Plug&Play Tractable Multimodal Probabilistic Learning

Jian Vora *jianv@cs.stanford.edu*
Department of Computer Science
Stanford University

Isabel Valera *ivalera@cs.uni-saarland.de*
Department of Computer Science
Saarland Informatics Campus

Guy Van den Broeck *guyvd@cs.ucla.edu*
Computer Science Department
University of California, Los Angeles

Antonio Vergari *avergari@ed.ac.uk*
School of Informatics
University of Edinburgh

Abstract

Data from multiple modalities is ubiquitous in many real world applications and learning a generative model over multimodal data has multiple use cases. Given M modalities or views of a data-point in the form of $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M$, we aim to learn a joint distribution $p_\theta(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M)$ that allows for exact inference queries. This is useful as it allows us to sample from a conditional model which allows us to do things like text guided image synthesis and image captioning from the same model itself (with image, text as the modalities). We can also train using missing data, a common issue in problems with multiple modalities.

It is known that probabilistic circuits as a generative model allow for tractable evaluation of the above inference queries. We train independent auto-encoders for each modality and train a tractable generative model on the joint latent space which is simply the concatenation of the latent spaces of individual modalities. We show that our method has a better performance, both quantitative and qualitative on MNIST-SVHN, CelebA-Attributes as compared to prior VAE based methods such as MMVAE (Shi et al., 2019) and MVAE (Wu & Goodman, 2018) which cannot compute inference queries and only rely on variational approximations for the same.

1 Introduction

Deep neural networks have shown impressive performance on a variety of tasks in domains such as vision (Dosovitskiy et al., 2020; Ramesh et al., 2021b), natural language (Brown et al., 2020a), speech (Hsu et al., 2021; Baevski et al., 2020). On many benchmarks such as image classification, question answering, and natural language understanding tasks, these models have shown to beat human performance. However, these models have been designed to perform well only on the modalities on which they are trained on. There has been recent interest in learning models on multiple modalities. With human learning being inherently multimodal, having models that can learn from different modalities seems to be the natural next step.

Multimodal data, i.e., data comprising different sets of features (modalities) belonging to different distributions, is ubiquitous: from collections of heterogeneous unstructured representations of objects (e.g. text, images, audio, categories, to describe a single data point) to sets of homogeneous features providing different views of samples (multi-view learning). The advantage of having multiple modalities is that different *views* of the same datapoint help in gathering additional information of the datapoint which can be useful for downstream tasks. For example, for the task of representation learning for speech, having a small amount of paired speech-text data helps in grounding the speech and text representations better and adds more semantic information in the speech representations. Given multimodal data, there are a variety of tasks one could do which include learning a unified encoder for all the modalities, learning a generative model over multimodal data or learn cross-modality tasks such as visual question-answering (VQA), text guided image synthesis, image captioning, etc. In this work, we aim to learn a joint distribution over multimodal data.

A principled probabilistic treatment of multimodal learning would allow not only to compactly represent multimodal distributions but also to perform inference over them. It would be possible to draw new samples from all or some modalities, or compute the likelihood of some joint assignment. For example, consider images and text (caption associated with the image) as two modalities over which we want to learn a joint distribution. We can use the *same* model for (a) image captioning, (b) image generation from text prompt, and (c) joint generation of an image with its caption. Although this example constitutes two modalities, the idea is more general and could encompass more than two modalities (for example visual, auditory, and text such as movies). Missing modalities is one of the common issues in multimodal learning. We want a treatment which helps us to handle missing modalities in a principled manner. One way to handle for missing modalities would be able to marginalize those modalities from the learnt joint distribution.

