LEARNING FROM PARTIALLY-OBSERVED MULTIMODAL DATA WITH VARIATIONAL AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning from only partially-observed data for imputation has been an active research area. Despite promising progress on unimodal data imputation (e.g., image in-painting), models designed for multimodal data imputation are far from satisfactory. In this paper, we propose variational selective autoencoders (VSAE) for this task. Different from previous works, our proposed VSAE learns only from partially-observed data. The proposed VSAE is capable of learning the joint distribution of observed and unobserved modalities as well as the imputation mask, resulting in a unified model for various down-stream tasks including data generation and imputation. Evaluation on both synthetic high-dimensional and challenging low-dimensional multi-modality datasets shows significant improvement over the state-of-the-art data imputation models.

1 INTRODUCTION

Learning from data is an integral part of machine learning and artificial intelligence. Modern deep learning techniques rely heavily on extracting information form large scale datasets. While such frameworks have been shown to be effective on various down-stream tasks such as classification, regression, representation learning, and prediction, it is typically crucial to have access to clean and complete training data. Complete data in this case can be either labeled data (for classification), or time-series data with no missing values (for regression), or simply image with no missing pixels (for generation). As such, if a model can only access partially-observed data, the performance will likely be much worse than those trained with fully-observed data, if not completely failing. In practical scenarios, however, it is usually costly to acquire clean and complete data due to the limited human resources and time. Having a model designed to learn and extract information from partially-observed data will not only largely increase the application spectrum of deep learning based models, but also provide benefit to new down-stream tasks, for example, data imputation.

Data imputation with deep generative models has been an active research area (Yoon et al., 2018; Ivanov et al., 2019; Nazabal et al., 2018). Despite promising progress, there are still challenges in learning effective models. First, some prior works focus on learning from fully-observed data while performing imputation on partially-observed data during test phase (Suzuki et al., 2016; Ivanov et al., 2019). Second, they usually have strong assumptions on missingness mechanism (see A.1) such as data is missing completely at random (MCAR) (Yoon et al., 2018). Third, mostly unimodal imputation such as image in-painting has been explored for high-dimensional data (Ivanov et al., 2019; Mattei & Frellsen, 2019). Unimodal data refers to data with only one modality such as image, video, or text. Modeling any combination of data modalities is not well-established yet, which apparently limits the potential of such models, since raw data in real-life is usually acquired in a multimodal manner (Ngiam et al., 2011) with more than one source of data gathered to represent a practical scenario. For example, a video can be uploaded by a user with its corresponding audio and textual descriptions. In practice, one or more of the modalities maybe be missing, leading to a challenging multimodal data imputation task.

In this work, we propose Variational Selective Autoencoder (VSAE) for multimodal data generation and imputation. Our proposed VSAE tries to address the challenges above by learning from partially-observed training data. By constructing an encoder for each modality independently, the latent representation selectively takes only the observed modalities as input, while a set of decoders maps the latent codes to not only *full data* (including both observed and unobserved modalities), but also a

mask representing the missingness scheme. Thus, it can model the joint distribution of the data and the mask together and avoid limiting assumptions such as MCAR. VSAE is optimized efficiently with a single variational objective. In our experimental validation, we evaluate our proposed VSAE on both synthetic high-dimensional multimodal data (MNIST+MNIST bimodal dataset) and challenging low-dimensional tabular data (UCI), and show that VSAE can outperform state-of-the-art baseline models for data imputation task. The contributions are summarized as follows:

(1) A novel framework VSAE to learn from partially-observed multimodal data.

(2) The proposed VSAE is capable of learning the joint distribution of observed and unobserved modalities as well as the imputation mask, resulting in a unified model for various down-stream tasks including data generation and imputation with relaxed assumptions on missingness mechanism.

(3) Evaluation on both synthetic high-dimensional and challenging low-dimensional multimodal datasets shows improvement over the state-of-the-art data imputation models.

2 Related Work

Our work is related to literature on *data imputation* and *multi-modal representation learning*. In this section, we briefly review recent models proposed in these two domains.

Data Imputation. Classical imputation methods such as MICE (Buuren & Groothuis-Oudshoorn, 2010) and MissForest (Stekhoven & Bühlmann, 2011) learn discriminative models to impute missing features from observed ones. With recent advances in deep learning, several deep imputation models have been proposed based on autoencoders (Vincent et al., 2008; Gondara & Wang, 2017; Ivanov et al., 2019), generative adversarial nets (GANs) (Yoon et al., 2018; Li et al., 2019), and autoregressive models (Bachman & Precup, 2015). GAN-based imputation method GAIN proposed by Yoon et al. (2018) assumes that data is missing completely at random. Moreover, this method does not scale to high-dimensional multimodal data. Several VAE based data imputation methods (Ivanov et al., 2019; Nazabal et al., 2018; Mattei & Frellsen, 2019) have been proposed in recent years. Ivanov et al. (2019) formulated variational autoencoders with arbitrary conditioning (VAEAC) for data imputation which allows generation of missing data conditioned on any combination of observed data. This algorithm needs complete data during training cannot learn from partially-observed data only. Nazabal et al. (2018) and Mattei & Frellsen (2019) modified VAE formulation to model the likelihood of the observed data only. However, they require training of a separate generative network for each dimension thereby increasing computational requirements. In contrast, our method aims to model joint distribution of observed and unobserved data along with the missingness pattern (imputation mask). This enables our model to perform both data generation and imputation even under relaxed assumptions on missingness mechanism (see Appendix A.1).

Learning from Multimodal Data. A class of prior works such as conditional VAE (Sohn et al., 2015) and conditional multimodal VAE (Pandey & Dukkipati, 2017) focus on learning the conditional likelihood of the modalities. However, these models requires complete data during training and cannot handle arbitrary conditioning. Alternatively, several generative models aim to model joint distribution of all modalities (Ngiam et al., 2011; Srivastava & Salakhutdinov, 2012; Sohn et al., 2014; Suzuki et al., 2016). However, multimodal VAE based methods such as joint multimodal VAE (Suzuki et al., 2016) and multimodal factorization model (MFM) (Tsai et al., 2019) require complete data during training. On the other hand, Wu & Goodman (2018) proposed another multimodal VAE (namely MVAE) can be trained with incomplete data. This model leverages a shared latent space for all modalities and obtains an approximate joint posterior for the shared space assuming each modalities to be factorized. However, if training data is complete, this model cannot learn the individual inference networks and consequently does not learn to handle missing data during test. Building over multimodal VAE approaches, our model aims to address the shortcomings above within a flexible framework. In particular, our model can learn multimodal representations from partially observed training data and perform data imputation from arbitrary subset of modalities during test. By employing a factorized multimodal representations in the latent space it resembles disentangled models which can train factors specialized for learning from different parts of data (Tsai et al., 2019).

3 METHOD

We introduce a novel VAE-based framework named Variational Selective Autoencoder (VSAE) to learn from partially-observed multimodal data. We first formalize our problem and then provide a detailed description of our model.

