Bridged Variational Autoencoders for Joint Modeling of Images and Attributes

Ravindra Yadav
IIT Kanpur
ravin@iitk.ac.in

Ashish Sardana
NVIDIA
asardana@nvidia.com

Vinay P Namboodiri
IIT Kanpur
vinaypn@iitk.ac.in

Rajesh M Hegde
IIT Kanpur
rhegde@iitk.ac.in

Abstract

Generative models have recently shown the ability to realistically generate data and model the distribution accurately. However, joint modeling of an image with the attribute that it is labeled with requires learning a cross modal correspondence between image and attribute data. Though the information present in a set of images and its attributes possesses completely different statistical properties altogether, there exists an inherent correspondence that is challenging to capture. Various models have aimed at capturing this correspondence either through joint modeling of a variational autoencoder or through separate encoder networks that are then concatenated. We present an alternative by proposing a bridged variational autoencoder that allows for learning cross-modal correspondence by incorporating cross-modal hallucination losses in the latent space. In comparison to the existing methods, we have found that by using a bridge connection in latent space we not only obtain better generation results, but also obtain highly parameter-efficient model which provide 40% reduction in training parameters for bimodal dataset and nearly 70% reduction for trimodal dataset. We validate the proposed method through comparison with state of the art methods and benchmarking on standard datasets.

1. Introduction

The ability to generate images from concepts is a challenging problem. In this problem, we are required to generate images just based on their attribute description. In recent years, generative models have been successful in unsupervised learning of data distributions [9, 3, 12, 2]. This is an appealing approach as the ability to generate samples implies that the distribution is learned and can be adapted for other machine learning applications. The other reason is that as the learning is unsupervised, this gives us the opportunity to make use of abundantly available data from various different sources. However, cross-modal generation has been more challenging. This is because, learning this would require learning correspondence between multiple modalities.

Recently, multimodal approaches have shown promising results for various different tasks like cross-modal retrieval [11, 18], localization [16], object identification [15] etc. Having multiple modalities during training can be seen as providing an extra source of information to the model, thus models trained using multiple modalities learn better latent representations than what is possible from using single modality only. Another very powerful technique has been using conditional generative adversarial networks for cross-modal representations [6, 26]. However, these rely on image to image level correspondence and further they do not provide a probabilistic ability to generate accurate likelihoods for generation.

In context of generative frameworks, various architectures have also been proposed that first extract high level feature representation of individual modality and then later combine them to form a joint representation (as shown in Fig.2(a)), the network is then trained end-to-end to produce minimum reconstruction loss for each individual modality. But, during testing phase, normally, the task is to do cross-modal generation therefore the model should be suited to handle situations when some input modalities are missing, but it should still be able to generate all modalities at output accurately. To solve this problem, previously proposed approaches make use of data augmentation [11, 13], or trains additional encoder networks retrofitted to a main joint encoder network [21, 19] (Fig.2(b)). We found that by using only the decoder layers of joint network and retrofitting new encoders layers results in suboptimal performance since we are inheriting only half of the information that the joint model, which had access to all the modalities, has learnt about the data.

Our model and Wu et al. [23] try to solve the same problem which retrofit models have, that is explosion in number of training parameters needed as we increase the number of modalities in the dataset. The difference is that while they proposed a different training objective which is product-of-expert of all modalities, we, on the other hand, have proposed a different model architecture to solve the same problem. But, in [23], as the author have mentioned in their

