation or retrieval generation (Guo et al., 2019), and therefore refer to them as missing. We show, in this scenario, that the variational lower bound limits mutual information (MI) between multi-modal representations and missing modalities. To counteract this limitation, we introduce a novel likelihood-free objective function that optimizes MI and also introduce a prior distribution for joint representations that is conditioned on the available modalities.

We show, through extensive experimentation, that by optimizing the MI between multi-modal representations and missing modalities, the latent representation learned by our proposed model does not show posterior collapse. We also show that its joint representations embed information from multiple data modalities, which is useful for downstream tasks. We have benchmarked different multi-modal learning models across different representative domains, e.g. image-to-image, acoustic-to-articulatory, image-to-annotation, and text-to-image. The empirical results from this show that CMMD achieves state-of-the-art results in downstream classification and in the generation of missing modalities at test time.

Our main contributions are:

- A new objective function that counteracts the restriction on MI between joint representations and the missing modalities
- A generative process that generates data from missing modalities at test time using a conditional prior[1]
- Insights into the effect of posterior collapse in downstream classification and in the generative process in multi-modal learning.

2. Multimodal learning

We use a common notation. Data modalities are represented by $\boldsymbol{x}$ and distinguished by a subscript. Joint latent representations are denoted by $\boldsymbol{z}$. In the following we provide an overview of the relevant multi-modal learning models to this work. See Guo et al. (2019) for a comprehensive review.

Deep neural networks. Deep canonical correlation analysis (Andrew et al., 2013) couples deep neural networks with canonical correlation analysis (Hotelling, 1936) to train neural networks $f(\cdot)$ and $g(\cdot)$ such that they can maximize the correlation $\rho(f(\boldsymbol{x}_1), g(\boldsymbol{x}_2))$ between views (modalities) $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. DCCA (Deep Canonical Correlation Analysis) not only handles non-linearities, but also captures high-level data abstractions in each of the multiple hidden layers. Its objective function is, however, a function of the entire data set and therefore does not scale to large data sets. To overcome this limitation, Wang et al. (2015a) developed the deep canonically correlated autoencoder (DCCAE), which is optimized using stochastic gradient descent. DCCAE also introduced reconstruction neural networks for the data modalities, which minimized their reconstruction error. This is in addition to maximizing the canonical correlation between the learned representations. Both DCCA and DCCAE use fully-connected neural networks to learn representations and, in the case of DCCAE, to reconstruct the data modalities using a bottleneck autoencoder-like architecture. These two models employ stochastic gradient descent as a means of optimization.

Variational inference. A problem with DCCAE is that the canonical correlation term in its objective function dominates the optimization procedure (Wang et al., 2015a). The reconstruction of the modalities is therefore poor. Wang et al. (2017) therefore developed a variational CCA (VCCA) model to overcome this problem. VCCA uses variational inference and deep generative models to generate latent representations of input modalities. As VCCA is a probabilistic model, the authors use fully-connected neural networks to parameterized the mean and variance parameters in the probability functions defining the inference and generative model in VCCA, and use stochastic gradient descent to maximize the evidence lower bound of the model.

Du et al. (2019) proposed DMDGM, a supervised extension of VCCA that combines multi-modal learning and classification in a unified framework. The classification in DMDGM uses available views and not joint representations. DMDGM is, however, not the only model that addresses classification in a unified objective function. Du et al. (2018) developed a semi-supervised deep generative model for missing modalities, the latent variable being shared across modalities. They also modeled the inference process as a Gaussian mixture model (GMM). Modeling the inference process as a GMM, however, harms the tightness of the lower bound, as the entropy of a GMM is intractable.

The model presented by Vedantam et al. (2017) focuses on cross-modality generation, using the product of experts (PoE) in the factorization of posterior inference distributions. Wu and Goodman (2018) similarly introduced MVAE, which assumes that the posterior distribution is proportional to the product of individual conditional posteriors $p(\boldsymbol{z}|\boldsymbol{x}_1) \cdots p(\boldsymbol{z}|\boldsymbol{x}_n)$ normalized by the prior distribution $p(\boldsymbol{z})$. The joint posterior distribution is therefore also a PoE. Shi et al. (2019), through applying a similar approach, used a mixture of experts (MoE) to develop MMVAE, the generative process of the model allowing conditioning modalities and generation modalities to be interchangeable. MoE and PoE provide elegant ways of cross-generation. The linear combination of marginal distributions, however, learn joint representations that might not be useful for downstream classification (see Section 4.2). Sutter et al. (2021) show that the MVAE models the joint posterior distribution as a geometric mean, while MMVAE models it as an arithmetic mean. Further, they generalize these two approaches in a Mixture-of-Products-of-Experts (MoPoE) VAE, which approximates the joint posterior of all subsets of modalities. MVAE, MMVAE and MoPoE use text as a data modality, which requires encoder and decoder architectures based on convolutional (MMVAE and MoPoE) or recurrent (MVAE) neural networks.

It is noteworthy that MVAE, MMVAE, and MoPoE approximate the joint posterior distribution, conditioned on all modalities, as a

[1] Conditional priors in variational autoencoders were introduced in Sohn et al. (2015). However, the focus was on the reconstruction of output data based on (always available) input data. On the other hand, our model uses conditional priors to generate latent representations in scenarios with missing data modalities. Hence, the prior distribution is modulated by the available modalities at test time and generates more informative representations than isotropic Gaussian priors.

function of unimodal posterior distributions. Such a modeling approach can deal with any combination of missing modalities simultaneously and, therefore, cross-modal generation can be done in any direction efficiently. However, none of these models are discriminative by nature and, as a consequence, can only deal with discriminative tasks in a two-steps fashion. CMMD is also able to model any combination of missing modalities, but one at a time. On the other hand, generative and discriminative models are trained end-to-end in the CMMD model.

The most recent multi-modal learning research has focused on different ways of learning flexible joint representations that are useful in cross-modality generation. For example, Theodoridis et al. (2020) describe the learning of joint representation by introducing a cross-modal alignment of the latent spaces by minimizing Wasserstein distances; Nedelkoski et al. (2020) couple normalizing flows and MVAE to learn more expressive representations; Liu et al. (2021) propose a variational information bottleneck lower bound to force the encoder to discard irrelevant information, keeping only relevant information to generate one modality. Chen and Zhu (2022) use generative adversarial networks to simultaneously align the different encoder distributions with the joint decoder distribution. None of these new methods, however, have been developed for downstream classification with missing modalities. Javaloy et al. (2022) focus on learning encoders and decoders that are impartial to the unimodal posterior distributions that generate latent representations. To achieve such impartial optimization (IO), the authors propose a novel optimization technique that modifies the gradients of each modality and, as a result, does not neglect the optimization of any specific modality.

Abrol et al. (2020) introduced a uni-modal method that uses, as in our proposed model, conditional priors to generate a discrete mixture of representations in the prior space. These are considered to be local latent variables. Continuous variables in the posterior distribution are considered to be global. Local and global variables are, for supervised data, aligned using maximum mean discrepancy (Gretton et al., 2007), which optimizes the mutual information of global latent variables and input data. However, our proposed CMMD model focuses, instead, on multi-modal data and uses conditional priors to generate representations when some modalities are missing. Further, its objective function arises from the restriction imposed by the Kullback–Leibler divergence in the evidence lower bound on mutual information.

3. Methods

3.1. Evidence lower bound

We have access to labeled multi-modal data $(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)$ during training. $\boldsymbol{x}_{\mathcal{O}} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ are n modalities that are always available and $\boldsymbol{x}_{\mathcal{M}} = (\boldsymbol{x}_{n+1}, \dots, \boldsymbol{x}_{n+m})$ are m modalities that are missing at test time.[2] Only $\boldsymbol{x}_{\mathcal{O}}$ is therefore available for downstream tasks, the label y and $\boldsymbol{x}_{\mathcal{M}}$ both missing. Our proposed model at test time generates latent representations, using a prior distribution $p(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})$ conditioned on the observed modalities. Latent representations $\boldsymbol{z} \sim p(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})$ are furthermore used in both the generative process $p(\boldsymbol{x}_{\mathcal{M}}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{z})$ and in the classifier model $p(y|\boldsymbol{z})$. This encourages the model to learn useful representations for classification, and to generate data from missing modalities at test time.

The joint distribution in our proposed model is, in this scenario, given by $p(\boldsymbol{x}_{\mathcal{M}}, y, \boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}) = p(\boldsymbol{x}_{\mathcal{M}}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{z})p(y|\boldsymbol{z})p(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})$, where $p(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})$ is a prior distribution conditioned on the always available modalities, $p(\boldsymbol{x}_{\mathcal{M}}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{z})$ is the generative process for the missing modalities at test time, and $p(y|\boldsymbol{z})$ is the density function for class labels. Note that the posterior distribution $p(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)$, the joint latent representation that we want to learn, requires a marginal distribution that is not available in closed form. We therefore approximate the true posterior distribution $p(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)$ using the parametric model, or encoder distribution, $q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)$.

The evidence lower bound (ELBO) $\mathcal{L}(\boldsymbol{x}_{\mathcal{M}}, \boldsymbol{x}_{\mathcal{O}}, y)$ of our proposed model is therefore

$$\begin{aligned}\log p(\boldsymbol{x}_{\mathcal{M}}, y|\boldsymbol{x}_{\mathcal{O}}) \geq& \mathbb{E}_{q(z|x_{\mathcal{O}},x_{\mathcal{M}},y)}\left[\log \frac{p(\boldsymbol{x}_{\mathcal{M}}, y, \boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})}{q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)}\right] \\ \equiv& \mathcal{L}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y),\end{aligned} \tag{1}$$

the inequality being a result of the concavity of log and Jensen's inequality. See Appendix A for details.

3.2. Maximizing mutual information

We can, in principle, optimize Eq. (1) using the stochastic variational gradient Bayes (SVGB) algorithm (Kingma & Welling, 2013). Eq. (1) does, however, include an average Kullback–Leibler divergence that is an upper bound on the conditional mutual information between $\boldsymbol{z}$ and $\boldsymbol{x}_{\mathcal{M}}$ (see Appendix B), i.e.

$$\mathbb{E}_{p(x_{\mathcal{M}},x_{\mathcal{O}},y)}[KL[q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)||p(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})] \geq I(\boldsymbol{x}_{\mathcal{M}}, \boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}). \tag{2}$$

Conditional mutual information and posterior collapse: We therefore introduced a conditional mutual information term $(1-\omega)I(\boldsymbol{x}_{\mathcal{M}}, \boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})$ in Eq. (1) to counteract the upper bound imposed by the Kullback–Leibler divergence, $\omega \in [0, 1]$ weighting the optimization on the mutual information term. Note that the consequence of the upper bound in Eq. (2) may result in latent representations that do not encode information about $\boldsymbol{x}_{\mathcal{M}}$, which is equivalent to generating $\boldsymbol{x}_{\mathcal{M}}$ based on the prior $p(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})$. This problem is called *posterior collapse* in the uni-modal literature (Dieng et al., 2019; Lucas et al., 2019), and it occurs when the variational posterior distribution matches the prior. It should be noted that the main motivation to optimize the mutual information term is to bypass the constraint imposed by the Kullback–Leibler divergence and, as a consequence of this choice, the latent representation learned by our proposed model does not show posterior collapse as shown in Section 4.2 and Fig. 6.