Notation: Upper-case letters X denote random variables (RVs) and lower-case letters their values, i.e., $x \sim X$. Similarly, we denote (ordered) sets of RVs as $\mathbf{X}$, and their corresponding values as $\mathbf{x}$. For a general discussion, assume that the set of RVs comprises M *modalities*—also referred to as *views* in the literature, i.e., we have a partitioning of the feature space as $\mathbf{X} = \bigcup_{i=1}^{M} \mathbf{X}_i$ and $\mathbf{X}_i \cap \mathbf{X}_j = \emptyset$ for any $i \neq j$, where $i, j \in \{1, \ldots, M\}$. When there will be the need to refer a particular modality, we will label its corresponding RV set accordingly, e.g., $\mathbf{X}_{\mathsf{txt}}$ for some structured representation of text data (like a bag-of-words representation, or some text embedding), $\mathbf{X}_{\mathsf{img}}$ for images, and so on.

Goal: We would like to learn a joint distribution $p_\theta(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M)$ which would allow us to do tractable inference on the learnt distribution. Concretely, we want to answer the following inference queries tractably:

1. Modality sampling ($\mathsf{M-SAM}$). We would like to sample from any subset of modalities given any other subset, that is

$$\boldsymbol{x}_{M_1}^{\mathsf{new}} \sim p_\theta(\mathbf{X}_{M_1} \mid \mathbf{X}_{M_2} = \boldsymbol{x}_{M_2}) \tag{1}$$

where $\mathbf{X}_{M_1}, \mathbf{X}_{M_2} \subset \mathbf{X}$ and $\mathbf{X}_{M_1} \cap \mathbf{X}_{M_2} = \emptyset$.

2. Modality MAR inference ($\mathsf{M-MAR}$). Here we would like to compute (log-)likelihoods while being able to marginalize over arbitrary sets of modalities (and potentially condition over arbitrary evidence). A prototypical query in the class looks like:

$$p_\theta(\boldsymbol{x}_{M_1} \mid \boldsymbol{x}_{M_2}) \tag{2}$$

where $\mathbf{X}_{M_1}, \mathbf{X}_{M_2} \subset \mathbf{X}$ and $\mathbf{X}_{M_1} \cap \mathbf{X}_{M_2} = \emptyset$ and we are marginalizing over $\mathbf{X} \setminus \{\mathbf{X}_{M_1} \cup \mathbf{X}_{M_2}\}$.

3. Modality MAP inference ($\mathsf{M-MAP}$). In this case, we would like to retrieve the mode of the conditional distribution obtained after conditioning on some subsets of modalities

$$\boldsymbol{x}_{M_1}^{*} = \mathrm{argmax}_{\boldsymbol{x}_{M_1}} p_\theta(\boldsymbol{x}_{M_1}, \boldsymbol{x}_{M_2}) \tag{3}$$

where $\mathbf{X}_{M_1} \cup \mathbf{X}_{M_2} = \mathbf{X}$ and $\mathbf{X}_{M_1} \cap \mathbf{X}_{M_2} = \emptyset$.

4. Modality marginal MAP inference ($\mathsf{M-MMAP}$). This case is similar to $\mathsf{M-MAP}$ but here we marginalize over some modalities which we don't care about (e.g., modalities missing).

In this work, we investigate how adopting tractable probabilistic models (TPMs), more specifically Sum Product Networks (Poon & Domingos, 2012), jointly model multimodal data that can enable tractable probabilistic inference over subsets of the modalities at hand as well as scaling multimodal learning.

2 Related Work

Generative Modeling: There has been a significant advances in the quality of generative models over the past decade. The task of these models is to learn a distribution $p_\theta(\mathbf{x})$ parameterized by θ given a dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$ where each $\mathbf{x}_i \sim p(\mathbf{x})$ where $p(.)$ is the true data-generating process. On a very high level, the goal is to make $p_\theta(\mathbf{x})$ to be close to $p(\mathbf{x})$ from the N samples present in the dataset $\mathcal{D}$. Models such as variational auto-encoders (VAEs) (Kingma & Welling, 2013), generative adversarial networks (GANs) (Goodfellow et al., 2014), autoregressive models, diffusion models (Dhariwal & Nichol, 2021) have been tremendously successful, and have widely been used for performing a number of tasks including image generation (Saharia et al., 2022; Ramesh et al., 2021a) and text generation (Brown et al., 2020b). While these methods have been quite powerful, each of these generative models is quite task-specific. For instance, a generative model trained for generating images from text can only perform this single task, and is unable to go in the backward direction - i.e. generate captions for images. This drawback limits the power of generative models, which we aim to address through our project by designing a generative model for multimodal data.