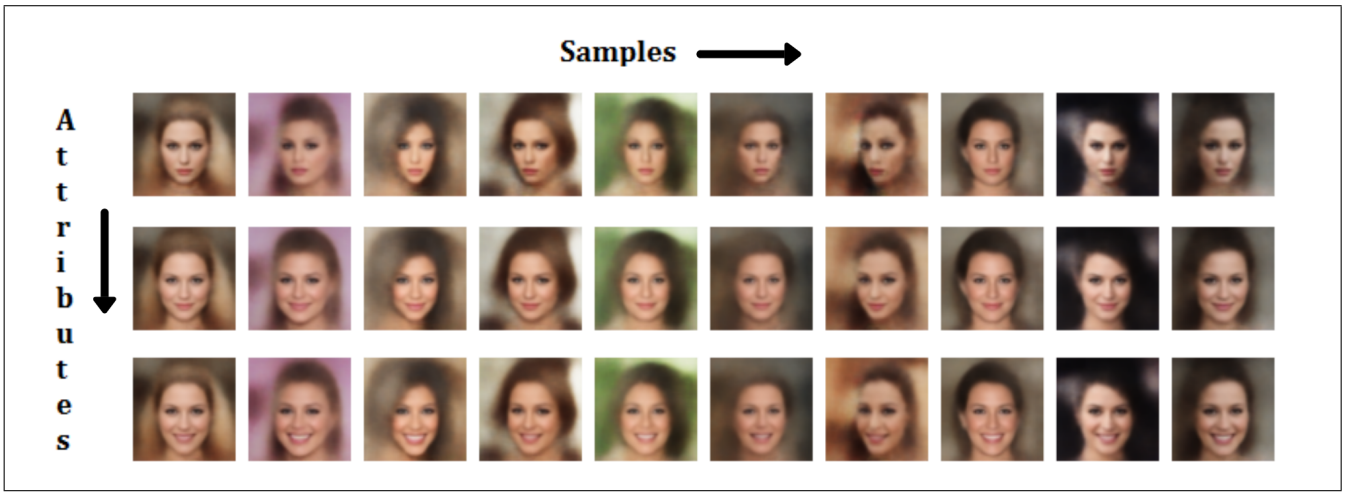

Figure 1. A Multimodal Generative Model should be able to generate *distinctive* samples of a particular modality for a given value of other modality. For example, above figure shows the variations in generated samples (along columns) for different attributes vectors (along rows) input to the network, obtained using proposed model. Best viewed in color.

model training the PoE inference model, does not train the individual inference networks well. Therefore, they have to resort to sub-sampling approach which results in loss function which is sum of numerous ELBO terms. To keep a balance between these individual elbo terms, they need different weighing hyper-parameters for each term, otherwise the overall loss will easily get biased towards the modality with the highest dimension. The number of hyper-parameters needed, in this case, will therefore combinatorially increases with number of modalities. Therefore, there approach achieves higher parameter-efficiency but at the cost of increase in amount of computations needed. Our proposed model, on the other hand, provides almost same amount of parameter-efficiency without any increase in amount of computations.

In this paper, we propose a novel multimodal Variational Autoencoders based architecture which retains both encoder and decoder networks of the joint model (Fig.2(d)), and does not need to train additional encoder networks, thus inheriting all the information that the joint model has learnt about the dataset. Though the proposed algorithm is evaluated on labeled datasets, where one of the modality is low-dimensional discrete attribute vector and other modality is a high-dimensional image vector. It is straightforward to extend the model to scenarios where both the modalities are high dimensional. Further, the evaluation is done on bimodal and trimodal datasets, however, the architecture allows for extension to more than three modalities.

In summary our contributions are as follows:

- Parameter-efficiency and Computational-efficiency: In comparison to the state-of-the-art models [23], our model provides almost 40% reduction in number of training parameters required, which for our trimodal dataset increases to nearly 70%, without any increase in amount of computations.
- Better transfer Learning: Another advantage that the proposed model provides is in terms of transfer learning where we don't have to train new set of encoder networks separately, but rather inherit *all* the learnt features of the joint encoder itself, which is, intuitively, important since joint encoder was trained using all the modalities.

2. Related Work

We start with making a brief introduction of Variational Autoencoders and later see some of the recent approaches towards multimodal learning using them.