Therefore, the following likelihood-free objective function for a single data point is therefore obtained[3]

$$\begin{aligned}\mathcal{J}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y) = \mathbb{E}_{q(z|x_{\mathcal{O}},x_{\mathcal{M}},y)}&[\log p(\boldsymbol{x}_{\mathcal{M}}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{z}) + \log p(y|\boldsymbol{z}) + \log p(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}) \\ &- \log q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)] + (1-\omega)I(\boldsymbol{x}_{\mathcal{M}}, \boldsymbol{z}) \\ &= \mathbb{E}_{q(z|x_{\mathcal{O}},x_{\mathcal{M}},y)}[\log p(\boldsymbol{x}_{\mathcal{M}}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{z}) + \log p(y|\boldsymbol{z})] \\ &-\omega KL[q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)||p(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})] \\ &-(1-\omega)KL[q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})||p(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})],\end{aligned} \tag{3}$$

where the last divergence term is called the marginal KL divergence (Hoffman & Johnson, 2016). The full derivation for Eq. (3) is given in Appendix A.

The first KL divergence term in Eq. (3) has an analytical solution. The second KL divergence is intractable due to the marginal distribution $q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})$. It can, however, be replaced by any strict divergence term (Zhao et al., 2017), e.g. maximum mean discrepancy divergence (MMD) (Gretton et al., 2007). We select the squared population MMD since it encourages the average posterior distribution to match the whole prior, which is

$$\text{MMD}[\mathcal{F}, p, q] = \mathbb{E}_{p(x,x')}[k(x, x')] - 2\mathbb{E}_{p(x),q(z)}[k(x, z)] + \mathbb{E}_{q(z,z')}[k(z, z')]. \tag{4}$$

Here $\mathcal{F}$ is a unit ball in a universal reproducing kernel Hilbert space $\mathcal{H}$, p and q are Borel probability measures, and $k(\cdot, \cdot)$ is a universal kernel.

[2] Hence, subscripts in $\boldsymbol{x}_{\mathcal{O}}$ and $\boldsymbol{x}_{\mathcal{M}}$ indicate whether modalities are **o**bserved or **m**issing at test time. Missing modalities refer to modalities that are costly to obtain or modalities generated using cross-modal generation.

[3] We write the objective function for a single data point to improve readability. The outer expectation in the objective function for the average conditional log-likelihood is approximated with the empirical data distribution.

We use a Gaussian kernel in our proposed model. Finally, the objective function for a single data point therefore becomes

$$
\begin{aligned}
\mathcal{J}(\boldsymbol{x}_{\mathcal{M}}, \boldsymbol{x}_{\mathcal{O}}, y) = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}},\boldsymbol{x}_{\mathcal{M}})}&[\log p(\boldsymbol{x}_{\mathcal{M}}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{z}) + \alpha \log p(y|\boldsymbol{z})] \\
&- \omega KL[q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)||p(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})] \\
&- (1-\omega)\lambda \mathrm{MMD}[q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}), p(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})],
\end{aligned}
\tag{5}
$$

where λ counteracts the loss imbalance between the $\boldsymbol{x}_{\mathcal{O}}$ and $\mathcal{Z}$ spaces and α controls the importance of the classification loss in the objective function.

Effect of $\boldsymbol{\omega}$ on the objective function: The first (term-by-term) KL divergence in Eq. (5) regularizes each posterior distribution towards its prior and is minimized when $q_i(\boldsymbol{z}|\boldsymbol{x}^i_{\mathcal{O}}, \boldsymbol{x}^i_{\mathcal{M}}, y^i) = p_i(\boldsymbol{z}|\boldsymbol{x}^i_{\mathcal{O}})$ for all i. The marginal MMD divergence, on the other hand, regularizes an *average posterior* distribution $q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}) = 1/N \sum_i q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}^i_{\mathcal{M}}, y^i)$ towards the prior distribution, without sacrificing model power (Hoffman & Johnson, 2016). Makhzani et al. (2015) show that the term-by-term KL divergence simply encourages the average posterior distribution to match the modes of the prior $p(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})$. However, the MMD term in Eq. (5) encourages the average posterior distribution to match the whole prior, giving an effect similar to the adversarial training proposed by Makhzani et al. (2015). Furthermore, setting the marginal MMD divergence to 0 may lead to representations from the prior that are useless for sculpting latent representations (Hoffman & Johnson, 2016). Setting the term-by-term KL divergence to 0 also implies that the joint posterior representation is independent of the modality $\boldsymbol{x}_{\mathcal{M}}$. Our proposed objective function therefore offers an elegant way of trading-off these effects through the ω parameter, recovering the variational lower bound for $\omega = 1$ and, for $1 > \omega \geq 0$, optimizing mutual information. The optimal ω value, as can be seen in Sections 4.3.1, 4.3.5 and 4.4, is specific to the learning task and, therefore, must be found by cross-validation.

CMMD finally assumes the following density functions for the prior distribution, the classifier, and the encoder

$$
\begin{aligned}
p(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}) &\sim \mathcal{N}(\boldsymbol{\mu} = f_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathcal{O}}), \boldsymbol{\sigma}^2 = f_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathcal{O}})), \\
p(y|\boldsymbol{z}) &\sim \mathrm{Cat}(\pi_{y|z} = f_{\boldsymbol{\theta}}(\boldsymbol{z})), \\
q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y) &\sim \mathcal{N}(\boldsymbol{\mu} = f_{\boldsymbol{\phi}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y), \boldsymbol{\sigma}^2 = f_{\boldsymbol{\phi}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)).
\end{aligned}
$$

The decoder network is parametrized as

$$
\begin{aligned}
&p(\boldsymbol{x}_{\mathcal{M}}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{z}) \sim \mathcal{N}(\boldsymbol{\mu} = f_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{z}), \boldsymbol{\sigma}^2 = f_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{z})), \\
&\qquad\qquad \text{or} \\
&p(\boldsymbol{x}_{\mathcal{M}}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{z}) \sim \mathrm{Bernoulli}(\boldsymbol{p} = f_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{z})),
\end{aligned}
\tag{6}
$$

where $\mathcal{N}$ and Cat denote the Gaussian and multinomial distributions respectively, and where $f(\cdot)$ is a multilayer perceptron (MLP) network (Rumelhart et al., 1985). This means that the density parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}^2$, p, and $\pi_{y|z}$ are parametrized by neural networks, learnable parameters being denoted by $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. Note that the classifier can handle binary, multi-class, and multi-label classification using a sigmoid, softmax, or multiple sigmoid activation function respectively at the output layer.

We observed that $\boldsymbol{z} \sim q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)$ leads to an unstable classification of y. We therefore fed the classifier $p(y|\boldsymbol{z})$ with $\boldsymbol{z} \sim p(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})$ during training and test time. We hypothesize that the prior distribution reproduces the test scenario more accurately than the posterior distribution. Fig. 2 shows the forward propagation during training and test time in our proposed methodology[4].

4. Experiments and results

The experiments in this section assess the impact of our proposed model, which optimizes mutual information, on the four criteria that a multi-modal generative model should satisfy, i.e., latent factorization (Sections 4.2, 4.3.1, 4.3.2, 4.3.3, 4.3.4), coherent joint (Section 4.3.2) and cross generation (Sections 4.2, 4.3.2, 4.3.3, 4.3.4), and synergy (Sections 4.2, 4.3.1, 4.3.3, 4.3.4). To that end, we compare the CMMD model we propose with different multi-modal learning algorithms in downstream classification and generative tasks across different domains: image-to-image using a multi-modal version of MNIST and SVHN, image-to-text using a text describing digits pairs of MNIST-SVHN images, acoustic-to-articulatory with the XRMB data set, and image-to-annotation using the MIR Flickr data set. The benchmark models are: CCA (Hotelling, 1936), DCCA (Andrew et al., 2013), DCCAE (Wang et al., 2015a), MVCL (Hermann & Blunsom, 2013), RBM-MDL (Ngiam et al., 2011), VCCA (Wang et al., 2017), MVAE (Wu & Goodman, 2018), MMVAE (Shi et al., 2019), and MoPoE (Sutter et al., 2021). In addition, we include a classifier model M − $\boldsymbol{x}_{\mathcal{O}}$ that only uses the always available modality, to allow the impact of joint representations for classification to be assessed. Finally, Section 4.2 also provides an analysis of the impact of mutual information optimization in the CMMD model on posterior collapse.

Network architectures and model training details are given in Appendix C. However, given the importance of the ω hyperparameter in the optimization of our proposed model, we mention here the value found by cross-validation in each experiment, unless otherwise specified. See Fig. C.1 for an overview over all ω values.

4.1. Data sets

In the following we explain the multi-modal data sets used in this research.

2-modality MNIST: This data set, introduced by Wang et al. (2015a), consists of 28 × 28 MNIST hand-written digit images (Deng, 2012). The images have been randomly rotated at angles in the interval $[-\pi/4, \pi/4]$, to generate $\boldsymbol{x}_{\mathcal{O}}$. The modality $\boldsymbol{x}_{\mathcal{M}}$ is generated by randomly selecting a digit from $\boldsymbol{x}_{\mathcal{O}}$ and adding noise uniformly sampled from $[0, 1]$ to each pixel in the non-rotated image. Each pixel is then truncated to the interval $[0, 1]$.

MNIST-SVHN: We randomly paired each instance of a MNIST digit ($\boldsymbol{x}_{\mathcal{O}}$) with one instance of the same digit class in the SVHN data set (Netzer et al., 2011) ($\boldsymbol{x}_{\mathcal{M}}$), which is composed of street-view house numbers, just as in Shi et al. (2019).

3-modality MNIST: This data set combines some of the modalities in the previous data sets, i.e. original MNIST, rotated MNIST, and SVHN digits. All of the same digit class.

MNIST-SVHN-Text: This data set was first introduced in Sutter et al. (2020) and it is based on the MNIST-SVHN data set. The additional string modality contains 8 characters where everything is a blank space except the digit word. Further, the starting position of the word is chosen randomly. The 8 character string is, finally, converted to a 71D one-hot-encoding, which corresponds to the length of possible characters in the dictionary used in Sutter et al. (2020). The experiments using this data set consider all possible combinations of missing and observable modalities, see Section 4.3.2.

XRMB: The original XRMB data set (Westbury, 1994) contains simultaneously recorded speech and articulatory measurements from 47 American English speakers. The modality $\boldsymbol{x}_{\mathcal{O}}$, the acoustic data, is composed of a 13D vector of mel-frequency cepstral coefficients (MFCCs). We also included their first and second derivatives. This 39D vector is concatenated over a 7-frame window around each frame, resulting in a 273D vector that corresponds to $\boldsymbol{x}_{\mathcal{O}}$. The modality $\boldsymbol{x}_{\mathcal{M}}$, the articulatory data, is formed by horizontal and vertical displacements of

[4] Full code of the model is available at https://github.com/rogelioamancisidor/cmmd.