In this report we investigate how adopting tractable probabilistic models (TPMs), more specifically Sum Product Networks (Poon & Domingos, 2012), jointly model multimodal data that can enable tractable probabilistic inference over subsets of the modalities at hand as well as scaling multimodal learning. We explore several scenarios where the employed TPMs, in the form of probabilistic circuits (PCs) can deliver implicit or explicit likelihood by aggregating latent representations for different modalities in a "plug&play" fashion.

Multimodal data has been investigated in a number of fashions for non-probabilistic modeling (deterministic mappings) of low-dimensional spaces for multi-view learning.

Among works dealing with probabilistic mappings, recent research lines involve deep generative models like GANs and VAEs. Both are primarly used as simulators (e.g. to sample) as they do not have an explicit likelihood model (GANs) or if computing the likelihood exactly is hard (VAEs). We briefly list some of them and also cite their main limitations.

2.1 VAEs

All models based on VAEs have issues in modeling a joint evidence lower bound (ELBO): many have to represent explicit inference networks for all subsets of modalities at hand, or resort to heuristics during training to let a single architecture adapt to missing (subsets of) modalities.

1. **Variational methods for Conditional Multimodal Deep Learning** (Pandey & Dukkipati, 2017): They introduce **CMMA** which learns one conditional distribution per modality as a conditional VAE.
2. **Deep Variational Canonical Correlation Analysis** (Wang et al., 2016): They introduce **BiVCCA**, a deep CCA requiring a network for each subset of modalities.
3. **Joint Multimodal Learning With Deep Generative Models** (Suzuki et al., 2016): **JMVAE** aims to represent each possible subset of modalities by an inference network.
4. **Generative Models of Visually Grounded Imagination** (Vedantam et al., 2017): They use triple ELBO (**TELBO**) but the method does not generalize to more than 3 modalities.
5. **Multimodal Generative Models for Scalable Weakly-Supervised Learning** (Wu & Goodman, 2018): They introduce **MVAE** as a joint VAE having a product of experts (PoE) formulation which helps dealing with missing modalities (setting each modality input to 0). This seems the best competitor so far. However, during training they still help the model deal with the missing modalities by generating K masks for random subsets of the M modalities and looking at the D marginals. They indeed optimize for:

$$\mathsf{ELBO}(\mathbf{X}_1, \ldots, \mathbf{X}_M) + \sum_{i=1}^{D} \mathsf{ELBO}(\mathbf{X}_i) + \sum_{j=1}^{K} \mathsf{ELBO}(\mathbf{X}_j)$$

6. **Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models** (Shi et al., 2019): They substitute the PoE in the MVAE with a mixture of univariate experts to have a joint posterior, delivering the **MMVAE**. While not requiring the additional terms in the ELBO as in MVAE, they have to resort to more expensive stratified sampling (Robert & Casella, 2005) to average over M modalities.

2.2 GANs

GANs-based models, on the other hand, have the classical issue of pesky adversarial training and we do not use GAN based methods for our comparison as they cannot answer any inference queries even approximately but still mention them here for the sake of completeness.

1. **Adversarially Learned Inference** (Dumoulin et al., 2016): The generation network maps samples from stochastic latent variables to the data space while the inference network maps training examples in data space to the space of latent variables.
2. **Triple Generative Adversarial Nets** (Li et al., 2017): Triple-GAN consists of three players—a generator, a discriminator and a classifier. Needs to model all conditional independencies.
3. **Triangle Generative Adversarial Networks** (Gan et al., 2017): Δ-GAN is developed for semi-supervised cross-domain joint distribution matching, an be considered as a combination of conditional GAN and ALI.