3.1 PROBLEM STATEMENT

Let $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2 ..., \mathbf{x}_M]$ be the complete data with M modalities, where $\mathbf{x}_i$ denotes the feature representation for the i-th modality. The size of each $\mathbf{x}_i$ varies and can be very high-dimensional like multimedia data or low-dimensional like tabular data. We define an M-dimensional binary mask variable $\mathbf{m} \in \{0, 1\}^M$ to represent the observed and unobserved modalities: $m_i = 1$ indicates the i-th modality is observed and $m_i = 0$ indicates the i-th modality is unobserved. Thus we have the set of *observed modalities* $\mathbb{O} = \{i | m_i = 1\}$, and the set of *unobserved modalities* $\mathbb{U} = \{i | m_i = 0\}$. $\mathbb{O}$ and $\mathbb{U}$ are complementary subsets of all modalities. Accordingly, we denote the data representation for the observed modalities with $\mathbf{x_o} = [\mathbf{x}_i | m_i = 1]$ and the data representation for unobserved modalities with $\mathbf{x_u} = [\mathbf{x}_i | m_i = 0]$. In this paper, we assume the data $\mathbf{x}$ and the mask $\mathbf{m}$ are dependent, and we aim to model the joint distribution of the data and mask together. As a result, the proposed VSAE model has higher capacity and can be used for both data imputation and data/mask generation.

The high-level overview is that the multimodal data is encoded to a latent space which factorizes w.r.t. the modalities. To handle training and test with partially observed data, the variational latent variable of each modality is modeled selectively to choose between a unimodal encoder if the corresponding modality is observed, or a multimodal encoder if the modality is unobserved. Next all the modalities and mask are reconstructed by decoding the aggregated latent codes through decoders.

3.2 BACKGROUND: VARIATIONAL AUTOENCODER

VAE (Kingma & Welling, 2013) is a probabilistic latent variable model to generate a random variable $\mathbf{x}$ from a latent variable $\mathbf{z}$ with a prior distribution $p(\mathbf{z})$ according to the marginalized distribution $p(\mathbf{x}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} p(\mathbf{x}|\mathbf{z}) = \int p(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) d\mathbf{z}$. However, this is intractable to compute, so the likelihood $\log p(\mathbf{x})$ is approximated by variational lower bound (ELBO) $\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x})$:

$$\log p(\mathbf{x}) \geq \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) = \mathbb{E}_{z \sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] - D_{\mathrm{KL}}[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) || p(\mathbf{z})]. \tag{1}$$

In this equation, $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ is a proposal distribution to approximate intractable true posterior $p(\mathbf{z}|\mathbf{x})$ and parameterized by a inference network (a.k.a encoder). $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ is the conditional likelihood parameterized by another generative network (a.k.a decoder). D_{KL} is the Kullback-Leibler (KL) divergence between the prior and the proposal distribution and functions as a regularizer term,

$$D_{\mathrm{KL}}[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) || p(\mathbf{z})] = \mathbb{E}_{z \sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) - \log p(\mathbf{z})]. \tag{2}$$

To train this model $\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x})$ is optimized on all training data with respect to the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. For more details see Appendix A.2.

3.3 PROPOSED MODEL: VARIATIONAL SELECTIVE AUTOENCODER

Our goal is to model the joint distribution of the data $\mathbf{x} = [\mathbf{x_o}, \mathbf{x_u}]$ and the mask $\mathbf{m}$ that is $p(\mathbf{x}, \mathbf{m}) = \int p(\mathbf{x}, \mathbf{m}|\mathbf{z}) p(\mathbf{z}) d\mathbf{z}$. Following the VAE formulation, we construct a proposal distribution $q(\mathbf{z}|\mathbf{x}, \mathbf{m})$ to approximate the intractable true posterior, and derive a lower bound for $\log p(\mathbf{x}, \mathbf{m})$ as:

$$\begin{aligned}
\mathcal{L}_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\theta},\boldsymbol{\epsilon}}(\mathbf{x}, \mathbf{m}) &= \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}|\mathbf{x},\mathbf{m})}[\log p_{\boldsymbol{\theta},\boldsymbol{\epsilon}}(\mathbf{x}, \mathbf{m}|\mathbf{z})] - D_{\mathrm{KL}}[q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}|\mathbf{x}, \mathbf{m}) || p(\mathbf{z})] \\
&= \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}|\mathbf{x},\mathbf{m})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{m}, \mathbf{z}) + \log p_{\boldsymbol{\epsilon}}(\mathbf{m}|\mathbf{z}) - \log q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}|\mathbf{x}, \mathbf{m}) + \log p(\mathbf{z})].
\end{aligned} \tag{3}$$

We assume the variational latent variables can be factorized with respect to the modalities $\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_M]$, which is a standard assumption for multimodal data (Tsai et al., 2019):

$$p(\mathbf{z}) = \prod_{i=1}^{M} p(\mathbf{z}_i), \qquad \log p(\mathbf{z}) = \sum_{i=1}^{M} \log p(\mathbf{z}_i), \tag{4}$$

$$q(\mathbf{z}|\mathbf{x}, \mathbf{m}) = \prod_{i=1}^{M} q(\mathbf{z}_i|\mathbf{x}, \mathbf{m}), \qquad \log q(\mathbf{z}|\mathbf{x}, \mathbf{m}) = \sum_{i=1}^{M} \log q(\mathbf{z}_i|\mathbf{x}, \mathbf{m}). \tag{5}$$

Given this, we define the proposal distribution for each modality as

$$q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}_i|\mathbf{x}, \mathbf{m}) = \begin{cases} q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathbf{x}_i) & \text{if } m_i = 1 \\ q_{\boldsymbol{\psi}}(\mathbf{z}_i|\mathbf{x}_\mathbf{o}, \mathbf{m}) & \text{if } m_i = 0 \end{cases} \tag{6}$$

This is based on the intuitive assumption that the latent space of each modality is independent of other modalities given its data is observed. But, if the data is missing for some modality, its latent space is constructed from the other observed modalities. We call this *selective proposal distribution*.

In the decoder, the probability distribution also factorizes over the modalities assuming that the reconstructions are conditionally independent given the complete latent variables of all modalities:

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{m}, \mathbf{z}) = \log p_{\boldsymbol{\theta}}(\mathbf{x}_\mathbf{o}, \mathbf{x}_\mathbf{u}|\mathbf{m}, \mathbf{z}) = \sum_{i \in \mathbb{O}} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{m}, \mathbf{z}) + \sum_{j \in \mathbb{U}} \log p_{\boldsymbol{\theta}}(\mathbf{x}_j|\mathbf{m}, \mathbf{z}) \tag{7}$$

Putting all this together, the ELBO in Equation 3 can be rewritten as

$$\begin{aligned} \mathcal{L}_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\theta},\boldsymbol{\epsilon}}(\mathbf{x}_\mathbf{o}, \mathbf{x}_\mathbf{u}, \mathbf{m}) =& \mathbb{E}_\mathbf{z}\left[\sum_{i \in \mathbb{O}} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{m}, \mathbf{z}) + \sum_{j \in \mathbb{U}} \log p_{\boldsymbol{\theta}}(\mathbf{x}_j|\mathbf{m}, \mathbf{z})\right] + \mathbb{E}_\mathbf{z}[\log p_{\boldsymbol{\epsilon}}(\mathbf{m}|\mathbf{z})] \\ & - \sum_{i=1}^{M} \mathbb{E}_{\mathbf{z}_i}[\log q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}_i|\mathbf{x}, \mathbf{m}) - \log p(\mathbf{z}_i)], \end{aligned} \tag{8}$$

where $\mathbf{z}_i \sim q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}_i|\mathbf{x}, \mathbf{m})$ according to the selective proposal distribution given in Equation 6.