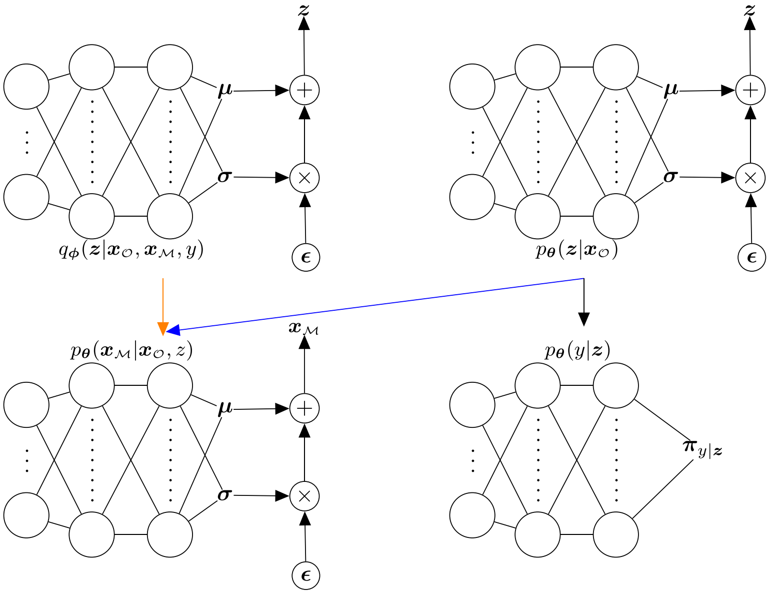

$\boldsymbol{z}$
$\boldsymbol{\mu}$
$\boldsymbol{\sigma}$
$\boldsymbol{\epsilon}$
$q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)$
$\boldsymbol{z}$
$\boldsymbol{\mu}$
$\boldsymbol{\sigma}$
$\boldsymbol{\epsilon}$
$p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})$
$\boldsymbol{x}_{\mathcal{M}}$
$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathcal{M}}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{z})$
$\boldsymbol{\mu}$
$\boldsymbol{\sigma}$
$\boldsymbol{\epsilon}$
$p_{\boldsymbol{\theta}}(y|\boldsymbol{z})$
$\boldsymbol{\pi}_{y|\boldsymbol{z}}$

Fig. 2. Forward propagation in our proposed CMMD model. The orange arrow indicates a forward pass during training, which is replaced by the blue arrow at test time, i.e. the input to $p(\boldsymbol{x}_{\mathcal{M}}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{z})$ is $\boldsymbol{z} \sim q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)$ during training, while $\boldsymbol{z} \sim p(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})$ at test time. The black arrow depicts a common forward propagation during training and testing, i.e. the input to $p(y|\boldsymbol{z})$ is always $\boldsymbol{z} \sim p(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})$.

8 pellets on the tongue, lips, and jaw, resulting in a 112D vector. The data set then finally contains 40 phone classes.

Flickr: The Flickr data set (Huiskes & Lew, 2008) contains 1 million images, 25 000 being labeled according to 24 classes. Note that each image can be assigned to multiple classes. Stricter labeling was also carried out for 14 of the classes, images only being annotated with a category where that category was salient. The data set therefore has 38 classes. We used the same 3857D feature vector ($\boldsymbol{x}_{\mathcal{O}}$) as used by Srivastava and Salakhutdinov (2012) to describe the images. The modality $\boldsymbol{x}_{\mathcal{M}}$ is composed of tags related to the image, the tags constrained to a vocabulary of the 2000 most frequent words.

Table 1
Error rates (%) and rmse (lower is best) for the experiment using 4 randomly chosen folds (speakers IDs $[(1, 3), (43, 45), (10, 13), (27, 29)]$) from the XRMB data set, where the shared latent representation is generated using the available modality at test time. Note that VCCA cannot generate missing modalities in the scenario considered in this experiment. We add a baseline classifier M$-\boldsymbol{x}_{\mathcal{O}}$ that only uses $\boldsymbol{x}_{\mathcal{O}}$.

Fold	M$-\boldsymbol{x}_{\mathcal{O}}$	VCCA	MVAE	MMVAE ELBO	MMVAE IWAE	CMMD
1	39.9 –	40.2 –	45.5 1.07	54.9 1.29	48.9 0.77	**32.4** **0.74**
2	36.5 –	40.6 –	44.2 1.06	49.4 1.28	47.0 0.79	**31.2** **0.75**
3	54.9 –	55.4 –	56.8 1.09	61.9 1.17	60.2 0.83	**47.4** **0.77**
4	48.0 –	47.2 –	51.7 1.07	59.1 1.23	53.7 0.82	**38.2** **0.80**
Avg.	44.8 –	45.8 –	49.6 1.07	56.3 1.24	52.5 0.80	**37.3** **0.76**

4.2. Posterior collapse in multimodal learning

This section evaluates the impact of posterior collapse in VCCA, MVAE, MMVAE and our proposed CMMD model. We therefore measured posterior collapse as the proportion of latent dimensions that are within ϵ KL divergence of the prior for at least 99% of the data sample, as introduced by Lucas et al. (2019).

We trained all models using a 4-fold cross-validation approach, each fold containing 2 speakers from the XRMB data set (Westbury, 1994). Table 1 shows that CMMD, optimized with $\omega = 0.8$, outperforms all other methods in terms of error rates and root mean square errors (rmse) for the generated missing modality. VCCA[5] surprisingly ranks number two in the classification task, despite having a simpler architecture than MVAE and MMVAE. MVAE has lower error rates than MMVAE, even when we train MMVAE using an importance weighted approach and $k = 10$ samples. MMVAE IWAE generates the missing modality more accurately than MMVAE ELBO, and achieves smaller error rates.

The first two diagrams on the left side of Fig. 3 show the posterior collapse between $\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}$ and $\boldsymbol{z}$, and between $\boldsymbol{z}|\boldsymbol{x}_{\mathcal{M}}$ and $\boldsymbol{z}$. They show, for both versions of MMVAE, that around 80% of the dimensions in the latent representations collapse to $\mathcal{N}(\mathbf{0}, \mathbf{1})$. This implies that the latent representation is independent of the observed modalities. MVAE, however, needs more than 5 nats when conditioned on the modality $\boldsymbol{x}_{\mathcal{O}}$, and more than 6 nats when conditioned on view $\boldsymbol{x}_{\mathcal{M}}$ before 80% of the latent dimensions collapse. None of the latent dimensions in VCCA and CMMD are within 6 nats, and their latent representations therefore embed more information on the observed modalities. This information on the modalities is useful for downstream classification and, for CMMD, for the generation of the missing modality. The third diagram finally shows posterior collapse between the representations generated using $\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}$ and $\boldsymbol{z}|\boldsymbol{x}_{\mathcal{M}}$. We want, in this case, $\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}$ to collapse into $\boldsymbol{z}|\boldsymbol{x}_{\mathcal{M}}$, this meaning that the model is able to learn joint representations that contain information on $\boldsymbol{x}_{\mathcal{M}}$. Note that, in MMVAE, the collapse between both marginal distributions is strong given that both collapsed to $\mathcal{N}(\mathbf{0}, \mathbf{1})$. On the other hand, the marginal distributions in MVAE embed information on the modalities (see first two diagrams). MVAE, however, fails to learn joint representation as suggested by the third diagram. CMMD does, however, counteract posterior collapse through the conditional prior and through directly optimizing mutual information, as shown by the first three diagrams.

We adapted the posterior collapse definition to the analysis of the variance parameters in the generative process, to allow us to understand the rmse results for the generated missing modality. This insight

[5] In this experiment we learn the variance parameters for the decoder networks in VCCA for fair comparisons. The authors of the original paper that introduced VCCA, used fixed variance parameters. The results are given in Table 2.

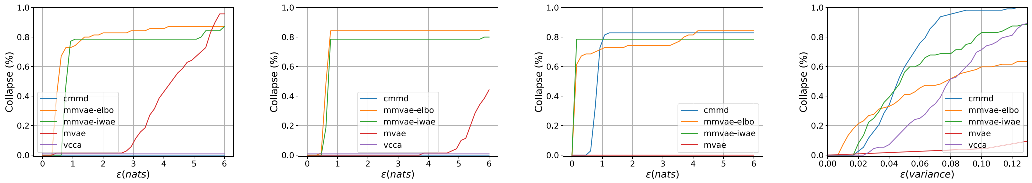


Fig. 3. Posterior collapse (from left to right: $KL[(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})||\boldsymbol{z}]$, $KL[(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{M}})||\boldsymbol{z}]$, and $KL[(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})||(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{M}})]$) in VCCA, MVAE, MMVAE, and CMMD. The far right diagram shows the adoption of the concept of posterior collapse to measure the variance parameters in the decoder generating $\boldsymbol{x}_{\mathcal{M}}$. For example, at $\epsilon = 0.06$ around 79% of the dimensions in $\hat{\boldsymbol{x}}_2$, generated by CMMD, have lower values than ϵ.

Table 2
We report error rates (lower is best) for experiments with MNIST and XRMB (average over speakers in the test dataset). For the Flickr data set, we report the mean average precision (mAP; higher is best). Results are based on Wang et al. (2017), except for values marked with † (which are from our own tests without pre-trained weights with Boltzmann machines) and results for CMMD.

Model name	Pretrain	MNIST error (%)	XRMB error (%)	Flickr mAP (%)
M-$\boldsymbol{x}_{\mathcal{O}}$	✗	13.1	37.6	48.0
DCCA	✓	2.9	–	–
DCCAE	✓	2.2	–	–
CCA	✗	19.1	29.4	52.9
DCCA	✗	4.7†	25.4	57.3
DCCAE	✗	4.4†	25.4	57.3
MVCL	✗	2.7	24.6	56.5
RBM-MDL	✗	11.7	29.4	47.7
VCCA	✗	3.0	28.0	60.5
VCCA-private	✗	**2.4**	25.2	61.5
MVAE	✗	6.0	39.8	**65.0**
MMVAE IWAE	✗	12.3	37.4	50.0
CMMD	✗	**2.4**	**21.1**	64.1

is shown in the last diagram of Fig. 3. For example at $\epsilon = 0.06$, around 79% of the dimensions in $\hat{\boldsymbol{x}}_2$ generated by CMMD, have lower values than ϵ. Furthermore, only 45% of the parameters learned by MMVAE ELBO have lower values than ϵ. We therefore hypothesize that MMVAE ELBO and MVAE overestimate the variance parameters in the decoder, resulting in higher rmse. The significant improvement for MMVAE IWAE seems to only change the decoder to a high capacity decoder and does not really improve the learned representations. Note that the variance collapse for VCCA is included for reference. It is actually generated using the modality $\boldsymbol{x}_{\mathcal{M}}$, which in theory is missing.