3 Methods

3.1 Probabilistic Circuits

Probabilistic circuits or sum-product networks (Poon & Domingos, 2012) are a class of generative models that model the joint distribution as a polynomial over the input leaf distributions. They allow for efficient tractable inference for the following kinds of queries exactly:

Probabilistic Circuits are implemented as computational graphs consisting of three types of nodes:

1. **Leaf Nodes**: Given a random variable/s at the input of the PC, these nodes parameterise a distribution over that random variable/s. For example, given an input $\mathcal{X}$, these distributions will output $p_{\mathcal{X}}$ for some parametric distribution p such as a Gaussian.
2. **Sum Nodes**: These nodes taken in multiple distributions as inputs and simply output a mixture of these distributions where the weights of the mixture are learnable.
3. **Product Nodes**: These nodes take in multiple distributions as inputs and output the product of these distributions which effectively models the variables in the scope of the product node to be independent.

So eventually, the only learnable parameters of the PC include the mixture weights of all the sum nodes and the parameters of the input distribution. An illustration of a simple PC can be found in Fig **??**. In our work, we use the Probabilistic Circuits to model the joint distribution of multiple modalities over their latent space instead of high-dimensional pixel spaces. PCs are trained using EM (expectation-maximization) by maximising the likelihood over the training data.

3.2 Regularized Autoencoders

Informative latent space encoding of the multimodal subspaces is crucial for training probabilistic circuits on multimodal data. Vanilla deterministic autoencoders have a spiky distribution of the latents making maximum likelihood training hard for probabilistic circuits. To learn a Probabilistic Circuit on the fused latent space, we attempt to enforce smoothness in the learnt latent space. To achieve this, instead of training autoencoders with MSE loss between the input and the reconstruction, we add two additional terms to the loss as shown in (Ghosh et al., 2020): l_2 norm on the latent space and l_2 norm on the decoder gradients while training.

Figure 1: Using the model to sample $\boldsymbol{x}_{k+1}, \dots, \boldsymbol{x}_M$ given $\boldsymbol{x}_1, \dots, \boldsymbol{x}_k$

3.3 PPPC: Plug&Play Probabilistic Circuits

3.3.1 PCs To “Glue" Modalities

We discuss the simplest generative model for multimodal data: it consists of a set of independent mechanisms, one for each modality $i = 1, \dots, M$. Each mechanism can be modeled as an autoencoder mapping a certain (lower-dimensional) latent space $\mathbf{Z}_i$ to $\mathbf{X}_i$, the corresponding set of observed RVs. A global joint generative mechanism for $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M$ is recovered by modeling the joint distribution over latent space $\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_M$ via some tractable model $p_{\mathcal{S}}$, e.g. a PC $\mathcal{S}$. A PC would give the flexibility to perform inference over subsets of the modalities flawlessly: it would *orchestrate* encoding-deconding over different autoencoders. We codename this architecture PPPC.

This is the most basic idea that was initially proposed for mixed sum-product networks (MSPNs) (Molina et al., 2018). The idea of using VAEs instead of PCs has been explored by in (Tan & Peharz, 2019) in the context of uni-modal data. While it might seem appealing to have a joint ELBO, we lose all the advantages of PPPC as the tractable properties discussed above cannot be achieved and we go back to a harder optimization problem (In (Tan & Peharz, 2019) they were not able to scale besides MNIST).

3.3.2 Training and Inference Procedure

Let M be the number of modalities for the multimodal data used for training. Our generative model consists of two major components: M encoder-decoder networks (which are deterministic) trained on each modality $[(\mathbf{E}_1, \mathbf{D}_1), (\mathbf{E}_2, \mathbf{D}_2), \dots, (\mathbf{E}_M, \mathbf{D}_M)]$ and one SPN network $\mathcal{S}$ that fits a generative model on the latent spaces of these encoders.