Variational Autoencoders: Variational Autoencoders [9, 17] are probabilistic latent variable models that explicitly try to maximize the marginal likelihood (called evidence) of each datapoint x in the training set under the entire generative process. The log marginal likelihood of any datapoint x is expressed as,

$$\log p_\theta(x) = \mathcal{D}(q_\phi(z|x)||p_\theta(z|x)) + \underbrace{\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \mathcal{D}(q_\phi(z|x)||p_\theta(z))}_{\mathcal{L}(\theta,\phi;\mathrm{x})} \tag{1}$$

where $\mathcal{L}(\theta, \phi; \mathrm{x})$ is the evidence lower bound, and $\mathcal{D}$ refers to Kullback Leibler divergence. Since, the first term on right hand side is intractable, but positive, therefore $\mathcal{L}(\theta, \phi; \mathrm{x})$ is

optimized to get closer to the evidence. The $q_\phi(z|x)$ and $p_\theta(x|z)$ are the inference (or encoder/recognition) and generative (or decoder) models respectively, and $p_\theta(z)$ is the prior over latent space. The entire network can be trained end-to-end using stochastic gradient descent, by applying reparameterization in the stochastic layers of the network.

Now, let us consider a bimodal dataset D = $\{(x_1, y_1)$, (x_2, y_2), ..., $(x_N, y_N))\}$ consisting of N data points, where each datum consists of two variables x and y corresponding to each modality. The objective of below multimodal models is to find correlation between the two modalities using these duplet data points.

Joint Multimodal Variational Autoencoder: The JMVAE model [19] is one of the first model, that had used VAEs for multimodal application. It consists of one joint multimodal inference network $q_\phi(z|x, y)$ and individual unimodal inference networks $q_\phi(z|x)$ and $q_\phi(z|y)$ to handle test scenarios where we want the model to do cross-modal generation. To get the same latent representation in missing modality situation as in when all modalities are present, for each modality it contains additional KL-Divergence terms in the training objective function.

$$\begin{aligned}\mathcal{L}_{\text{JMVAE-kl}} = \mathcal{L}_{\text{JMVAE-zero}} - \alpha[&\mathcal{D}(q_\phi(z|x,y)||q_{\phi_\text{x}}(z|x)) + \\ &\mathcal{D}(q_\phi(z|x,y)||q_{\phi_\text{y}}(z|y))]\end{aligned} \tag{2}$$

where,

$$\begin{aligned}\mathcal{L}_{\text{JMVAE-zero}} = &-\mathcal{D}(q_\phi(z|x,y)||p_\theta(z)) + \\ &\mathbb{E}_{q_\phi(z|x,y)}[\log p_{\theta_\text{x}}(x|z) + \log p_{\theta_\text{y}}(y|z)]\end{aligned} \tag{3}$$

There are some drawbacks in this extension: first for each modality there need to be a multiple inference networks $q_\phi(z|x_2)$, $q_\phi(z|x_2, x_3)$, $q_\phi(z|x_2, x_3, x_4)$ and so on to handle cases when modality x_1 is missing, and second it not explicitly clear as to how minimizing $\mathcal{D}(q_\phi(z|x, y)||q_{\phi_\text{x}}(z|x))$ is equivalent to minimizing $\mathcal{D}(q_{\phi_\text{x}}(z|x)||p_{\theta_x}(z|x))$.

VAEs with Product-of-experts: In order to solve the problems in above JMVAE model, in [21] author explicitly minimizes $\mathcal{D}(q_{\phi_\text{x}}(z|x)||p_{\theta_x}(z|x))$ and $\mathcal{D}(q_{\phi_\text{y}}(z|y)||p_{\theta_y}(z|y))$, thus proposing a new training objective function (abbr. TELBO) consisting of three elbo terms. Assuming that modalities factorizes over individual attributes $p(x|z) = \Pi_k p(x_k|z)$, they incorporated Product-of-Experts model [4] in joint VAEs architecture (eq.3) and thus have shown that more abstract results can be generated.