A mixture of experts and product of experts provide an elegant cross-generation in multi-modal learning, the joint posterior distribution being a linear combination of marginal parameters or distributions. Our approach to learning the posterior distribution is, however, to use a single encoder network, which can capture interactions between all modalities. The model we propose handles missing modalities using a conditional prior modulated by the available modalities. VCCA presents an interesting alternative to learning joint representations, the generative process embedding information on modalities into $\boldsymbol{z}$. VCCA cannot, however, generate missing modalities, which its generative model requires. Note that only CMMD has lower error rates than the baseline model M$-\boldsymbol{x}_{\mathcal{O}}$, which indicates that current variational multi-modal models are not suitable for learning useful joint representations for downstream classification. CMMD should therefore be preferred over VCCA, MVAE and MMVAE given that, in the setting of this research, CMMD outperforms concurrent models in downstream classification and in the generation of missing modalities at test time.

4.3. Classification and generating the missing modality

4.3.1. Image-to-image with MNIST

Table 2 shows that the performance of our proposed CMMD model is on a par with state-of-the-art models, including those that use pre-trained weights. We observed (practically) the same model performance for this data set at different ω values, our best model using a value of 0.4. Note that both DCCA and DCCAE use pre-trained weights with Boltzmann machines (BMs) (Salakhutdinov & Hinton, 2009). We therefore, for completeness, also retrained DCCA and DCCAE without using pre-trained weights. 2D t-SNEs of the latent space can be found in Appendix F.

We used (in a second analysis) the original version of MNIST as $\boldsymbol{x}_{\mathcal{O}}$ and the SVHN data set as $\boldsymbol{x}_{\mathcal{M}}$. Our best model used $\omega = 0.1$ and achieved a higher accuracy than MVAE and MMVAE, as shown in Table 3.

To show that CMMD can handle more than one missing and observed modality, we construct a 3-modality data set matching the class labels in: MNIST ($\boldsymbol{x}_1$), rotated MNIST ($\boldsymbol{x}_2$), and SVHN ($\boldsymbol{x}_3$). We used the same model parameters as were used in the previous experiment, and considered two test scenarios: (i) rotated MNIST and SVHN are both missing, i.e. $\boldsymbol{x}_{\mathcal{O}} = \boldsymbol{x}_1$ and $\boldsymbol{x}_{\mathcal{M}} = (\boldsymbol{x}_2, \boldsymbol{x}_3)$ and (ii) SVHN is missing, i.e. $\boldsymbol{x}_{\mathcal{O}} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$ and $\boldsymbol{x}_{\mathcal{M}} = \boldsymbol{x}_3$. The top (bottom) row in Table 4 shows the classification performance for the test scenario, in which two (one) modalities are missing. Generated modalities are shown in Appendix G.

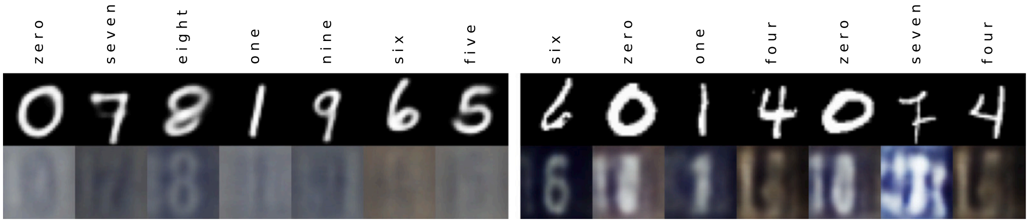
4.3.2. Image-to-text with MNIST and SVHN

Note that given $M = 3$ modalities, there are $2^M - 1 = 7$ combinations of observable modalities $\boldsymbol{x}_{\mathcal{O}}$.[6] We generate multimodal representations, conditioned on all of the possible combinations of observed modalities, with the CMMD model. After training, we randomly choose 500 representations from the training data set to train a multiclass logistic regression to classify true digits. Table 5 compares the classification performance of the CMMD model (see Fig. C.1 and Table H.2 to know the ω values used in these experiments), under this two-step classification approach, with that of MVAE, MMVAE, and MoPoE in similar experiments to those in Sutter et al. (2021) and Javaloy et al. (2022). We report model accuracy averaged over all 7 combinations of observable modalities and 5 different runs. Models ending with *IO* are trained with the impartial optimization approach introduced and reported in Javaloy et al. (2022).

We can see that IO increases the classification accuracy of all three models, especially for MMVAE. However, CMMD achieves higher discriminative power in all scenarios of observable modalities. Furthermore, Fig. 4 shows some examples of the images generated by CMMD. The left panel shows generated images for MNIST and SVHN modalities conditioned on the Text modality, while the right panel shows generated images for the SVHN modality conditioned on both Text and MNIST modalities. Note that the SVHN images in the right panel are sharper, compared to the left panel, since the generative model is conditioned on more observed modalities in that case. Details on model architectures and hyperparameter values are in Appendix H.

Finally, using the same models as before, we evaluate the quality of the generated missing modalities conditioned on all different combinations of observable modalities, i.e. conditional or cross-modal generation. To that end, we use the generative coherence metric, first introduced in Shi et al. (2019). Following previous works and same architectures as in Sutter et al. (2021), we train a classifier on the original

[6] These combinations are {M},{S},{T},{M, S},{M, T},{S, T}, and {M, S, T}, where M, S, and T refer to the MSNIT, SVHN, and Text modalities, respectively.



(a) MNIST and SVHN are missing

(b) SVHN is missing

Fig. 4. The left panel shows generated images for the MNIST and SVHN modalities conditioned on the observed Text modality using the CMMD model. The right panel shows generated images for the SVHN modality conditioned on Text and MNIST modalities, which are assumed to be observed at test time.

Table 3
Accuracy results for downstream classification with MNIST-SVHN. Results for MVAE and MMVAE are based on Shi et al. (2019).

Model name	MNIST-SVHN accuracy (%)
MVAE	95.7
MMVAE	91.3
CMMD	**97.6** ± 0.08%

Table 4
Accuracy results for 3-modality MNIST. The first experiment classifies using representations generated with $\boldsymbol{x}_{\mathcal{O}} = \boldsymbol{x}_1$, while the second experiment uses $\boldsymbol{x}_{\mathcal{O}} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$.

Missing modality	Accuracy (%)
$\boldsymbol{x}_{\mathcal{M}} = (\boldsymbol{x}_2, \boldsymbol{x}_3)$	97.5 ± 0.25%
$\boldsymbol{x}_{\mathcal{M}} = \boldsymbol{x}_3$	98.9 ± 0.13%

Table 5
Accuracy performance (%), averaged over all 7 combinations of observable modalities and 5 different runs with the MNIST-SVHN-Text data set, of the MVAE, MMVAE, MoPoE, and CMMD models. Additionally, we include the results from Javaloy et al. (2022) where MVAE, MMVAE, and MoPoE are trained with impartial optimization.

Model	Javaloy et al. (2022)	Sutter et al. (2021)	Ours
MVAE	69.7	83.1	–
MVAE-IO	70.0	–	–
MMVAE	87.6	89.0	–
MMVAE-IO	90.8	–	–
MoPoE	89.9	95.1	–
MoPoE-IO	91.5	–	–
CMMD	–	–	**96.5**

unimodal training data set to classify the generated modalities. If the classifier detects the same attributes in the generated samples, it is a coherent generation. Further, we use classification accuracy to measure the quality of generated samples. Table 6 shows accuracy values of the conditionally generated modalities averaged over 5 different runs. The letter at the top indicates the modality being generated based on the different sets of modalities below, where M, S, and T stands for MNIST, SVHN, and Text modalities. CMMD achieves higher accuracy in most of the conditional generation scenarios.

4.3.3. Acoustic-to-articulatory with XRMB

The experimental setup and the data pre-processing used in this section are based on (Wang et al., 2017). Table 2 shows average error rates for all test speakers, CMMD outperforming previous models without a domain-specific classifier. For this experiment, CMMD is optimized with $\omega = 0.7$. Note that Wang et al. (2017) used the tandem speech recognizer (Hermansky et al., 2000) as classifier model in all the experiments they conducted. The tandem speech recognizer successfully couples neural networks and Gaussian mixtures models for word recognition and, in the benchmark results of Hermansky et al. (2000), reduced speech classification error rates by 35%. Wang et al. (2017) also used the 39D vector of MFCCs and the joint data representations as input data for the tandem recognizer for all experiments. We hypothesize that this further improves the performance of the tandem recognizer. The CMMD model we propose, however, only uses the shared data representations for classification.[7]

4.3.4. Image-to-annotation with Flickr

We use the same data set in this section as that used in Srivastava and Salakhutdinov (2012). Most of the Flickr data corresponds to unlabeled images. We therefore used a two-stage training approach. Firstly, we trained our proposed model, but without the classifier and omitting the class label in the encoder, i.e. $q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}})$. Secondly, we used the weights from the first stage in the corresponding networks of Eq. (5) and used random weights at initialization for y in the encoder $q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)$.

Following the standards set by previous research, we use the mean average precision (mAP) to measure the classification performance of our proposed CMMD model for 10 000 randomly selected images. Table 2 shows that CMMD, optimized with $\omega = 0.5$, and MVAE outperform previously proposed image classification methods.

4.3.5. Acoustic inversion and annotation generation

We tested the generative process $p(\boldsymbol{x}_{\mathcal{M}}|\boldsymbol{x}_{\mathcal{O}}, z)$ in CMMD in image-to-annotation mapping and also acoustic-to-articulatory (called acoustic inversion (AI)). The scarce availability of articulatory data (Badino et al., 2017) makes acoustic inversion an important field. Table 8 shows, on the test data set, that CMMD outperforms the rmse for AI reported in Wang et al. (2015b), which is based on the training and validation data set (1.17 and 1.96, respectively). Our results also outperform the average rmse of 2.14 obtained on the test dataset of Badino et al. (2017).

The second experiment involves generating tags, which can be costly to obtain, that describe a given picture in the Flickr data set. We used our trained model from the previous section and compared it with the deep Boltzmann machine (DBM) model (Srivastava & Salakhutdinov, 2012), MVAE, and MMVAE. We furthermore tested all models on different images and with different levels of complexity. Table 7 shows some of the generated tags. More examples are given in Appendix D.

The generative process in CMMD generates quality articulatory and annotation samples at test time. The results suggest that the prior distribution in our proposed model learns joint representations through the optimization of our proposed objective function, which maximizes mutual information between representations and the missing modality at test time.

[7] In the current version of the data set, it is not possible to identify the 39D vector of MFCCs in the 273D vector for the modality $\boldsymbol{x}_{\mathcal{O}}$. We could therefore not concatenate MFCCs and joint representations for training the classifier.

Table 6
Accuracy values of the conditionally generated modalities averaged over 5 different runs. The letter at the top indicates the modality being generated based on the different sets of modalities below, where M, S, and T stands for MNIST, SVHN, and Text modalities, respectively.