More formally, let $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_M$ represent a *single* datapoint consisting of the M modalities. Let $\boldsymbol{z}_i = \mathbf{E}_i(\boldsymbol{x}_i)$ denote the latent representations of the input. Denoting $\boldsymbol{z} = \textsc{Concat}(\boldsymbol{z}_1, \boldsymbol{z}_2, \dots, \boldsymbol{z}_M)$, the SPN network predicts the likelihood of $\boldsymbol{z}_1, \boldsymbol{z}_2, \dots, \boldsymbol{z}_M$ as $\mathcal{S}(\boldsymbol{z})$. Thus, our complete likelihood model is simply $\mathcal{S}(\textsc{Concat}(\mathbf{E}_1(\boldsymbol{x}_1), \mathbf{E}_2(\boldsymbol{x}_2), \dots, \mathbf{E}_M(\boldsymbol{x}_M))$.

As an example (figure 1), suppose we want to perform inference queries such as predicting $\boldsymbol{x}_{k+1}, \dots, \boldsymbol{x}_M$ given $\boldsymbol{x}_1, \dots, \boldsymbol{x}_k$. Such a query is useful in the case of text to image generation, where image and text are two of the modalities ($M = 2$ and $k = 1$ in this case). To perform this query, first we encode the given attributes, i.e., compute $\boldsymbol{z}_1, \dots \boldsymbol{z}_k$ in the latent space. Next, we utilize tractable inference of SPNs to compute $\mathcal{S}(\boldsymbol{z}_{k+1}, \dots, \boldsymbol{z}_M | \boldsymbol{z}_1, \dots \boldsymbol{z}_k)$. Finally, we sample from this distribution to obtain $\hat{\boldsymbol{z}}_{k+1}, \dots, \hat{\boldsymbol{z}}_M$ and then utilize the decoders $\mathbf{D}_{k+1}, \dots, \mathbf{D}_M$ to compute $\hat{\boldsymbol{x}}_i = \mathbf{D}_i(\hat{\boldsymbol{z}}_i), i \in \{k+1, \dots, M\}$. Thus, we are able to perform efficient sampling in multimodal data.

3.3.3 Advantages of Plug&Play Learning

The main advantage of PPPC is that each (R)AE could be trained independently from others. This i) greatly simplifies a potentially tough joint optimization problem and ii) provides a single inference machine for all possible inference “directions” (e.g., while conditioning, overcoming the need of other GAN- and VAE-based competitors that have to either train different architectures for each directions or sampling subsets of modalities); iii) allows to leverage sota AE architectures, *out-of-the-box* each tailored for each modalities, while iv) plugging them in (without retraining), out or swapping them more easily. Moreover, v) PCs can flawlessly deal with heterogeneous embeddings. Indeed, we can devise an online learning scheme where we train the PC over $\mathbf{Z}$ by adding some modalities at the time, reusing the partial distributions previously learned.

4 Datasets

(a) (b) (c) (d)

Figure 2: The above figure shows unconditional samples for (a): GMM, (b): SPN, (c): MMVAE, (d): MVAE arranged as a column drawn from the joint distribution $p(x_{\text{mnist}}, x_{\text{svhn}})$. These samples are arranged in order of their likelihoods in a decreasing order

We work with the following datasets for this project:

1. We combined digit pairs of the same kind from MNIST (Lecun et al., 1998) and SVHN (Netzer et al., 2011) to create a multimodal dataset for initial experimentation. We pair each image from the MNIST dataset with 20 images from the SVHN dataset. The MNIST dataset is divided into 50k training images and 10k validation and test images each.

2. We also use the CelebA dataset (Liu et al., 2015) which consists of images and binary attributes. The dataset consists of around 200k celebrity faces, each annotated with 40 attributes. We only work with 4 selected attributes - glasses, hair color, gender, open mouth.

5 Results and Discussion

5.1 Label Paired MNIST-SVHN

In this section, we shall show the results on the MNIST-SVHN multiview dataset which involves pairing of MNIST and SVHN images having the same class label.

5.1.1 Method Details

For this particular experiment, the latent space size for both the modalities was chosen to be 32 which led to a dimensionality of 64 over which the probabilistic circuit was to be learnt. The autoencoder architecture was a simple conv arch with 6 layers and Swish activations which lead to smoother image reconstructions as shown in (Ghosh et al., 2020). The latent space was restricted to lie in a hypercube of $[-1, 1]^{64}$ which shall make it easier to learn the PC and avoid any numerical instability issues. For the PC, we use a randomized GPU implementable architecture as shown in (Peharz et al., 2020) along with Gaussian leaves. For qualitative samples, we also compare with learning a Gaussian Mixture Model (GMM) on the latent space which is a very simple PC with only 1 layer and one sum node.