Similar to JMVAE model [19], to handle missing modality scenarios, they also still need additional encoder networks for each modality $q_{\phi_x}(z|x)$ and $q_{\phi_y}(z|y)$, along with the joint model $q_\phi(z|x, y)$.

The objective function is expressed as a sum of three ELBO terms,

$$\begin{aligned}\mathcal{L}_{\text{Triple ELBO}} &= \mathcal{L}_{\text{JMVAE-zero}} - \alpha[\mathcal{D}(q_{\phi_\text{x}}(z|x)||p_{\theta_x}(z|x)) + \\ &\quad \mathcal{D}(q_{\phi_\text{y}}(z|y)||p_{\theta_y}(z|y))] \\ &\propto \mathcal{L}_{\text{JMVAE-zero}} + \mathcal{L}(\phi_\text{x}, \theta_\text{x}; \text{x}) + \mathcal{L}(\phi_\text{y}, \theta_\text{y}; \text{y})\end{aligned} \tag{4}$$

In MVAE model [23], product-of-experts approach has been extended to modalities instead on the attributes. The advantage in doing this is that now they do not need large number of side inference networks to handle missing modality situations. The entire network is trained end-to-end using a single ELBO term that is build upon product-of-experts of modalities.

$$\begin{aligned}\mathcal{L}_{\text{Single ELBO}} = \mathbb{E}_{q_\phi(z|X)}[&\sum_{x_i \in X} \lambda_i \log p_\theta(x_i|z) - \\ &\beta\mathcal{D}(q_\phi(z|X)||p_\theta(z))]\end{aligned} \tag{5}$$

where x_i denote individual modalities, and $X \subseteq \{x_1, x_2, ..x_N\}$ is the subset of observed modalities.

But, as the author have mentioned, training using a product of networks never train the individual inference networks (or small sub-networks). Therefore, they have used a sub-sampling approach, where model is trained with each combination of input modalities, resulting in a ELBO term for each of these combination. The final ELBO term is given by weighted combination of these individual ELBO terms.

Unfortunately, due to difference in dimensionality of the given input modalities, to keep a balance between these individual ELBO terms, the model require a separate weighing hyper-parameters for each of these individual ELBO term, otherwise the overall loss will easily get biased towards the modality with the highest dimension. And these number of hyper-parameters needed, therefore, combinatorially increases with number of modalities.

However, the main advantage there model provides is in terms of higher parameter-efficiency compared to the retrofit models. But, using the PoE approach decrease the learning ability of the model. Our proposed model, on the other hand, provides almost same parameter-efficiency with improved learning ability.

Bi-Variational Canonical Correlation Analysis: In [22], author have suggested learning of independent Variational Autoencoders for each modality with interacting inference networks, such that each VAE network is able to reconstruct all observed modalities taking only single modality as input. Thus, the overall ELBO function is a convex combination of two lower bounds,

$$\mathcal{L} = \mu\mathcal{L}_{q_\phi(z|x)}(x, y) + (1 - \mu)\mathcal{L}_{q_\phi(z|y)}(x, y) \tag{6}$$

This approach, however, results in generating a mean image for a given attribute vector, thus lacking the ability to produce any variations in the samples. This is because the VAE with the low-dimensional attribute vector