In order to learn the model, the ELBO should be maximized over training data. However under partially-observed setting, $\mathbf{x}_\mathbf{u}$ is missing and unavailable even during training. Thus, we define the objective function for training by taking expectation over $\mathbf{x}_\mathbf{u}$

$$\mathcal{L}'_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\theta},\boldsymbol{\epsilon}}(\mathbf{x}_\mathbf{o}, \mathbf{m}) = \mathbb{E}_{\mathbf{x}_\mathbf{u}}[\mathcal{L}_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\theta},\boldsymbol{\epsilon}}(\mathbf{x}_\mathbf{o}, \mathbf{x}_\mathbf{u}, \mathbf{m})] \tag{9}$$

Only one term in Equation 8 is dependent on $\mathbf{x}_\mathbf{u}$, so the final objective function is obtained as

$$\begin{aligned} \mathcal{L}'_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\theta},\boldsymbol{\epsilon}}(\mathbf{x}_\mathbf{o}, \mathbf{m}) =& \mathbb{E}_\mathbf{z}\left[\sum_{i \in \mathbb{O}} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{m}, \mathbf{z}) + \sum_{j \in \mathbb{U}} \mathbb{E}_{\mathbf{x}_j}[\log p_{\boldsymbol{\theta}}(\mathbf{x}_j|\mathbf{m}, \mathbf{z})]\right] + \mathbb{E}_\mathbf{z}[\log p_{\boldsymbol{\theta}}(\mathbf{m}|\mathbf{z})] \\ & - \sum_{i=1}^{M} \mathbb{E}_{\mathbf{z}_i}[\log q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}_i|\mathbf{x}, \mathbf{m}) - \log p(\mathbf{z}_i)], \text{where } \mathbf{z}_i \sim q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}_i|\mathbf{x}, \mathbf{m}) \end{aligned} \tag{10}$$

In the proposed algorithm, we approximate $\mathbb{E}_{\mathbf{x}_j}[\log p_{\boldsymbol{\theta}}(\mathbf{x}_j|\mathbf{m}, \mathbf{z})], j \in \mathbb{U}$ by sampling from the prior network. Our experiments show that even a single sample is sufficient to learn the model effectively. In fact, the prior network can be used as a self supervision mechanism to find the most likely samples which dominate the other samples when taking the expectation.

In Equation 10, $p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{m}, \mathbf{z})$ is the decoding term of corresponding modality $\mathbf{x}_i$ and the type of distribution depends on the data type. For the mask decoding term $p_{\boldsymbol{\theta}}(\mathbf{m}|\mathbf{z})$ we consistently use Bernoulli distribution to model the binary mask vector. Also for the prior distribution we simply use standard normal distribution $p(\mathbf{z}) = \prod_{i=1}^{M} p(\mathbf{z}_i) = \prod_{i=1}^{M} \mathcal{N}(\mathbf{z}_i; \mathbf{0}, \mathbf{I})$ which is fully-factorized.

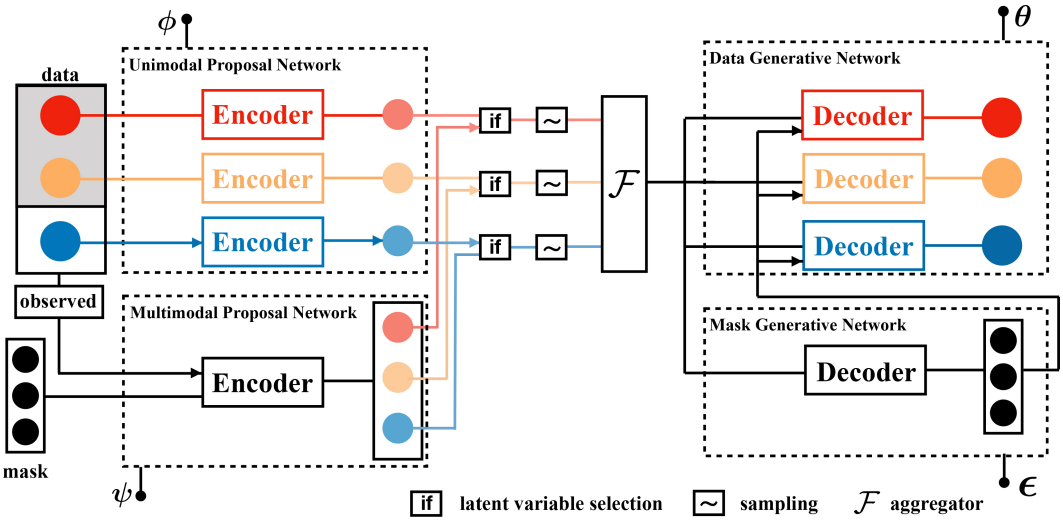

Figure 1: **Overall architecture.** The unimodal proposal network and multimodal proposal network are employed by selection. Different modalities are denoted by different colors. Unobserved modalities are shaded. (i.e. blue is observed while red/yellow are unobserved.) The selected variables are indicated by the arrows. Standard normal prior is not plotted for simplicity. ϕ, ψ, θ and ϵ are the parameters of unimodal proposal network, multimodal proposal network, data generative network and mask generative network respectively. All components are trained jointly.

3.4 Network Modules

We construct each module of our model using neural networks and optimize the parameters via backpropagation techniques. Following the terms in standard VAE, our model is composed of encoders and decoders. The architecture is shown in Figure 1 with different modalities denoted by different colors. The data space of unobserved modalities is shaded to differentiate from observed modalities. The whole architecture can be viewed as an integration of two auto-encoding structures: the top-branch data-wise encoders/decoders and the bottom-branch mask-wise encoders/decoder. The selective proposal distribution chooses between the unimodal and multimodal encoders if the data is observed or not. The outputs of all encoders are aggregated and a common latent space is shared among all decoders. In the rest of this section we explain different modules in the proposed model. For more details about architecture and implementation see Appendix B.

Selective Factorized Encoders Standard proposal distribution of VAEs depends on the whole data and can not handle incomplete input when the data is partially-observed. To overcome this, we introduce our selective proposal distribution, which is factorized with respect to the modalities. As defined in Equation 6, $q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathbf{x}_i)$, named as the *unimodal proposal distribution*, is inferred only from each observed individual modality of data. However, if the modality is unobserved, the *multimodal proposal distribution* $q_{\boldsymbol{\psi}}(\mathbf{z}_i|\mathbf{x}_{\mathbf{o}}, \mathbf{m})$ is used to infer corresponding latent variables from other observed modalities and mask. Hence, the learned model can impute the missing information by combining unimodal proposal distribution of observed modalities and multimodal proposal distribution of the unobserved modalities. The condition on the mask could make the model aware of the missing pattern and could help the model to attend to observed modalities.

For each modality $\mathbf{x}_i$, we have a separate encoder to infer its unimodal proposal distribution parameterized by $\boldsymbol{\phi}$. For the multimodal proposal distribution, however, we use a single encoder parameterized by $\boldsymbol{\psi}$. This encoder outputs the latent codes for all modalities, and we simply obtain the latent variable for each modality by slicing the output vector to M sequential vectors. We simply model all the proposal distributions as normal distributions by setting the outputs of all encoders as mean and variance of a normal distribution. For the unimodal proposal distributions, we have $q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathbf{x}_i) = \mathcal{N}(\mathbf{z}_i; \boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}_i), \boldsymbol{\Sigma}_{\boldsymbol{\phi}}(\mathbf{x}_i))$, where $\boldsymbol{\mu}_{\boldsymbol{\phi}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}$ are deterministic neural networks parameterized by $\boldsymbol{\phi}$ that output the mean and covariance respectively. Similarly, the multimodal proposal