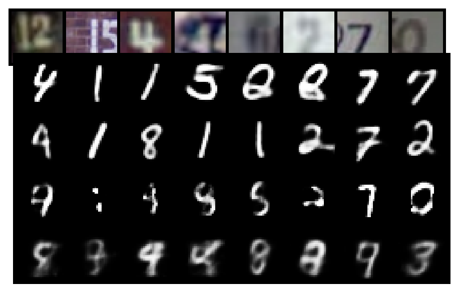
5.1.2 Unconditional Inference Evaluation

Once we learn the joint distribution, we can sample from the joint $p_\theta(\boldsymbol{x}_{\text{mnist}}, \boldsymbol{x}_{\text{svhn}})$ to generate a pair of same class images. We can evaluate the joint query by the following two metrics:

1. Measure the FID scores of the images generated with their corresponding datasets which gives an indication of how well the images look when compared to other samples from the same distribution.

2. Measure coherence, which means how well do the class labels of both the sampled images match with each other. For this, we train two classifiers on MNIST and SVHN separately and the joint predictions of these classifiers are used to measure coherence. The results are in Fig 2.

(a) Generation from $p(x_{svhn}|x_{mnist})$



(b) Generation from $p(x_{mnist}|x_{svhn})$

Figure 3: Samples drawn from conditional distribution with each row indicating a model in the order of (1): GMM, (2): SPN, (3): MMVAE, (4): MVAE. Qualitatively samples from SPNs are the most in sync with the digit classes and also show the least blurry artifacts unlike MVAE/MMVAE

5.1.3 Conditional Inference Evaluation

To understand cross-modality generation capabilities of all the models, we condition on one of the modality and sample the other from the conditional distribution $p(x_{\text{mnist}}|x_{\text{svhn}})$ and $p(x_{\text{svhn}}|x_{\text{mnist}})$. For PCs, we perform an approximate MAP query to assign the most likely image from the conditional distribution. From the samples in figure 3, it can be clearly seen that the samples from the plug-and-play models offer a better coherence in aligning modalities which is further reinforced by the quantitative scores later.

5.1.4 Quantitative Evaluation

In this subsection, we look at the FID scores and the coherence accuracies of the generated samples and compare those with MVAE and MMVAE. We see PPPC outperforms the competitors by a huge margin except of the coherence in the MNIST to SVHN modality case.

The final row AE stands for the autoencoder and can be thought of as the groundtruth/best result that we can obtain for that particular column. Those numbers are indicative of the quality of the learnt autoencoders and the per modality classifiers and hence we cannot expect any generative model to outperform them.

Model	Joint Qua(↓)	Joint Coh(↑)	$\mathsf{Mod}_1 \rightarrow \mathsf{Mod}_2$ FID(↓)	$\mathsf{Mod}_1 \rightarrow \mathsf{Mod}_2$ Acc(↑)	$\mathsf{Mod}_2 \rightarrow \mathsf{Mod}_1$ FID(↓)	$\mathsf{Mod}_2 \rightarrow \mathsf{Mod}_1$ Acc(↑)
MVAE	220.56	28.15	92.58	54.60	94.27	27.45
MMVAE	112.49	34.75	101.58	72.25	35.98	**59.02**
PPPC	**87.71**	**38.10**	**64.60**	**75.67**	**21.34**	47.35
AE	76.49	78.38	55.52	81.48	17.42	98.15

Table 1: Quantitative Evaluation of generative capacities of various models using FID scores and classification accuracies

5.2 CelebA-Attributes

In this section, we shall show the results on the CelebA-Attributes dataset which involves a CelebA image (64×64) and associated 4 binary attributes (blonde hair, gender, mouth open or not, glasses on or not) and hence we have a total of 5 modalities.