Model	M			S			T		
	S	T	S, T	M	T	M, T	M	S	M, S
MVAE	0.24	0.20	0.32	0.43	0.30	0.75	0.28	0.17	0.29
MVAE-IO	0.11	0.26	0.28	0.50	0.33	0.30	0.61	0.12	0.64
MMVAE	**0.75**	0.99	0.87	0.31	0.30	0.30	0.96	**0.76**	0.84
MMVAE-IO	0.49	0.79	0.64	**0.87**	0.76	0.82	0.97	0.58	0.77
MoPoE	0.74	0.99	0.94	0.36	0.34	0.37	0.96	**0.76**	0.93
MoPoE-IO	0.11	0.63	0.52	0.28	0.47	0.43	0.80	0.11	0.90
CMMD	**0.75**	**1.00**	**1.00**	0.66	**0.87**	**0.87**	**0.98**	0.69	**0.98**

Table 7
Tags describing images are generated with the multi-modal learning deep Boltzmann machine (DBM) (Srivastava & Salakhutdinov, 2012) and with CMMD. DBM fails to generate coherent tags in the first 3 images. CMMD is, however, able to generate meaningful tags. In the last image, both models generate coherent tags.

Generated tags DBM	water, glass, wine, drink, beer, bubbles, splash, drops, drop	portrait, women, soldier, postcard soldiers, army	nikon, d200, tamron, d300, f28, sb600, d60 nikkor, d50, d90	foliage, autumn, trees, leaves, fall, forest, woods, path
Generated tags MVAE	–	statue	car, performance	a700
Generated tags MMVAE IWAE	canon, night, 2007	nikon, green, lion	flower	trees, autumn
Generated tags CMMD	sign, fisheye	animal, lion, outdoors, zoo, k10d, challengeyouwinner, boston, wildlife	apple, food	nature, light autumn, leaves wood, path, forest

Table 8
We report rmse for AI and error rates (%) for downstream classification in a speaker-independent experiment for eight speakers. Average and standard deviation (std) values are shown at the bottom.

XRMB					
Speaker ID	AI rmse	Classification error (%)	Speaker ID	AI rmse	Classification error (%)
7	0.80	15.8	23	0.79	26.8
16	0.76	21.4	28	0.57	22.5
20	0.73	16.2	31	0.77	20.6
21	0.78	25.8	35	0.78	20.2
average 8 speakers	0.75	21.2	std 8 speakers	0.07	3.7

4.4. Analysis of the objective function

In this section, we train the CMMD model using the XRBM data set for all speakers in Table 8 using a speaker-dependent approach, i.e. 70%–30% of the data for each speaker used for training-testing, unless otherwise specified. We furthermore trained the CMMD model in two ways: (i) we fine-tuned ω in the range $[0, 0.1, \ldots, 1]$ and (ii) we used $\omega = 0.7$ (which is the optimal value in the previous section) and fixed the variance parameters in the decoder network $p(\boldsymbol{x}_{\mathcal{M}}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{z})$ to the same value as in Wang et al. (2017), i.e. $\sigma^2 = 0.01$.

Impact of fixed variance parameters:. The second diagram in Fig. 5 shows some between-variability in the modality $\boldsymbol{x}_{\mathcal{M}}$. Fixing the variance parameters in $p(\boldsymbol{x}_{\mathcal{M}}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{z})$ therefore deteriorates error rates, as shown in the first diagram.

Should we optimize the ELBO?. The third panel in Fig. 5 compares error rates for the ELBO (dashed line), recovered for $\omega = 1$, and for our proposed objective function (Eq. (5)) with fine-tuned ω. Our proposed objective function achieves lower error rates for all speakers. The rmse for the generated features in $\boldsymbol{x}_{\mathcal{M}}$ are also smaller when we optimize our proposed objective function.

The top panel of Fig. 6 shows the posterior collapse in the CMMD model for $\omega = 0$ and $\omega = 1$; the latter optimizes the ELBO, while the former optimizes mutual information (in addition to the generative and classifier models). Recall that the main motivation to include the mutual information term $I(\boldsymbol{x}_{\mathcal{M}}, \boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})$ is to counteract the posterior collapse problem and, from the figure, it is clear that CMMD avoids the posterior collapse problem by optimizing mutual information. However, as shown by the four panels in the middle and bottom of Fig. 6 (in which we add the relatively more complex learning task presented in Section 4.2, but varying ω) optimizing only mutual information harms the performance of the generative and classifier models, reflected in the rmse and error rate respectively. Note that only optimizing mutual information accounts for the minimization of an average MMD divergence measure. That is, we only minimize the divergence from the average conditional posterior $q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})$ to the conditional prior. Our results confirm that minimizing an average divergence measure makes the prior distribution, which is used for downstream tasks, unable to sculpt latent representations as suggested by Hoffman and Johnson (2016). On the other hand, only optimizing the term-by-term KL divergence leads to latent representations $\boldsymbol{z}$ that are independent from $\boldsymbol{x}_{\mathcal{M}}$, which turns out to be relatively less harmful for downstream tasks. Fortunately, our proposed objective function offers a way of trading-off these two effects and, as can be seen in the middle and bottom rows in Fig. 6, there is an ω region in which the generative and classifier models

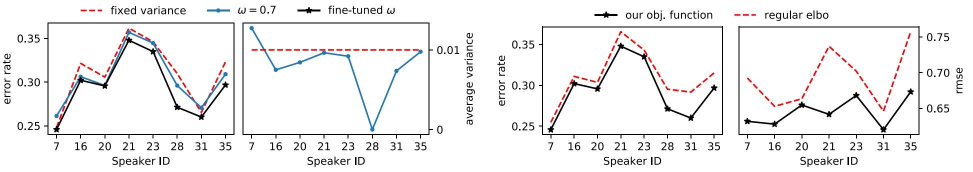


Fig. 5. The 1st and 3rd plot show error rates for the speaker-dependent experiments (Section 4.4). The 2nd plot shows average variances of all generated $x_{\mathcal{M}}$ features. The last plot compares rmse for their generated values. Speaker 28 was removed as the rmse in both cases is roughly 0.

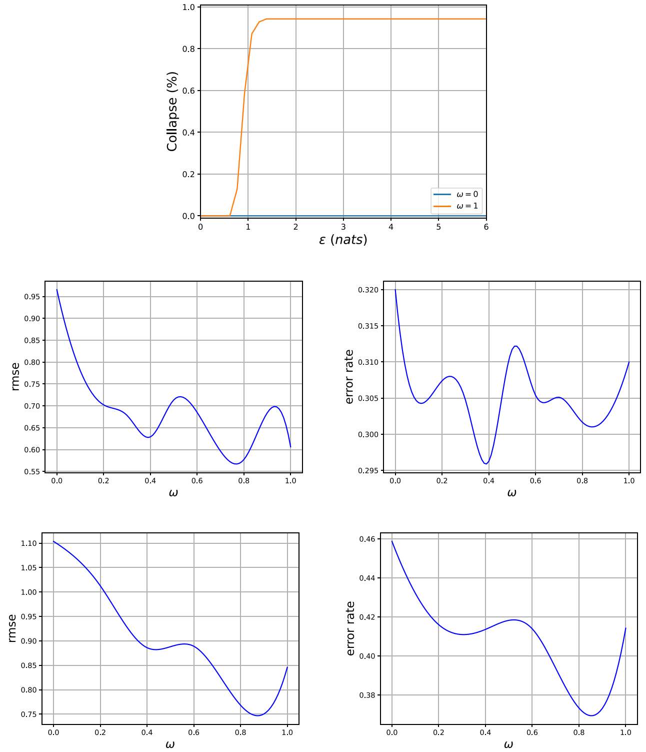


Fig. 6. The top panel shows the posterior collapse in the CMMD model for $\omega = 0$ and $\omega = 1$. In both cases, we use data for speaker 7 in the XRMB data set. The two panels in the middle show average rmse and error rate as a function of ω for speakers 7, 16, 20, 21, 23, 28, 31, and 35 in the XRMB data set. Finally, the two panels in the bottom show average rmse and error rate values in the cross-validation approach introduced in Section 4.2.

achieve higher performance. Hence, the optimal ω value is specific to the learning task and must be found by cross-validation.

How much overhead does mutual information optimization add? We use the MNIST-SVHN-Text data set to measure training time for $\omega = 0$ and $\omega = 1$. The average training time for processing one batch with 256 observations is 10.59 ms if the ELBO is optimized. On the other hand, the average training time to optimize our proposed objective function, including mutual information, is 11.04 ms, which is the same training time for $1 > \omega > 0$. Therefore, our proposed objective function does not add significant overhead and is able to achieve higher performance in the downstream tasks considered in this research.

5. Conclusion

This research shows that the variational lower bound on the conditional likelihood has a Kullback–Leibler divergence that limits the amount of information on the modalities embedded in the joint representation. We, to counteract this effect, propose a novel likelihood-free objective function that optimizes the mutual information between joint representations and the modalities that we are interested in generating at test time. Our proposed CMMD model furthermore uses an informative prior distribution that is conditioned on the modalities that are always available. We analyze the negative effects of posterior collapse in downstream classification and in the generative process of multi-modal learning models.

The empirical results show that the objective function we propose achieves higher downstream classification performance and lower rmse in the generated modalities than the regular variational lower bound. The model we propose also successfully counteracts the posterior collapse problem by optimizing mutual information and by using an informative prior. Finally, the higher performance of our proposed CMMD model with respect to the state-of-the-art is consistent across different representative multi-modal problems.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Objective function

The joint distribution in the CMMD model is $p(\boldsymbol{x}_{\mathcal{M}}, y, \boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}) = p(\boldsymbol{x}_{\mathcal{M}}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{z})p(y|\boldsymbol{z})p(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})$ and, under this model specification, the posterior distribution $p(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)$ is intractable. Therefore, CMMD uses VI and approximates the true posterior distribution with a variational density $q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)$. Hence, the variational lower bound on the marginal log-likelihood of a single observation is

$$\begin{aligned}
\log p(\boldsymbol{x}_{\mathcal{M}}, y|\boldsymbol{x}_{\mathcal{O}}) &= \log \int p(\boldsymbol{x}_{\mathcal{M}}, y, \boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})d\boldsymbol{z} \\
&= \log \int q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)\frac{p(\boldsymbol{x}_{\mathcal{M}}, y, \boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})}{q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)}d\boldsymbol{z} \\
&= \log \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)}\frac{p(\boldsymbol{x}_{\mathcal{M}}, y, \boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})}{q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)} \\
\geq & \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)}\left[\log \frac{p(\boldsymbol{x}_{\mathcal{M}}, y, \boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})}{q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)}\right] \\
&= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)}[\log p(\boldsymbol{x}_{\mathcal{M}}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{z}) + \log p(y|\boldsymbol{z}) + \log p(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}) \\
& - \log q(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, y)], \qquad \text{(A.1)}
\end{aligned}$$

where the inequality is a result of the concavity of log and Jensen's inequality.