	Categorical		Numerical	
	Phishing	Mushroom	Yeast	Glass
AE	0.348 ± 0.002	0.556 ± 0.009	0.737 ± 0.036	1.651 ± 0.049
VAE	0.293 ± 0.003	0.470 ± 0.017	0.461 ± 0.001	1.409 ± 0.011
CVAE w/ mask	0.241 ± 0.003	0.445 ± 0.004	0.449 ± 0.001	1.498 ± 0.0013
MVAE	0.308 ± 0.015	0.586 ± 0.019	0.442 ± 0.018	1.572 ± 0.035
VSAE (ours)	$\mathbf{0.237 \pm 0.001}$	$\mathbf{0.396 \pm 0.008}$	$\mathbf{0.409 \pm 0.012}$	$\mathbf{1.312 \pm 0.021}$
CVAE w/ data	0.301 ± 0.005	0.485 ± 0.034	0.449 ± 0.0001	1.380 ± 0.045
VAEAC	0.240 ± 0.006	0.403 ± 0.006	0.447 ± 0.0016	1.432 ± 0.027

Table 1: **Feature Imputation on UCI datasets.** Missing ratio is 0.5. Categorical and numerical datasets are respectively evaluated by PFC and NRMSE. Last two rows are trained with fully-observed data, potentially serving as an upper bound for imputation models. We show mean and standard deviation over 3 independent runs. For both lower is better.

distribution $q_{\psi}(\mathbf{z}_i|\mathbf{x_o}, \mathbf{m}) = \mathcal{N}(\mathbf{z}_i; \boldsymbol{\mu}_{\psi}(\mathbf{x_o}, \mathbf{m}), \boldsymbol{\Sigma}_{\psi}(\mathbf{x_o}, \mathbf{m}))$ can be modeled by a neural network with $\mathbf{x_o}$ and $\mathbf{m}$ as the inputs. The reparameterization in standard VAE is used for end-to-end training.

Decoding through Latent Variable Aggregator $\mathcal{F}$ After selecting and sampling from proper proposal distributions for all modalities, the variational latent codes can be fed to the downstream decoders even when the observation is incomplete. To do this, the information from different modalities interact by aggregating their stochastic latent codes before going through the decoders:

$$p_{\boldsymbol{\epsilon}}(\mathbf{m}|\mathbf{z}) = p_{\boldsymbol{\epsilon}}(\mathbf{m}|\mathcal{F}(\mathbf{z})),$$
$$p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{z}, \mathbf{m}) = p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathcal{F}(\mathbf{z}), \mathbf{m})), \tag{11}$$

Here we simply choose the aggregator $\mathcal{F}(\cdot) = \text{concat}(\cdot)$, i.e., concatenating the latent codes as one single vector. One may also use other aggregation functions such as max/mean pooling or matrix fusion (Veit et al., 2018) to combine latent codes from all modalities. The decoders take the shared aggregated variational latent codes as input to generate data and mask.

Mask Vector Encoding and Decoding The mask variable $\mathbf{m}$ is encoded into the latent space through the multimodal proposal network. The latent space is shared by the mask and data decoders. The mask decoder is an MLP parameterized by $\boldsymbol{\epsilon}$ in our implementation. It maps the aggregated latent codes from the selective proposal distributions to a reconstruction of the M-dimensional binary mask vector. We assume each dimension of the mask variable is an independent Bernoulli distribution.

Training With reparameterization trick (Kingma & Welling, 2013), we can jointly optimize the objective derived in Equation 10 with respect to these parameters defined above on training set:

$$\max_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\epsilon}} \mathbb{E}_{\mathbf{x_o},\mathbf{m}}[\mathcal{L}'_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\epsilon}}(\mathbf{x_o}, \mathbf{m})] \tag{12}$$

Since Equation 12 only requires the mask and observed data during training, this modified ELBO $\mathcal{L}'_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\epsilon}}(\mathbf{x_o}, \mathbf{m})$ can be optimized without the presence of unobserved modalities. The KL-divergence term is calculated analytically for each factorized term. The conditional log-likelihood term is computed by negating reconstruction loss function. (See Section 4 and Appendix B.2.)

Inference The learned model can be used for both data imputation and generation. For imputation, the observed modalities $\mathbf{x_o}$ and mask $\mathbf{m}$ are fed through the encoders to infer the selective proposal distributions. Then the sampled latent codes are decoded to estimate the unobserved modalities $\mathbf{x_u}$. All the modules in Figure 1 are used for imputation. For generation, since no data is available at all, the latent codes are sampled from the prior and go through the decoders to generate the data and the mask. In this way, only modules after the aggregator are used without any inference modules.

4 Experiment

To demonstrate the effectiveness of our model, we evaluate our model on high-dimensional multi-modal data and low-dimensional tabular data. We provide an extensive comparison with state-of-the-

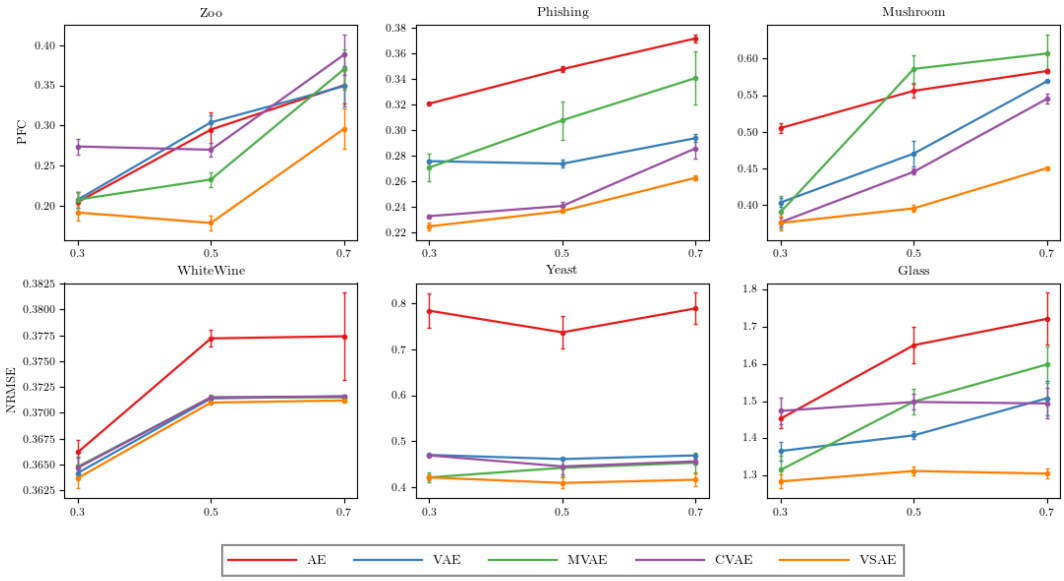

Figure 2: **Feature Imputations on UCI datasets.** Missing ratios is 0.3, 0.5, 0.7. Categorical (top row) and numerical (bottom row) datasets are evaluated by PFC and NRMSE respectively (lower is better for both). We show mean and standard deviation over 3 independent runs.

art latent variable models. To test the robustness of our model, we evaluate our model under various challenging missingness mechanisms.

Baselines. There are two categories of deep latent variable models. First, models that observe the masked information during training phase and learn the conditional probability of unobserved modalities given observed modalities $p(\mathbf{x_u}|\mathbf{x_o})$. Intuitively, this serves as an upper bound of the imputation model since it has complete access to the full data. In this category, we report the results of VAEAC (Ivanov et al., 2019) and conditional VAE (Sohn et al., 2015). Second, models that only partially observe the data during both training and test phase. These models can only observe partial modalities $\mathbf{x_o}$. Here, distribution of $\mathbf{x_u}$ can be approximated from $\mathbf{x_o}$ based on the assumption that $\mathbf{x_u}$ and $\mathbf{x_o}$ are sampled from different regions of the same space modeling the data. Models that belong to this category can learn from partially-observed data without any requirement of full data at any stage. Our model VSAE falls in this category since it can learn the joint distribution of $p(\mathbf{x_o}, \mathbf{x_u}, \mathbf{m})$ given observed information solely. We use deterministic Autoencoder (AE), VAE (Kingma & Welling, 2013), conditional VAE (Sohn et al., 2015) conditional on mask and MVAE (Wu & Goodman, 2018) as baselines and compare our model with these baselines. For fair comparison, we compare VSAE with models trained only with $\mathbf{x_o}$, but also report numbers from models trained with $\mathbf{x_u}$ and $\mathbf{x_o}$. Additional information about experimental details can be found in Appendix. B.