5.2.1 Method Details

For this particular experiment, the latent space size for images was chosen to be 64 and all the binary attributes did not have any encoder or decoder associated with them and were simply appended in the joint latent space $\mathbf{Z}$ which led to learning a PC over a dimensionality of 68. Here, as the last 4 variables were binary, we modified the PC leaves such that they encoded a Bernoulli distribution for these variables and a Gaussian distribution for the latent codes obtained from the image encoder. The image latent space was also constrained in $[-1, 1]^{64}$ and we also tested with the GMM training on the joint latent space.

5.2.2 Unconditional Inference Evaluation

We sample from the joint distribution and observe the images generated after passing through the decoder in Fig 4. As one can clearly observe, PPPC leads to better unconditional samples and are sharper compared to other methods.

(a) (b) (c)

Figure 4: The above figure shows unconditional samples (only the image modality) for (a): PPPC, (b): MMVAE, (c): MVAE drawn from the joint distribution arranged in order of decreasing likelihoods for each method

5.2.3 Conditional Inference Evaluation

We start off with just conditioning on a single attribute being true and observe the generated image and visually see the coherence between the conditioned attribute and image sampled in Fig 5. Again the coherence of our model is much better compared other counterparts.

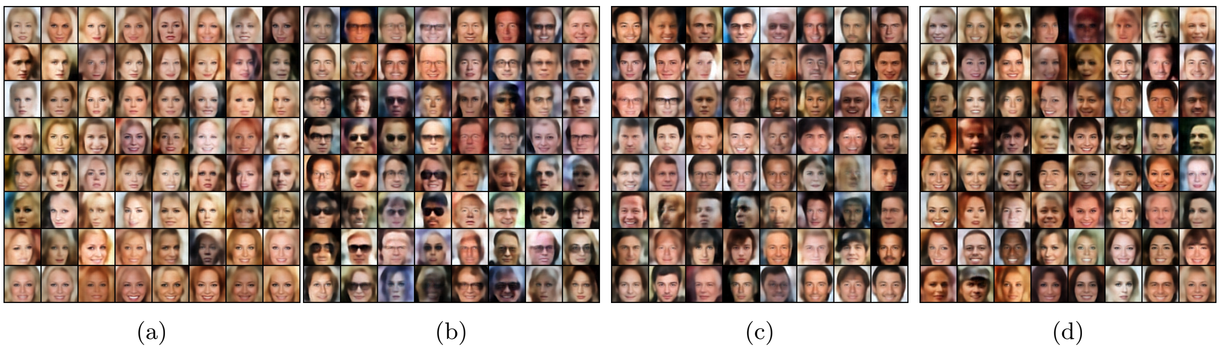


Figure 5: The above figure shows conditional samples for (a): Blond Hair, (b): Glasses, (c): Male, (d): Mouth Open attributes true for PPPC. Samples from MVAE and MMVAE are in the Appendix as their performance is significantly worse.

We next condition over multiple modalities which was the main motivating factor of using a tractable probabilistic model. We get the following samples where each row denotes the following features as denoted in figure 6:

1. Blond Hair + Glasses
2. Male + Mouth Open
3. Blond Hair + Woman + Mouth Close
4. Blond Hair + Woman + Mouth Open

Past methods such as MVAE & MMVAE use variational approximations, which leads to strange images.

Figure 6: Samples generated by conditioning on 1) blond hair + glasses, 2) male + mouth open, 3) blond hair + woman + mouth close, 4) blond hair + woman + mouth open, arranged row-wise from top to bottom respetively

5.2.4 Quantitative Evaluation

For unconditional queries, we measure the image quality by measuring the FID scores of the generated samples. We also want to measure the coherence between the generated image and the generated attributes for the same. For this, we train 4 separate binary classifiers which take in the image as an input and outputs the class of the image for each of the attribute. Thus, to measure coherence, we pass the generated image through each of the 4 classifiers and get corresponding labels. As we are sampling from the joint distribution, the model also generates labels and we compare these labels to the labels given by the classifier to measure coherence. Joint coherence is measured by finding the Hamming Loss between these two sets of binary labels (4 in total). The results for unconditional sampling are shown in Table 2.