Now we can write the conditional mutual information term $I_e(\boldsymbol{x}_{\mathcal{M}}, \boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})$ (which depends on the functional form of the encoder as denoted by the subscript), as follows

$$I_e(\boldsymbol{x}_{\mathcal{M}}, \boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}}) = \mathbb{E}_{p(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{M}}, \boldsymbol{z})}\left[\log \frac{p_e(\boldsymbol{x}_{\mathcal{M}}, \boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})}{p_e(\boldsymbol{x}_{\mathcal{M}}|\boldsymbol{x}_{\mathcal{O}})p_e(\boldsymbol{z}|\boldsymbol{x}_{\mathcal{O}})}\right]$$

$$
\begin{aligned}
&= \mathbb{E}_{p(\mathbf{x}_{\mathcal{O}},\mathbf{x}_{\mathcal{M}},z)}\left[\log \frac{p_e(z|\mathbf{x}_{\mathcal{O}},\mathbf{x}_{\mathcal{M}})p_e(\mathbf{x}_{\mathcal{M}}|\mathbf{x}_{\mathcal{O}})}{p_e(\mathbf{x}_{\mathcal{M}}|\mathbf{x}_{\mathcal{O}})p_e(\mathbf{z}|\mathbf{x}_{\mathcal{O}})}\right] \\
&= \mathbb{E}_{p(\mathbf{x}_{\mathcal{O}},\mathbf{x}_{\mathcal{M}},z)}[\log p_e(z|\mathbf{x}_{\mathcal{O}},\mathbf{x}_{\mathcal{M}}) - \log p_e(\mathbf{z}|\mathbf{x}_{\mathcal{O}}) + \log p(\mathbf{z}|\mathbf{x}_{\mathcal{O}}) \\
&- \log p(\mathbf{z}|\mathbf{x}_{\mathcal{O}})] \\
&= \mathbb{E}_{p(\mathbf{x}_{\mathcal{M}},\mathbf{x}_{\mathcal{O}})}\left[KL[p_e(z|\mathbf{x}_{\mathcal{O}},\mathbf{x}_{\mathcal{M}})||p(\mathbf{z}|\mathbf{x}_{\mathcal{O}})]\right] \\
&- \mathbb{E}_{p(\mathbf{x}_{\mathcal{O}})}\left[KL[p_e(\mathbf{z}|\mathbf{x}_{\mathcal{O}})||p(\mathbf{z}|\mathbf{x}_{\mathcal{O}})]\right],
\end{aligned}
\tag{A.2}
$$

where $p(\mathbf{x}_{\mathcal{O}},\mathbf{x}_{\mathcal{M}},\mathbf{z}) = p(\mathbf{x}_{\mathcal{O}})p(\mathbf{x}_{\mathcal{M}}|\mathbf{x}_{\mathcal{O}})p(\mathbf{z}|\mathbf{x}_{\mathcal{O}})$, $p(\mathbf{x}_{\mathcal{M}},\mathbf{z}|\mathbf{x}_{\mathcal{O}}) = p(\mathbf{z}|\mathbf{x}_{\mathcal{M}},\mathbf{x}_{\mathcal{O}})$ $p(\mathbf{x}_{\mathcal{M}}|\mathbf{x}_{\mathcal{O}})$, $\int p(\mathbf{z}|\mathbf{x}_{\mathcal{M}},\mathbf{x}_{\mathcal{O}})p(\mathbf{x}_{\mathcal{M}})d\mathbf{x}_{\mathcal{M}} \approx 1/N\sum_n p_e(\mathbf{z}|\mathbf{x}_{\mathcal{O}},\mathbf{x}^n_{\mathcal{M}}) = p_e(\mathbf{z}|\mathbf{x}_{\mathcal{O}})$, and all probability density functions are approximated by variational approximations (the encoder and prior distribution in our proposed model). The expectations $\mathbb{E}_{p(\mathbf{x}_{\mathcal{M}},\mathbf{x}_{\mathcal{O}})}$ and $\mathbb{E}_{p(\mathbf{x}_{\mathcal{O}})}$ are finally estimated using the empirical data distribution $\tilde{p}_D$.

Adding the conditional mutual information term $(1-\omega)I_e(\mathbf{x}_{\mathcal{M}},\mathbf{z}|\mathbf{x}_{\mathcal{O}})$ to the lower bound in Eq. (A.1) (mutual information optimization being controlled by $\omega \in [0,1]$) and replacing p_e with the encoder $q(\mathbf{z}|\mathbf{x}_{\mathcal{O}},\mathbf{x}_{\mathcal{M}},y)$[8] gives the likelihood-free objective function for a single data point

$$
\begin{aligned}
&\mathcal{L}(\mathbf{x}_{\mathcal{O}},\mathbf{x}_{\mathcal{M}},y) + (1-\omega)I_e(\mathbf{x}_{\mathcal{M}},\mathbf{z}|\mathbf{x}_{\mathcal{O}}) \\
&\quad = \mathbb{E}_{q(z|\mathbf{x}_{\mathcal{O}},\mathbf{x}_{\mathcal{M}},y)}[\log p(\mathbf{x}_{\mathcal{M}}|\mathbf{x}_{\mathcal{O}},\mathbf{z}) + \log p(y|\mathbf{z}) + \log p(\mathbf{z}|\mathbf{x}_{\mathcal{O}}) \\
&\quad - \log q(\mathbf{z}|\mathbf{x}_{\mathcal{O}},\mathbf{x}_{\mathcal{M}},y)] \\
&+(1-\omega)KL[q(z|\mathbf{x}_{\mathcal{O}},\mathbf{x}_{\mathcal{M}},y)||p(\mathbf{z}|\mathbf{x}_{\mathcal{O}})] - (1-\omega)KL[q(\mathbf{z}|\mathbf{x}_{\mathcal{O}})||p(\mathbf{z}|\mathbf{x}_{\mathcal{O}})] \\
&\quad = \mathbb{E}_{q(z|\mathbf{x}_{\mathcal{O}},\mathbf{x}_{\mathcal{M}},y)}[\log p(\mathbf{x}_{\mathcal{M}}|\mathbf{x}_{\mathcal{O}},\mathbf{z}) + \log p(y|\mathbf{z})] \\
&\quad - \omega KL[q(\mathbf{z}|\mathbf{x}_{\mathcal{O}},\mathbf{x}_{\mathcal{M}},y)||p(\mathbf{z}|\mathbf{x}_{\mathcal{O}})] \\
&-(1-\omega)KL[q(\mathbf{z}|\mathbf{x}_{\mathcal{O}})||p(\mathbf{z}|\mathbf{x}_{\mathcal{O}})] \\
&\quad \equiv \mathcal{J}(\mathbf{x}_{\mathcal{O}},\mathbf{x}_{\mathcal{M}},y).
\end{aligned}
\tag{A.3}
$$

Note that we can obtain unbiased samples from $q(\mathbf{z}|\mathbf{x}_{\mathcal{O}})$ by first randomly sampling tuples $(\mathbf{x}_{\mathcal{M}},y) \sim \tilde{p}_D$ and then $\mathbf{z} \sim q(\mathbf{z}|\mathbf{x}_{\mathcal{O}},\mathbf{x}_{\mathcal{M}},y)$. These are used to estimate the MMD divergence term in Eq. (5).

Appendix B. Upperbound on mutual information

Using the last line in Eq. (A.2) and replacing p_e with the encoder $q(\mathbf{z}|\mathbf{x}_{\mathcal{O}},\mathbf{x}_{\mathcal{M}},y)$, which acknowledges the access to a labeled data set, it follows that

$$
\mathbb{E}_{p(\mathbf{x}_{\mathcal{M}},\mathbf{x}_{\mathcal{O}},y)}[KL[q(\mathbf{z}|\mathbf{x}_{\mathcal{O}},\mathbf{x}_{\mathcal{M}},y)||p(\mathbf{z}|\mathbf{x}_{\mathcal{O}})]] \geq I(\mathbf{x}_{\mathcal{M}},\mathbf{z}|\mathbf{x}_{\mathcal{O}}) \tag{B.1}
$$

given that the KL divergence is strictly positive. The expectation can be estimated using the empirical data distribution $\tilde{p}_D$.

Appendix C. Model training and architectures

We minimized Eq. (5) using SVGB and automatic differentiation routines in Theano (Team et al., 2016). Note that the reconstruction term of Eq. (5) can be efficiently estimated using the *reparameterization trick* (Kingma & Welling, 2013). The KL divergence term has a closed-form expression (Kingma & Welling, 2013; Mancisidor et al., 2020), and the MMD divergence is approximated numerically by drawing samples, as explained in Section A. This is the method suggested by Zhao et al. (2017) and Rezaabad and Vishwanath (2020).

CMMD architectures are, to provide a fair comparison in all experiments, chosen to resemble previous works. We furthermore use softplus activation functions in all hidden layers, using dropout (Srivastava et al., 2014) with 0.2 probability. We use the same α and λ parameter values for all CMMD models, which are set to 10 and 1000 respectively. We furthermore tune the hyperparameter ω over the grid $[0, 0.1, 0.2, \ldots, 1]$. Fig. C.1 shows the optimal value of ω for each experiment in this research, which is found by cross-validation. Finally, we use the Adam optimizer (Kingma & Ba, 2014) with a 10^{-4} learning rate in all experiments. Our model is implemented in Theano and trained on a GeForce GTX 1080 GPU.

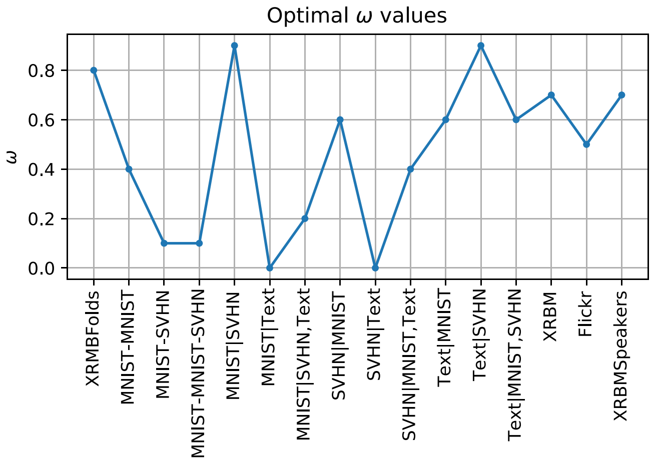


Fig. C.1. Optimal ω value, found by cross-validation, for each of the experiments in this research. Experiments are ordered chronologically.

Image-to-Image with MNIST: The encoder network uses 3 hidden layers of 2500 neurons. Both the prior distribution and the decoder use 3 layers of 1024 neurons. The latent variable is a 50D vector and the classifier uses 2 hidden layers of 50 neurons. We assume, given that the second view is almost a continuous variable, that it is Gaussian distributed.

Image-to-Image with MNIST-SVHN: The encoder, decoder and prior distribution in this experiment have 1 hidden layer of 400 neurons. The latent representation is a 20D vector and the classifier has 2 hidden layers of 50 neurons each.

3-modality MNIST: We use the same encoder in this experiment as in the "image-to-image with MNIST" experiment. The decoder architecture is shown in Table H.3 (Decoder columns), which is the same architecture as in Shi et al. (2019). We add an extra layer, such as the one at the bottom of Table H.3, but with 1 stride to generate two missing modalities ($\mathbf{x}_2$ and $\mathbf{x}_3$). Note that we, for the rotated MNIST images, pad the images to a 32 × 32 matrix during training, and crop-back to a 28 × 28 matrix at test time. The decoder loss is, finally, the sum of two cross entropy terms, one for each missing modality.