4.1 Low-dimensional Tabular Data Imputation

We choose UCI repository datasets to demonstrate the effectiveness of our model on tabular data. It contains different tabular datasets with numerical or categorical variables. In our experiments, we randomly sample from independent Bernoulli distributions with pre-defined missing ratio to simulate the masking mechanism. All masks for data points are fixed during training and test. We use min-max normalization to preprocess the numerical data and replace the unobserved dimensions by standard normal noise. Training/test set split is $80\%/20\%$. We also split 20% of training set as validation set to choose the best model. We use Mean Square Error, Cross Entropy and Binary Cross Entropy as reconstruction loss for numerical and categorical variables and mask variables respectively. We report the standard measures: **NRMSE** (i.e. RMSE normalized by the standard deviation of the feature and averaged over all features) for numerical datasets and **PFC** (i.e. proportion of falsely classified attributes of each feature and averaged over all features) for categorical datasets.

	α-MSE/784	β-MSE/784	Bimodal Error
AE	0.1077 ± 0.0002	0.1070 ± 0.0008	0.2147 ± 0.0008
VAE	0.0734 ± 0.0002	0.0682 ± 0.0001	0.1396 ± 0.0002
CVAE w/ mask	0.0733 ± 0.0004	0.0679 ± 0.0003	0.1412 ± 0.0007
MVAE	0.0760 ± 0.0003	0.0802 ± 0.0005	0.1562 ± 0.0003
VSAE	$\mathbf{0.0712 \pm 0.0001}$	$\mathbf{0.0663 \pm 0.0001}$	$\mathbf{0.1376 \pm 0.0002}$
CVAE w/ data	0.0694 ± 0.0001	0.0646 ± 0.0003	0.1340 ± 0.0003
VAEAC	0.0693 ± 0.0003	0.0645 ± 0.0001	0.1338 ± 0.0003

Table 2: **Imputation on MNIST+MNIST.**. Missing ratio is 0.5. α and β denote two modalities. Last two rows are trained with fully-observed data. We show mean and standard deviation over 3 independent runs (lower is better).

Results and Analysis. Table 1 indicates it outperforms other methods on both numerical and categorical data. The first five rows are trained in partially-observed setting, while the last two trained with fully-observed data. We observe that models trained with partially-observed data can outperform those models trained with fully-observed data on some datasets. We argue this is due to two potential reasons: (1) the mask provides a natural way of dropout on the data space, thereby, helping the model to generalize; (2) If the data is noisy or has outliers, which is common in low-dimensional data, learning from partially-observed data can improve the results by ignoring these data. However, we do not claim our model is the state of the art for fully-observed data imputation and these models potentially serve as upper bound if the data is clean.

Figure 2 illustrates that our model generally has lower error with lower variance for all missing ratios. With higher missing ratio (more data is unobserved), our model achieves more stable imputation performance on most of the datasets. On the contrary, there is a performance drop along with higher variance in the case of baselines. We believe this is becaause of the proposal distribution based encoder selection in VSAE. When the missing ratio goes up, the input to unimodal encoders stays same while other encoders have to learn to focus on the useful information in data.

4.2 High-dimensional Multimodal Data Imputation

We synthesize two bimodal datasets using MNIST and SVHN datasets. MNIST contains 28-by-28 gray images (0-9 handwritten digits) with training/test of 60000/10000 samples; SVHN contains 32-by-32 RGB images (0-9 house numbers) with training/test of 73257/26032 samples. We synthesize our datasets as follows.

MNIST+MNIST bimodal dataset: We randomly pair two digits in MNIST as [0, 9], [1, 8], [2, 7], [3, 6], [4, 5]. The training/test/validation sets respectively contain 23257/4832/5814 samples.

MNIST+SVHN bimodal dataset: We pair one digit in MNIST with the random same digit in SVHN. The training/test/validation sets respectively contain 44854/10000/11214 samples. For both datasets, we synthesize mask vectors over each modality by sampling from Bernoulli distribution. All mask are fixed after synthesis process. All original data points are only used once.

Results and Analysis. For MNIST+MNIST, our model has better imputation results in both modalities (refer to Table 2). With masks sampled with different missing ratios, the combined errors (i.e. sum of MSE in each modality averaged over its dimensions) of our model are 0.1371 ± 0.0001, 0.1376 ± 0.0002 and 0.1379 ± 0.0001 under missing ratio of 0.3, 0.5 and 0.7 (Additional results are in Appendix C.2). This indicates that our VSAE is robust under different missing ratios, whereas other baselines are sensitive to the missing ratio. Figure 3 presents the qualitative results of imputations on MNIST+MNIST in Figure 3. For experimental results on MNIST+SVHN, see Appendix C. VSAE has better performance for imputation task on all modalities with lower variance (see Table 2). We believe this is because of the underlying mechanism of proper proposal distribution selection and sharing priors. The separate structure of unimodal/multimodal encoders helps the model to attend to the observed data. It limits the input of unimodal encoders to observed single modality. Thus it is more robust to the missingness. On the other hand, baseline methods have only single proposal distribution inferred from the whole input. Furthermore, VSAE can easily ignore

Figure 3: **Imputation on MNIST+MNIST.** Top row visualizes observed modality, middle row unobserved modality, and bottom row shows the imputation of unobserved modality from VSAE.

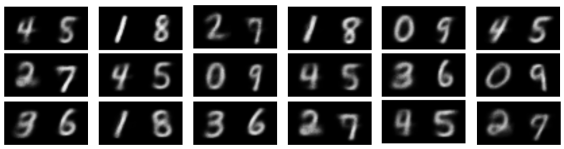

Figure 4: **Generation on MNIST+MNIST.** Generated Samples w/o conditional information. As shown, the correspondence between modalities (predefined pairs) are preserved while stochastic multimodal generation.

unobserved noisy modalities and attends on observable useful modalities, while baselines methods rely on neural networks to learn to useful information from the whole data (which is dominated by missing information in case of high missing ratio).

For partially-observed training setting, unobserved data is not available even during training. However, the unobserved modality in one data sample could be the observed modality in another data sample. Therefore, over the whole training set, the multimodal encoders are able to construct the mapping from observable information to unobservable information. Multimodal encoders also include the mask vector as input. This allows the multimodal encoders to be aware of the shape of the missingness and forces it to focus on the useful information in the observed modalities.

Figure 3 shows that the model can impute the unobserved modality preserving the rule of pre-defined pairs. This is done by the late aggregator to fuse the information of selected unimodal encoders (for observed information) and multimodal encoders (for unobserved information).

4.3 Data and Mask Generation

Unlike conventional methods that model $p(\mathbf{x_u}|\mathbf{x_o})$, our method is to model the joint probability $p(\mathbf{x_o}, \mathbf{x_u}, \mathbf{m})$. Thus our model can impute missing features and also generate data and mask from scratch. Figure 4 shows the model learns the correlation between different modalities to pair the digits as predefined in the dataset without giving any labels in partially-observed setting. Our proposed VSAE has a mask conditional log-likelihood term is included in the objective ELBO. This term allows the latent space to have information from mask variables and be able to reconstruct (or generate if sample the prior) the mask vector.