	Joint	
Model	Qua(↓)	Coh(↑)
MVAE	70.264	0.252
MMVAE	93.031	0.236
PPPC	**66.713**	**0.126**
AE	58.871	0.0365

Table 2: Quantitative Evaluation of generative capacities of various models using FID scores and classification accuracies

	Blond		Glasses		Male		Mouth Open	
Model	FID(↓)	Acc(↑)	FID(↓)	Acc(↑)	FID(↓)	Acc(↑)	FID(↓)	Acc(↑)
MVAE	81.16	20.4	120.74	6.5	93.59	2.9	69.61	19.3
MMVAE	118.02	6.2	137.35	7.1	104.55	41.8	99.72	34.7
PPPC	**72.47**	**81.2**	**89.27**	**50.7**	**74.64**	**87.4**	**63.83**	**57.4**

Table 3: Quantitative Evaluation of generative capacities of various models using FID scores and classification accuracies by conditioning on single attributes

For evaluating conditional queries quantitatively, we use the FID scores and classification accuracies (condition on only 1 attribute at a time). We use the same classifiers mentioned above to calculate the classification accuracies. The results are shown in Table 3.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020b. URL https://arxiv.org/abs/2005.14165.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. URL https://arxiv.org/abs/2105.05233.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

Zhe Gan, Liqun Chen, Weiyao Wang, Yuchen Pu, Yizhe Zhang, Hao Liu, Chunyuan Li, and Lawrence Carin. Triangle generative adversarial networks. In *Advances in neural information processing systems*, pp. 5247–5256, 2017.

Partha Ghosh, Mehdi S. M. Sajjadi, Antonio Vergari, Michael Black, and Bernhard Scholkopf. From variational to deterministic autoencoders. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1g7tpEYDS.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013. URL https://arxiv.org/abs/1312.6114.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998. ISSN 0018-9219. doi: 10.1109/5.726791.

Chongxuan Li, Taufik Xu, Jun Zhu, and Bo Zhang. Triple generative adversarial nets. *Advances in neural information processing systems*, 30, 2017.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Alejandro Molina, Antonio Vergari, Nicola Di Mauro, Sriraam Natarajan, Floriana Esposito, and Kristian Kersting. Mixed sum-product networks: A deep architecture for hybrid domains. In *AAAI*, 2018.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

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.

Robert Peharz, Steven Lang, Antonio Vergari, Karl Stelzner, Alejandro Molina, Martin Trapp, Guy Van den Broeck, Kristian Kersting, and Zoubin Ghahramani. Einsum networks: Fast and scalable learning of tractable probabilistic circuits. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, jul 2020. URL http://starai.cs.ucla.edu/papers/PeharzICML20.pdf.

Hoifung Poon and Pedro Domingos. Sum-product networks: A new deep architecture, 2012.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021a. URL https://arxiv.org/abs/2102.12092.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8821–8831. PMLR, 18–24 Jul 2021b. URL https://proceedings.mlr.press/v139/ramesh21a.html.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 0387212396.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. URL https://arxiv.org/abs/2205.11487.

Yuge Shi, N Siddharth, Brooks Paige, and Philip Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Advances in Neural Information Processing Systems*, pp. 15692–15703, 2019.

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

Ping Liang Tan and Robert Peharz. Hierarchical decompositional mixtures of variational autoencoders. In *International Conference on Machine Learning*, pp. 6115–6124, 2019.

Ramakrishna Vedantam, Ian Fischer, Jonathan Huang, and Kevin Murphy. Generative models of visually grounded imagination. *arXiv preprint arXiv:1705.10762*, 2017.

Weiran Wang, Xinchen Yan, Honglak Lee, and Karen Livescu. Deep variational canonical correlation analysis. *arXiv preprint arXiv:1610.03454*, 2016.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, pp. 5575–5585, 2018.