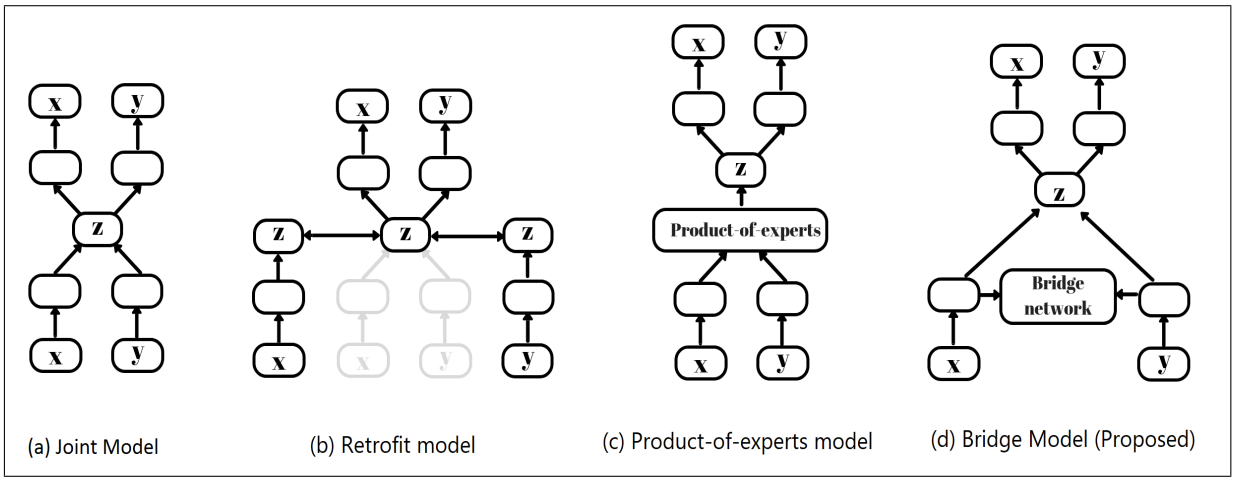

Figure 2. An Encoder-Decoder networks for bimodal datasets. (a) **Joint model:** It requires access to both the modalities, therefore cannot be used for cross-modal generation application. (b) **Retrofitted model:** It involves training additional encoder networks along side with the Joint model, requiring them to be as similar as to the Joint model encoder network. (c) **Product-of-experts model:** The model combines individual sub-inference networks using product-of-experts approach, and therefore do not require a joint model at the center. (d)**Bridged model (Proposed):** The proposed model consists of a bridge encoder, which does all the cross-modal mapping. The model consists of a single Joint model only, therefore do not train additional encoder networks like retrofit models does.

input has to map any given attribute vector to multiple different high-dimensional images, thus due to one-to-many mapping this VAE instead learns to generate the mean value for high-dimensional output modality for that particular attribute vector.

We now present our multimodal VAE architecture. Consider the joint model, in Fig 2(a), since it always has access to both the modalities during training, therefore it is reasonable to expect that representations learnt by it are optimum. If for a moment we ignore the stochastic layer present at the center of the model, the model consists of a set of deterministic encoder layers and decoder layers. The deterministic encoder layers generates a high level feature representation of the inputs, which are then combined (via concatenation) and processed by additional non-linear layers to obtain the parameters of the posterior distribution. The deterministic decoder layers, on the other hand, takes a common sample from posterior distribution and use it to predict the parameters of the likelihood functions of both modalities.

The retrofitted models, in [21, 19], takes the parameters of trained deterministic decoder layers of the joint model and use them to train new set of encoders for cross-modal generation. The main drawback of this approach is that we throw away half the information learnt by joint model by not using its encoders, while training completely new encoders. Another drawback is that these models needs to train individual encoder blocks for every combination of the given inputs, this is because at testing time any of the input(s) can be unavailable and the task will be to reconstruct all the modalities that are used during training. The proposed model provides a solution to both these problems, by inheriting decoders as well as the encoders both of the joint model only. The model builds upon the architecture of joint model only, of Fig 2(a), where the entire cross-modal generation processing is done on high level feature representations of the given modalities using a bridge network in latent space.

3. The Proposed Model

3.1. Model Architecture And Training

In this section, we describe the various blocks and training procedure of the model. For convenience, let us use following notations: Considering a bimodal dataset, let x and y be the two input modalities available to us, Enc_x and Enc_y denotes the encoder blocks for modalities x and y respectively, z_x and z_y are the outputs of the two encoders blocks, Enc_BR is the central encoder block that takes z_x and z_y as inputs and outputs the parameters μ and σ of the posterior distribution (assumed diagonal Gaussian, $p_\theta(z|x,y) = \mathcal{N}(z|\mu,\sigma)$). The detailed encoder architecture is shown in Fig.3.