In UCI repository experiments, the mask variable follows Bernoulli distribution. After training, we sample from the prior to generate the mask. We calculate the proportion of the unobserved dimensions in generated mask vectors (averaged over 100 samples of the output). Averaging over this proportion on all datasets, we get 0.3123 ± 0.026, 0.4964 ± 0.005, 0.6927 ± 0.013 for missing ratio of 0.3, 0.5, 0.7. It indicates that our model can learn the mask distribution. We also observe that conditions on the reconstructed mask vector in the data decoders improve the performance. We believe this is because the mask vector can inform the data decoder about the missingness in the data space since the latent space is shared by both observed and unobserved modalities thereby allowing it to generate data from the selective proposal distribution.

5 Conclusion

In this paper, we propose a VAE framework to learn from partially-observed data. Learning from partially-observed data is important but previous deep latent variable models cannot work well on this problem. The proposed model differentiates the observed and unobserved information by selecting a proper proposal distribution. The experimental results show the model can consistently outperform other baselines on low-dimensional tabular data and high-dimensional multimodal data. The model also can generate data with mask directly from prior without any conditions.

REFERENCES

Philip Bachman and Doina Precup. Data generation as sequential decision making. In *Advances in Neural Information Processing Systems*, pp. 3249–3257, 2015.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *International Conference on Learning Representations*, 2016.

S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pp. 1–68, 2010.

Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 2018.

Lluis Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Learning aligned cross-modal representations from weakly aligned data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Hong-Min Chu, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Deep generative models for weakly-supervised multi-label classification. In *The European Conference on Computer Vision (ECCV)*, September 2018.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 2016.

Lovedeep Gondara and Ke Wang. Multiple imputation using deep denoising autoencoders. *arXiv preprint arXiv:1705.02737*, 2017.

Jiawei He, Yu Gong, Joseph Marino, Greg Mori, and Andreas Lehrmann. Variational autoencoders with jointly optimized latent dependency structure. In *International Conference on Learning Representations*, 2019.

Oleg Ivanov, Michael Figurnov, and Dmitry Vetrov. Variational autoencoder with arbitrary conditioning. In *International Conference on Machine Learning*, 2019.

Vikas Jain, Nirbhay Modhe, and Piyush Rai. Scalable generative models for multi-label learning with missing labels. In *Proceedings of the 34th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2017.

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1857–1865. JMLR. org, 2017.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

Diederik P. Kingma, Danilo Jimenez Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models. *CoRR*, abs/1406.5298, 2014.

Steven Cheng-Xian Li, Bo Jiang, and Benjamin Marlin. Misgan: Learning from incomplete data with generative adversarial networks. In *International Conference on Learning Representations*, 2019.

Roderick J A Little and Donald B Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., New York, NY, USA, 1986. ISBN 0-471-80254-9.

Pierre-Alexandre Mattei and Jes Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International Conference on Machine Learning*, pp. 4413–4423, 2019.

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3994–4003, 2016.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pp. 169–176. ACM, 2011.

Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using vaes. *arXiv preprint arXiv:1807.03653*, 2018.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *International Conference on Machine Learning*, ICML'11, pp. 689–696, USA, 2011. Omnipress. ISBN 978-1-4503-0619-5.

Gaurav Pandey and Ambedkar Dukkipati. Variational methods for conditional multimodal deep learning. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 308–315. IEEE, 2017.

Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Casper Kaae Sø nderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sø nderby, and Ole Winther. Ladder variational autoencoders. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 3738–3746. Curran Associates, Inc., 2016.

Kihyuk Sohn, Wenling Shang, and Honglak Lee. Improved multimodal deep learning with variation of information. In *Advances in neural information processing systems*, pp. 2141–2149, 2014.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems 28*, pp. 3483–3491. 2015.

Yuhang Song, Chao Yang, Zhe Lin, Xiaofeng Liu, Qin Huang, Hao Li, and C.-C. Jay Kuo. Contextual-based image inpainting: Infer, match, and translate. In *The European Conference on Computer Vision (ECCV)*, September 2018.

Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pp. 2222–2230, 2012.

Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2011.

Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *International Conference on Learning Representations*, 2019.

Andreas Veit, Maximilian Nickel, Serge Belongie, and Laurens van der Maaten. Separating self-expression and visual content in hashtag supervision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *international conference on Machine learning*, pp. 1096–1103. ACM, 2008.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 2018.

Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Gain: Missing data imputation using generative adversarial nets. In *International Conference on Machine Learning*, 2018.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114. Association for Computational Linguistics, 2017. doi: 10.18653/v1/D17-1115.

Amir Zadeh, Yao-Chong Lim, Paul Pu Liang, and Louis-Philippe Morency. Variational auto-decoder: Neural generative modeling from partial data. *arXiv preprint arXiv:1903.00840*, 2019.

Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *The European Conference on Computer Vision (ECCV)*, September 2018.

Yu Zhang, Ying Wei, and Qiang Yang. Learning to multitask. In *Advances in Neural Information Processing Systems*, pp. 5771–5782, 2018.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2017.

A BACKGROUND

A.1 IMPUTATION PROCESS AND MISSINGNESS MECHANISMS

Following (Little & Rubin, 1986), the imputation process is to learn a generative distribution for unobserved missing data. To be consistent with notations in Section **??**, let $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M]$ be the complete data of all modalities where $\mathbf{x}_i$ denote the feature representation for the i-th modality. We also define $\mathbf{m} \in \{0, 1\}^M$ as the binary mask vector, where $m_i = 1$ indicates if the i-th modality is observed, and $m_i = 0$ indicates if it is unobsrved:

$$\begin{aligned} \mathbf{x} &\sim p_{\text{data}}(\mathbf{x}), \\ \mathbf{m} &\sim p(\mathbf{m}|\mathbf{x}). \end{aligned} \tag{13}$$

Given this, the observed data $\mathbf{x_o}$ and unobserved data $\mathbf{x_u}$ are represented accordingly:

$$\begin{aligned} \mathbf{x_o} &= [\mathbf{x}_i|m_i = 1], \\ \mathbf{x_u} &= [\mathbf{x}_i|m_i = 0]. \end{aligned} \tag{14}$$

In the standard maximum likelihood setting, the unknown parameters are estimated by maximizing the following marginal likelihood, integrating over the unknown missing data values:

$$p(\mathbf{x_o}, \mathbf{m}) = \int p(\mathbf{x_o}, \mathbf{x_u}, \mathbf{m}) d\mathbf{x_u} = \int p(\mathbf{x_o}, \mathbf{x_u}) p(\mathbf{m}|\mathbf{x_o}, \mathbf{x_u}) d\mathbf{x_u} \tag{15}$$

Little & Rubin (1986) characterizes the missingness mechanism $p(\mathbf{m}|\mathbf{x_o}, \mathbf{x_u})$ in terms of independence relations between the complete data $\mathbf{x} = [\mathbf{x_o}, \mathbf{x_u}]$ and the mask $\mathbf{m}$:

- Missing completely at random (MCAR): $p(\mathbf{m}|\mathbf{x_o}, \mathbf{x_u}) = p(\mathbf{m})$,
- Missing at random (MAR): $p(\mathbf{m}|\mathbf{x_o}, \mathbf{x_u}) = p(\mathbf{m}|\mathbf{x_o})$,
- Not missing at random (NMAR): $p(\mathbf{m}|\mathbf{x_o}, \mathbf{x_u}) = p(\mathbf{m}|\mathbf{x_u})$ or $p(\mathbf{m}|\mathbf{x_o}, \mathbf{x_u})$.