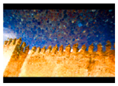
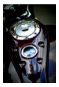
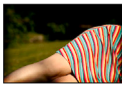
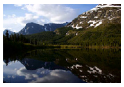
Acoustic-to-Articulatory with XRMB: We trained our model using the same 35 speakers used by Wang et al. (2015a, 2017). The current version of the test data set, however, only contains 8 speakers without silence frames (silence frames were removed in the other 35 speakers). Our model is, for this reason, tested on 8 speakers in a speaker-independent downstream classification task (Table 2).

We use an encoder with 3 hidden layers of 3000 neurons. The prior distribution and decoder each have 3 hidden layers of 1500 neurons. The classifier model has 2 hidden layers of 100 neurons and the latent shared representation is a 70D vector. We assume a Gaussian distribution for modality $\mathbf{x}_{\mathcal{M}}$ in this case. The ω parameter has, for this data set, a significant impact on downstream classification and our best model uses $\omega = 0.7$.

Image-to-Annotation with Flickr: We use an encoder with 4 hidden layers of 2048 neurons each. The prior distribution and decoder use 4 hidden layers of 1024 neurons. We, given that the modality $\mathbf{x}_{\mathcal{M}}$ corresponds to tags, use a Bernoulli decoder. The shared representation is a 1024D vector and our best model uses $\omega = 0.5$. We deal with multi-label classification in this data set. The classifier for this model therefore

[8] In this case the encoder is a variational approximation that can take any arbitrary form as long as it is a valid probability distribution (Sutter et al., 2021).

Table C.1
Tags generated using our proposed CMMD for some labeled images in the Flickr data set.

					
Generated tags	reflection, themoulinrouge	beauty, stone, sculpture	car	chrome	white, yellow, abstract, bus, lines, graphic
Generated tags	flower, holiday, vacation, red	flower, layers, textures, iris, soe	food, vegan, cupcake	d80	landscape, explore, flickr
					
Generated tags	abigfave, minimal, buenosaires, wall, impressedbeauty	macro, tamron, nikond40, india	flower, explore, flores	bike, red, tree, nyc, ny,	stripes
					
Generated tags	nature, water abigfave, landscape, reflection, mirror	horses, grey, friends	macro, garden, closeup	shoes, explore, selfportrait, 365days, toronto	california

uses 2000 neurons with sigmoid activations in the output layer and 2 hidden layers of 1550 neurons. Note that we follow previous works and exclude unlabeled images if they have less than 2 tags, given that we are interested in finding joint representations for both data-modalities. We finally standardize all features in the modality $\boldsymbol{x}_{\mathcal{O}}$.

Appendix D. Generating tags

Table C.1 shows tags generated using our proposed CMMD model for some labeled images in the Flickr data set. Note that none of the images have any tag in the original data set.

Appendix E. Additional details on posterior collapse

We use the posterior collapse definition introduced in Lucas et al. (2019). This, in our experiments, is $Pr(KL[q(\cdot)||p(\cdot)] < \epsilon) \geq 1-\delta$, where $\delta = 0.01$ and $\epsilon \in [0, 6]$. We therefore measure the proportion of latent dimensions i that are within ϵ KL divergence for at least $1-\delta$ of the data points. The MMVAE, MVAE, and VCCA models are, in our experiments, trained using the authors' publicly available codes[9].

Fig. 3 shows different measures of collapse for Fold 1[10] for Section 4.2 experiments. The far left diagram shows posterior collapse $Pr(KL[(z_i|\boldsymbol{x}_{\mathcal{O}})||(z_i)] < \epsilon) \geq 1-\delta)$, where $(\boldsymbol{z}_i) \sim \mathcal{N}(0, 1)$ and $(z_i|\boldsymbol{x}_{\mathcal{O}})$ are drawn from the prior distribution, the joint MoE posterior, the joint PoE posterior, and the shared inference distribution for CMMD, MMVAE, MVAE, and VCCA respectively.

The second diagram in Fig. 3 calculates $Pr(KL[(z_i|\boldsymbol{x}_{\mathcal{M}})||(z_i)] < \epsilon) \geq 1-\delta)$. However, $z_i \sim N(0, 1)$ and $(z_i|\boldsymbol{x}_{\mathcal{M}})$ are, for this, drawn from the inference posterior distribution, the joint MoE posterior, the joint PoE posterior, and the inference private distribution in CMMD, MMVAE, MVAE, and VCCA respectively. Finally, the third diagram in Fig. 3 calculates $Pr(KL[(z_i|\boldsymbol{x}_{\mathcal{O}})||(z_i|\boldsymbol{x}_{\mathcal{M}})] < \epsilon) \geq 1-\delta)$, where $(z_i|\boldsymbol{x}_{\mathcal{O}})$ and $(z_i|\boldsymbol{x}_{\mathcal{M}})$ are drawn as explained above.

Appendix F. Latent space - MNIST

Fig. F.1 shows 2D t-SNEs (Van der Maaten & Hinton, 2008) of the latent space learned using CMMD, MVAE and VCCA. The t-SNEs for both CMMD and VCCA show well separated class labels. Note that the class label variability is larger for the CMMD embeddings than VCCA. The t-SNEs for MVAE, however, show some overlapping class labels.

Appendix G. Generating multiple missing modalities

Fig. G.1 compares the missing modality/modalities generated at test time by the decoders in CMMD, MMVAE, and MVAE. In panel (a) we assume SVHN digits are missing at test time, in panel (b) both rotated-MNIST and SVHN are missing modalities at test time. In panel (c) and (d) we train MMVAE, optimizing the evidence lower bound (ELBO) and its importance weighted autoencoder (IWAE) version respectively, and generate the missing modality at test time (SVHN digits). For completeness, panel (e) shows the SVHN digits generated using MVAE reported in Shi et al. (2019). Both CMMD (panel (a)) and MMVAE-IWAE (panel (d)) generate quality and coherent SVHN digits, matching the MNIST digit in all cases. MVAE (panel (e)), however, generates low quality SVHN digits and it is difficult to see whether the generated image matches the MNIST digit. CMMD generates two missing modalities in panel (b), which is clearly a more challenging task. Only digits 7 and 0 are generated correctly for both missing modalities. Finally, it is interesting to compare the results obtained with MMVAE using two objective functions. If MMVAE optimizes the evidence lower bound, then the generated SVHN images have relatively low quality and do not match the MNIST class.

Appendix H. Cross-modal generation with MNIST-SVHN-text

The network architectures used in the experiments using the MNIST-SVHN-Text data set are shown in Tables H.3, H.4, and H.5, which are the same architectures used in Shi et al. (2019), Sutter et al. (2020, 2021) and Javaloy et al. (2022). The only difference is that we use the encoder architecture in the aforementioned methods for the prior distribution in the CMMD model. The encoder architecture in the CMMD model is a fully-connected neural network with 3 hidden layers,

[9] MMVAE: https://github.com/iffsid/mmvae,
MVAE: https://github.com/mhw32/multimodal-vae-public,
VCCA: https://ttic.uchicago.edu/∼wwang5/.

[10] The other 3 folds show the same pattern.

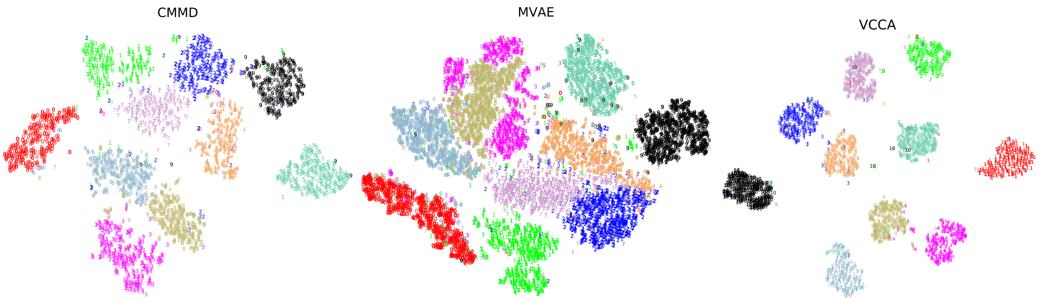
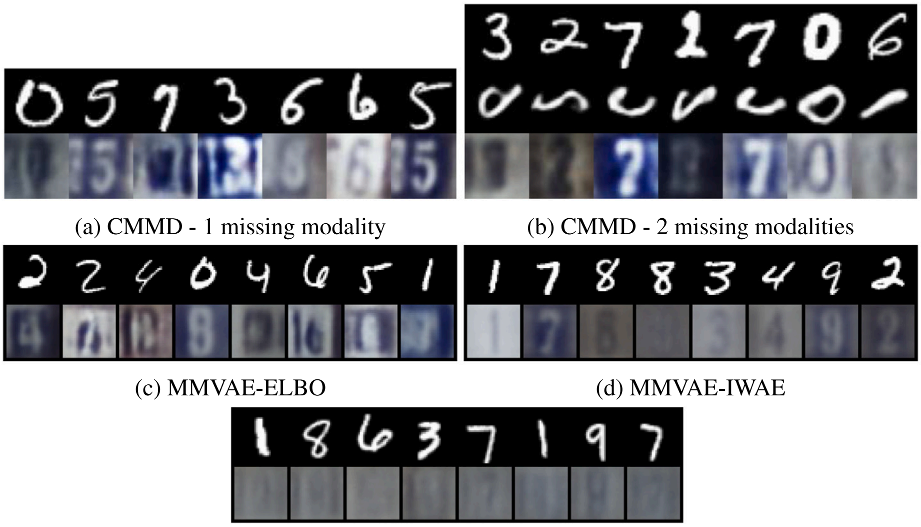


Fig. F.1. 2D t-SNEs of the latent space in CMMD, MVAE and VCCA. The scatter color is assigned by the class label.



(a) CMMD - 1 missing modality

(b) CMMD - 2 missing modalities

(c) MMVAE-ELBO

(d) MMVAE-IWAE

(e) MVAE

Fig. G.1. Generated images using CMMD (top row), MMVAE (middle row), and MVAE (bottom row). We, for all models, use the original MNIST digits to draw latent representations, which are further used to generate SVHN digits. Note that the MVAE images are taken from Shi et al. (2019).

Table H.1
Accuracy performance, averaged over 5 different runs, for all subsets of observable modalities. We do not include results for the method introduced in Javaloy et al. (2022) given that the authors only provided average values over the different set of observable modalities (see Table 5).