Most previous data imputation methods works under MCAR or MAR assumptions since $p(\mathbf{x_o}, \mathbf{m})$ can be factorized into $p(\mathbf{x_o})p(\mathbf{m}|\mathbf{x_o})$ or $p(\mathbf{x_o})p(\mathbf{m})$. With such decoupling, we do not need missing information to marginalize the likelihood, and it provides a simple but approximate framework to learn from partially-observed data.

A.2 VARIATIONAL AUTOENCODER

Variational Autoencoder (VAE) (Kingma & Welling, 2013) is a probabilistic generative model, where data is constructed from a latent variable $\mathbf{z}$ with a prior distribution $p(\mathbf{z})$. It is composed of an inference network and a generation network to encode and decode data. To model the likelihood of data, the true intractable posterior $p(\mathbf{z}|\mathbf{x})$ is approximated by a proposal distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, and the whole model is trained until ideally the decoded reconstructions from the latent codes sampled from the approximate posterior match the training data. In the generation module, $p_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}|\mathbf{z})$, a decoder realized by a deep neural network parameterized by $\boldsymbol{\theta}$, maps a latent variable $\mathbf{z}$ to the reconstruction $\tilde{\mathbf{x}}$ of observation $\mathbf{x}$. In the inference module, an encoder parameterized by $\boldsymbol{\phi}$ produces the sufficient statistics of the approximation posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ (a known density family where sampling can be readily done). In vanilla VAE setting, by simplifying approximate posterior as a parameterized diagonal normal distribution and prior as a standard diagonal normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, the training criterion is to maximize the following evidence lower bound (ELBO) w.r.t. $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$.

$$\log p(\mathbf{x}) \geq \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] - D_{\text{KL}}[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] \tag{16}$$

where D_{KL} denotes the Kullback-Leibler (KL) divergence. Usually the prior $p(\mathbf{z})$ and the approximate $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ are chosen to be in simple form, such as a Gaussian distribution with diagonal covariance, which allows for an analytic calculation of the KL divergence. While VAE approximates $p(\mathbf{x})$, conditional VAE (Sohn et al., 2015) approximates the conditional distribution $p(\mathbf{x}|\mathbf{y})$. By simply introducing a conditional input, CVAE is trained to maximize the following ELBO:

$$\log p(\mathbf{x}|\mathbf{y}) \geq \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{y})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}, \mathbf{y})] - D_{\text{KL}}[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{y})||p_{\boldsymbol{\psi}}(\mathbf{z}|\mathbf{y})] \tag{17}$$

B IMPLEMENTATION DETAILS

B.1 ARCHITECTURE

In all models, all the layers are modeled by MLP without any skip connections or resnet modules. Basically, the unimodal encoders take single modality data vector as input to infer the unimodal proposal distribution; the multimodal encoders take the observed data vectors and mask vector as as input to infer the multimodal proposal distributions. The input vector to multimodal encoders should have same length for the neural network. Here we just concatenate all modality vectors and replace the unobserved modality vectors with some noise. In UCI repository experiment, we replace the unobserved modality vectors as standard normal noise. In Bimodal experiment, we simply replace the pixels of unobserved modality as zero. Note that all the baselines has encoders/decoders with same or larger number of parameters than our method. We implement our model using PyTorch.

Unimodal Encoders In UCI repository experiment, the unimodal encoders for numerical data are modeled by 3-layer 64-dim MLPs and the unimodal encoders for categorical data are modeled by 3-layer 64-dim MLPs, all followed by Batch Normalization and Leaky ReLU nonlinear activations. In MNIST+MNIST bimodal experiment, the unimodal encoders are modeled by 3-layer 128-dim MLPs followed by Leaky ReLU nonlinear activations; In MNIST+SVHN bimodal experiment, the unimodal encoders are modeled by 3-layer 512-dim MLPs followed by Leaky ReLU nonlinear activations. We set the latent dimension as 20-dim for every modality in UCI repository experiments and 256-dim for every modality in Bimodal experiments.
UCI data unimodal encoder: Linear(1, 64)→ BatchNorm1d(64)→ LeakyReLU→ Linear(64, 64)→ LeakyReLU→ Linear(64, 64)→ LeakyReLU→ Linear(64, 20);
MNIST+MNIST synthetic unimodal encoder: Linear(data-dimension, 128)→ LeakyReLU→ Linear(128,128)→ LeakyReLU→ Linear(128, 128)→ LeakyReLU→ Linear(128, 256);
MNIST+SVHN synthetic unimodal encoder: Linear(data-dimension, 512)→ LeakyReLU→ Linear(512,512)→ LeakyReLU→ Linear(512, 512)→ LeakyReLU→ Linear(512, 256);

Multimodal Encoders In general, any model capable of multimodal fusion (Zadeh et al., 2017; Morency et al., 2011) can be used here to map the observed data $\mathbf{x_o}$ and the mask $\mathbf{m}$ to the latent variables $\mathbf{z}$. However, in this paper we simply use an architecture similar to unimodal encoders. The difference is that the input to unimodal encoders are lower dimensional vectors of an individual modalities. But, the input to the multimodal encoders is the complete data vector with unobserved modalities replaced with noise or zeros. As the input to the multimodal encoders is the same for all modalities (i.e., $q(\mathbf{z}_i|\mathbf{x_o}, \mathbf{m})\ \forall i$), we model the multimodal encoders as one single encoder to take advantage of the parallel matrix calculation speed. Thus the multimodal encoder for every experiment has the same structure as its unmidal encoder but with full-dimensional input.

Aggregator In our models, we simply use vector concatenation as the way of aggregating.

Mask Decoder UCI mask decoder: Linear(20*data-dimension, 64)→ BatchNorm1d(64)→ LeakyReLU→ Linear(64, 64)→ LeakyReLU→ Linear(64, 64)→ LeakyReLU→ Linear(64, mask-dimension)→Sigmoid;
MNIST+MNIST synthetic mask decoder: Linear(512, 16)→ BatchNorm1d(16)→ LeakyReLU→ Linear(16,16)→ LeakyReLU→ Linear(16, 16)→ LeakyReLU→ Linear(16, 2)→Sigmoid;
MNIST+SVHN synthetic mask encoder: Linear(512, 16)→ BatchNorm1d(16)→ LeakyReLU→ Linear(16,16)→ LeakyReLU→ Linear(16,16)→ LeakyReLU→ Linear(16,2)→Sigmoid;

Data Decoder As the output is factorized over modalities and for every decoder the input is shared as the latent codes sampled from the selective proposal distribution. We implement all the decoders of the data modalities as one single decoder for parallel speed. UCI data decoder: Linear(20*data-dimension, 128)→ BatchNorm1d(128)→ LeakyReLU→ Linear(128)→ Linear(128, 128)→ Linear(128, data-dimension);
MNIST+MNIST synthetic data decoder: Linear(512, 128)→ BatchNorm1d(128)→ LeakyReLU→ Linear(128,128)→ Linear(128, 128)→ Linear(128, 784)→Sigmoid;
MNIST+SVHN synthetic mask encoder: Linear(512, 512)→ BatchNorm1d(512)→ LeakyReLU→ Linear(512,512)→ Linear(512,512)→ Linear(512,784/3072)→Sigmoid;

B.2 TRAINING

We use Adam optimizer for all models. For UCI numerical experiment, learning rate is 1e-3 and use validation set to find a best model in 1000 epochs. For UCI categorical experiment, learning rate is 1e-2 and use validation set to find a best model in 1000 epochs. For bimodal experiments, learning rate is 1e-4 and use validation set to find a best model in 1000 epochs. All modules in our models are trained jointly.
In our model, we calculate the conditional log-likelihood of unobserved modality by generating corresponding modalities from prior. We initially train the model for some (empirically we choose 20) epochs without calculating the conditional log-likelihood of $\mathbf{x_u}$. And then first feed the partially-observed data to the model and generate the unobserved modality $\tilde{\mathbf{x}}_\mathbf{u}$ without calculating any loss; then feed the same batch for another pass, calculate the conditional log-likelihood using real $\mathbf{x_o}$ and generated $\mathbf{x_u}$ as ground truth.

C ADDITIONAL RESULTS

C.1 UCI REPOSITORY DATASETS

	Phishing	Zoo	Mushroom
AE	0.348 ± 0.002	0.295 ± 0.022	0.556 ± 0.009
VAE	0.293 ± 0.003	0.304 ± 0.009	0.470 ± 0.017
CVAE w/ mask	0.241 ± 0.003	0.270 ± 0.023	0.445 ± 0.004
MVAE	0.308 ± 0.015	0.233 ± 0.013	0.586 ± 0.019
VSAE	$\mathbf{0.237 \pm 0.001}$	$\mathbf{0.213 \pm 0.004}$	$\mathbf{0.396 \pm 0.008}$
CVAE w/ data	0.301 ± 0.005	0.323 ± 0.032	0.485 ± 0.034
VAEAC	0.240 ± 0.006	0.168 ± 0.006	0.403 ± 0.006

Table 3: **Imputation on Categorical datasets**. Missing ratio is 0.5. Last two rows are trained with fully-observed data. Evaluated by PFC, lower is better.

	Yeast	White Wine	Glass
AE	0.737 ± 0.036	0.3772 ± 0.0008	1.651 ± 0.049
VAE	0.461 ± 0.001	0.3714 ± 0.0001	1.409 ± 0.011
CVAE w/ mask	0.449 ± 0.001	0.3716 ± 0.0001	1.498 ± 0.0013
MVAE	0.442 ± 0.018	$\mathbf{0.3722 \pm 0.0009}$	1.572 ± 0.035
VSAE	$\mathbf{0.409 \pm 0.012}$	$\mathbf{0.3711 \pm 0.0002}$	$\mathbf{1.312 \pm 0.021}$
CVAE w/ data	0.449 ± 0.0001	0.3567 ± 0.0016	1.380 ± 0.045
VAEAC	0.447 ± 0.0016	0.3647 ± 0.0039	1.432 ± 0.027

Table 4: **Imputation on Numerical datasets**. Missing ratio is 0.5. Last two rows are trained with fully-observed data. Evaluated by NRMSE, lower is better.

C.2 MNIST+MNIST BIMODAL DATASET

	0.3	0.5	0.7
AE	0.2124 ± 0.0012	0.2147 ± 0.0008	0.2180 ± 0.0008
VAE	0.1396 ± 0.0002	0.1416 ± 0.0001	0.1435 ± 0.0006
CVAE w/ mask	0.1393 ± 0.0002	0.1412 ± 0.0006	0.1425 ± 0.0012
MVAE	0.1547 ± 0.0012	0.1562 ± 0.0003	0.1579 ± 0.0006
VSAE	$\mathbf{0.1371 \pm 0.0001}$	$\mathbf{0.1376 \pm 0.0002}$	$\mathbf{0.1379 \pm 0.0001}$
CVAE w/ data	0.1336 ± 0.0003	0.1340 ± 0.0003	0.1343 ± 0.0002
VAEAC	0.1333 ± 0.0004	0.1338 ± 0.0003	0.1344 ± 0.0001

Table 5: **Imputation on MNIST+MNIST.** Missing ratio is 0.3, 0.5 and 0.7. Last two rows are trained with fully-observed data. Evaluated by combined errors of two modalities, lower is better.

Figure 5: **Imputation on MNIST+MNIST.** Top row visualizes observed modality, middle row unobserved modality, and bottom row shows the imputation of unobserved modality from VSAE.

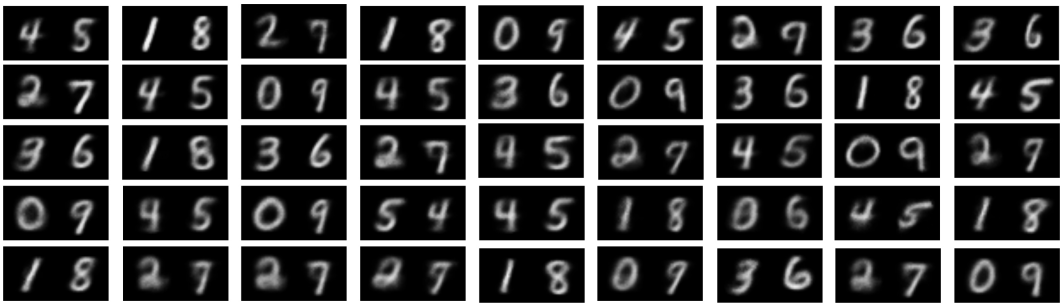

Figure 6: **Generation on MNIST+MNIST.** Generated Samples w/o conditional information. As shown, the correspondence between modalities (pre-defined pairs) are preserved while generation.

C.3 MNIST+SVHN BIMODAL DATASET

	MNIST-MSE/784	SVHN-MSE/3072	Bimodal Error
AE	0.0867 ± 0.0001	0.1475 ± 0.0006	0.2342 ± 0.0007
VAE	0.0714 ± 0.0001	0.0559 ± 0.0027	0.1273 ± 0.0003
CVAE w/ mask	0.0692 ± 0.0001	0.0558 ± 0.0003	0.1251 ± 0.0005
MVAE	0.0707 ± 0.0003	0.602 ± 0.0001	0.1309 ± 0.0005
VSAE	$\mathbf{0.0682 \pm 0.0001}$	$\mathbf{0.0516 \pm 0.0001}$	$\mathbf{0.1198 \pm 0.0001}$
CVAE w/ data	0.0716 ± 0.0002	0.0550 ± 0.0007	0.1266 ± 0.0005
VAEAC	0.0682 ± 0.0001	0.0523 ± 0.0001	0.1206 ± 0.0001

Table 6: **Imputation on MNIST+SVHN.** Missing ratio is 0.5. Last two rows are trained with fully-observed data. Evaluated by combined errors of two modalities, lower is better.

	0.3	0.5	0.7
AE	0.1941 ± 0.0006	0.2342 ± 0.0007	0.2678 ± 0.0012
VAE	0.1264 ± 0.0001	0.1273 ± 0.0003	0.1322 ± 0.0005
CVAE w/ mask	0.1255 ± 0.0002	0.1251 ± 0.0005	0.1295 ± 0.0006
MVAE	0.1275 ± 0.0029	0.1309 ± 0.0005	0.1313 ± 0.0013
VSAE	$\mathbf{0.1217 \pm 0.0002}$	$\mathbf{0.1198 \pm 0.0001}$	$\mathbf{0.1202 \pm 0.0002}$
CVAE w/ data	0.1288 ± 0.0011	0.1266 ± 0.0005	0.1248 ± 0.0003
VAEAC	0.1218 ± 0.0002	0.1206 ± 0.0001	0.1211 ± 0.0001

Table 7: **Imputation on MNIST+SVHN.** Missing ratio is 0.3, 0.5 and 0.7. Last two rows are trained with fully-observed data. Evaluated by combined errors of two modalities, lower is better.