Model	M	S	T	M, S	M, T	S, T	M, S, T
MVAE	0.90 ± 0.01	0.44 ± 0.01	0.85 ± 0.10	0.89 ± 0.01	0.97 ± 0.02	0.81 ± 0.09	0.96 ± 0.02
MMVAE	0.95 ± 0.01	0.79 ± 0.05	0.99 ± 0.01	0.87 ± 0.03	0.93 ± 0.03	0.84 ± 0.04	0.86 ± 0.03
MoPoE	0.95 ± 0.01	0.80 ± 0.03	0.99 ± 0.01	0.97 ± 0.01	0.98 ± 0.01	0.99 ± 0.01	0.98 ± 0.03
CMMD	0.98 ± 4E-3	0.80 ± 4E-3	1.00 ± 0.00	0.99 ± 2E-3	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 2E-3

Table H.2
Accuracy values (generation coherence), averaged over 5 different runs, of the modalities conditionally generated by the CMMD model, together with the optimal ω values found by cross-validation.

	M			S			T		
	S	T	S, T	M	T	M, T	M	S	M, S
avg coherence	0.75	1.00	1.00	0.66	0.87	0.87	0.98	0.69	0.98
std coherence	0.02	0	0	0.07	0.07	0.10	2E-3	0.02	4E-3
ω	0.9	0	0.2	0.6	0	0.4	0.6	0.9	0.6

each with 2500 units. All layers in the encoder use softplus activation functions and a dropout layer with 0.2 probability. Following previous work, the multimodal representation is a 20D latent variable, and we use the same values for α and λ as in the other experiments, which are 10 and 1000 respectively. It is noteworthy that the 3 modalities are vectorized and concatenated before sending them through the encoder.

To make a fair comparison with previous methods, we implemented a two-step classification using the multinomial logistic regression model implemented by scikit-learn with default values. The logistic regression model is trained using 500 latent variables, which are generated and randomly selected from the train data set. Finally, we test the predictive power of the trained logistic regression on the entire test data set. Table H.1 shows the average accuracy over 5 different runs, for all subsets of observed modalities. For the cross-modal generation experiments, we train classifier models for each of the original unimodal modalities. The network architectures are the same as in the conditional prior column of Table H.3, H.4, and H.5, which again are the same architectures used in the aforementioned methods. Table H.2 shows the average coherence

Table H.3
SVHN conditional prior and decoder layers. The last column for each model specifies the kernel size, stride, padding, and dilation. All layers are 2D convolutional (conv) and upconvolutional (upconv) in the encoder and decoder, respectively, with ReLU activations. Finally, the number of input and output dimensions in each layer is shown in the columns #F.In and #F.Out, respectively.

Conditional prior					Decoder				
Layer	Type	#F.In	#F.Out	Spec.	Layer	Type	#F.In	#F.Out	Spec.
1	conv	3	32	(4, 2, 1, 1)	1	linear	20	128	
2	conv	32	64	(4, 2, 1, 1)	2	upconv	128	64	(4, 2, 0, 1)
3	conv	64	64	(4, 2, 1, 1)	3	upconv	64	64	(4, 2, 1, 1)
4	conv	64	128	(4, 2, 0, 1)	4	upconv	64	32	(4, 2, 1, 1)
5a	linear	128	20		5	upconv	32	3	(4, 2, 1, 1)
5b	linear	128	20						

Table H.4
MNIST conditional prior and decoder layers. All layers are linear with ReLU activations. Finally, the number of input and output dimensions in each layer is shown in the columns #F.In and #F.Out, respectively.

Conditional prior				Decoder			
Layer	Type	#F.In	#F.Out	Layer	Type	#F.In	#F.Out
1	linear	784	400	1	linear	20	400
2a	linear	400	20	2	linear	400	784
2b	linear	400	20				

Table H.5
Text conditional prior and decoder layers. The last column for each model specifies the kernel size, stride, padding, and dilation. All layers are 1D convolutional (conv) and upconvolutional (upconv) in the encoder and decoder, respectively, with ReLU activations. Finally, the number of input and output dimensions in each layer is shown in the columns #F.In and #F.Out, respectively.

Conditional prior					Decoder				
Layer	Type	#F.In	#F.Out	Spec.	Layer	Type	#F.In	#F.Out	Spec.
1	conv	71	128	(1, 1, 0, 1)	1	linear	20	128	
2	conv	128	128	(4, 2, 1, 1)	2	upconv	128	128	(4, 1, 0, 1)
3	conv	128	128	(4, 2, 0, 1)	3	upconv	128	128	(4, 2, 1, 1)
4a	linear	128	20		4	conv	128	71	(1, 1, 0, 1)
4b	linear	128	20						

and standard deviation over 5 different runs, together with the optimal ω value found by cross-validation in each experiment.

References

Abrol, V., Sharma, P., & Patra, A. (2020). Improving generative modelling in VAEs using multimodal prior. *IEEE Transactions on Multimedia*.

Andrew, G., Arora, R., Bilmes, J., & Livescu, K. (2013). Deep canonical correlation analysis. In *International conference on machine learning* (pp. 1247–1255).

Badino, L., Franceschi, L., Arora, R., Donini, M., & Pontil, M. (2017). A speaker adaptive DNN training approach for speaker-independent acoustic inversion. In *Proceedings of the annual conference of the international speech communication association, Vol. 2017* INTERSPEECH, (pp. 984–988). International Speech Communication Association (ISCA).

Chen, W., & Zhu, J. (2022). Multimodal adversarially learned inference with factorized discriminators. In *Proceedings of the AAAI conference on artificial intelligence. Vol. 36, no. 6* (pp. 6304–6312).

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, *29*(6), 141–142.

Dieng, A. B., Kim, Y., Rush, A. M., & Blei, D. M. (2019). Avoiding latent variable collapse with generative skip models. arXiv:1807.04863. [cs, stat].

Du, C., Du, C., Wang, H., Li, J., Zheng, W.-L., Lu, B.-L., & He, H. (2018). Semi-supervised deep generative modelling of incomplete multi-modality emotional data. In *2018 ACM multimedia conference on multimedia conference* (pp. 108–116). ACM.

Du, F., Zhang, J., Hu, J., & Fei, R. (2019). Discriminative multi-modal deep generative models. *Knowledge-Based Systems*, *173*, 74–82.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., & Smola, A. J. (2007). A kernel method for the two-sample-problem. In *Advances in neural information processing systems* (pp. 513–520).

Guo, W., Wang, J., & Wang, S. (2019). Deep multimodal representation learning: A survey. *IEEE Access*, *7*, 63373–63394.

Hermann, K. M., & Blunsom, P. (2013). Multilingual distributed representations without word alignment. arXiv preprint arXiv:1312.6173.

Hermansky, H., Ellis, D. P., & Sharma, S. (2000). Tandem connectionist feature extraction for conventional HMM systems. In *2000 IEEE international conference on acoustics, speech, and signal processing. proceedings (Cat. No. 00CH37100). Vol. 3* (pp. 1635–1638). IEEE.

Hoffman, M. D., & Johnson, M. J. (2016). Elbo surgery: Yet another way to carve up the variational evidence lower bound. In *Workshop in advances in approximate bayesian inference. Vol. 1* (p. 2).

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*(3/4), 321–377.

Huiskes, M. J., & Lew, M. S. (2008). The MIR flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on multimedia information retrieval* (pp. 39–43).

Javaloy, A., Meghdadi, M., & Valera, I. (2022). Mitigating modality collapse in multimodal VAEs via impartial optimization. arXiv preprint arXiv:2206.04496.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Liu, X., Zhao, J., Sun, S., Liu, H., & Yang, H. (2021). Variational multimodal machine translation with underlying semantic alignment. *Information Fusion*, *69*, 73–80.

Lucas, J., Tucker, G., Grosse, R., & Norouzi, M. (2019). Don't blame the ELBO! a linear VAE perspective on posterior collapse. arXiv:1911.02469. [cs, stat].

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., & Frey, B. (2015). Adversarial autoencoders. arXiv preprint arXiv:1511.05644.

Mancisidor, R. A., Kampffmeyer, M., Aas, K., & Jenssen, R. (2020). Deep generative models for reject inference in credit scoring. *Knowledge-Based Systems*, Article 105758.

Nedelkoski, S., Bogojeski, M., & Kao, O. (2020). Learning more expressive joint distributions in multimodal variational methods. In *International conference on machine learning, optimization, and data science* (pp. 137–149). Springer.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., & Ng, A. Y. (2011). Reading digits in natural images with unsupervised feature learning.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In *ICML*.

Rezaabad, A. L., & Vishwanath, S. (2020). Learning representations by maximizing mutual information in variational autoencoders. In *2020 IEEE international symposium on information theory* (pp. 2729–2734). IEEE.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation*: *Technical report*, California Univ San Diego La Jolla Inst for Cognitive Science.

Salakhutdinov, R., & Hinton, G. (2009). Deep Boltzmann machines. In *Artificial intelligence and statistics* (pp. 448–455).

Shi, Y., Siddharth, N., Paige, B., & Torr, P. (2019). Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Advances in neural information processing systems* (pp. 15718–15729).

Sohn, K., Lee, H., & Yan, X. (2015). Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems* (pp. 3483–3491).

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Srivastava, N., & Salakhutdinov, R. R. (2012). Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems* (pp. 2222–2230).

Sutter, T., Daunhawer, I., & Vogt, J. (2020). Multimodal generative learning utilizing Jensen-Shannon-divergence. *Advances in Neural Information Processing Systems*, *33*, 6100–6110.

Sutter, T. M., Daunhawer, I., & Vogt, J. E. (2021). Generalized multimodal ELBO. In *International conference on learning representations*.

Suzuki, M., Nakayama, K., & Matsuo, Y. (2016). Joint multimodal learning with deep generative models. arXiv preprint arXiv:1611.01891.

Team, T. T. D., Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., Bastien, F., Bayer, J., & Belikov, A. (2016). Theano: A Python framework for fast computation of mathematical expressions. arXiv preprint arXiv:1605.02688.

Theodoridis, T., Chatzis, T., Solachidis, V., Dimitropoulos, K., & Daras, P. (2020). Cross-modal variational alignment of latent spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 960–961).

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11).

Vedantam, R., Fischer, I., Huang, J., & Murphy, K. (2017). Generative models of visually grounded imagination. arXiv preprint arXiv:1705.10762.

Wang, W., Arora, R., Livescu, K., & Bilmes, J. (2015a). On deep multi-view representation learning. In *International conference on machine learning* (pp. 1083–1092).

Wang, W., Arora, R., Livescu, K., & Bilmes, J. A. (2015b). Unsupervised learning of acoustic features via deep canonical correlation analysis. In *2015 IEEE international conference on acoustics, speech and signal processing* (pp. 4590–4594). IEEE.

Wang, W., Yan, X., Lee, H., & Livescu, K. (2017). Deep variational canonical correlation analysis. arXiv preprint arXiv:1610.03454.

Westbury, J. R. (1994). *X-ray microbeam speech production database user's handbook version 1.0*. Madison, WI: Waisman Center on Mental Retardation & Human Development, University of Wisconsin.

Wu, M., & Goodman, N. (2018). Multimodal generative models for scalable weakly-supervised learning. In *Advances in neural information processing systems* (pp. 5575–5585).

Zhao, S., Song, J., & Ermon, S. (2017). Infovae: Information maximizing variational autoencoders. arXiv preprint arXiv:1706.02262.