The Bridge encoder (Enc_BR) is indeed the main building block in the proposed model. It consists of one fully connected network ($\text{FC}_{xy \rightarrow z}$) which takes z_x and z_y as input and generates the parameters μ and σ of the posterior distribution, in addition, it also contains a set of fully connected networks, $\text{FC}_{x \rightarrow y}$ and $\text{FC}_{y \rightarrow x}$, which take z_x as input and output z_y and vice-versa. Apart from the bridge encoder we have two encoder blocks (Enc_x and Enc_y), that generates

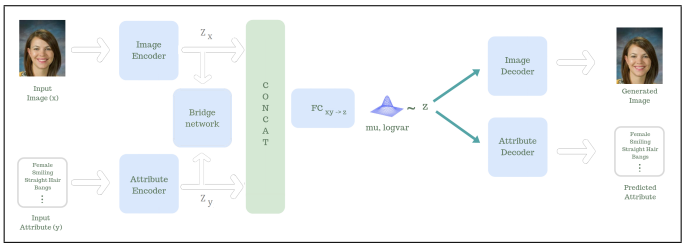

Figure 3. **Proposed Model:** The proposed model consists of two encoders (Image encoder (Enc_x) and Attribute encoder (Enc_y)) and two decoders (Image decoder (Dec_x) and Attribute decoder (Dec_y)).The Bridge network consists of two fully-connected networks $\text{FC}_{x \to y}$ to predict z_y given z_x, and $\text{FC}_{y \to x}$ to predict z_x given z_y.

high level feature representations of x and y. The decoder blocks are mirror symmetric to the corresponding encoder blocks, both blocks takes a common sample from posterior distribution $q_\phi(z|x,y)$ and map it to the likelihood functions $p_{\theta_x}(x|z)$ and $p_{\theta_y}(y|z)$ of both modalities. The training of encoders (Enc_x, Enc_y), decoders and $\text{FC}_{xy \to z}$ layers, is based on maximization of joint log-likelihood of both the given modalities, and can be expressed as,

$$\underset{\theta,\phi}{\text{maximize}} \quad \mathbb{E}_{q_\phi(z|x,y)}[\log p_{\theta_\text{x}}(x|z) + \log p_{\theta_\text{y}}(y|z)] - \mathcal{D}(q_\phi(z|x,y)||p_\theta(z)) \tag{7}$$

where, ϕ jointly denote encoders and $\text{FC}_{xy \to z}$ parameters, while $\theta = \{\theta_x, \theta_y\}$ denotes decoders parameters. The prior is standard normal distribution, $p_\theta(z) = \mathcal{N}(z|0, I)$.

To train $\text{FC}_{x \to y}$ and $\text{FC}_{y \to x}$ layers we first freeze the layers trained above. Later, the $\text{FC}_{x \to y}$ layers then takes feature representation of input x (i.e., z_x) and based on it predict the feature representation of y (i.e., z_y), conditioned that these two representation combined together produces posterior parameters μ and σ of a Gaussian distribution (using $\text{FC}_{xy \to z}$), whose sample z can be mapped to the data likelihood $p_{\theta_x}(x|z)$ through the above trained decoder network. Mathematically, it can be stated as,

$$\underset{FC_{x \to y}}{\text{maximize}} \quad \mathbb{E}_{q_\phi(z|x)}[\log p_{\theta_\text{x}}(x|z)] - \mathcal{D}(q_\phi(z|x)||p_\theta(z)) \tag{8}$$

where, input to $\text{FC}_{xy \to z}$ is [z_x, $\text{FC}_{x \to y}(z_x)$].

Similarly, the training objective for $\text{FC}_{y \to x}$ layers is expressed as,

$$\underset{FC_{y \to x}}{\text{maximize}} \quad \mathbb{E}_{q_\phi(z|y)}[\log p_{\theta_\text{y}}(y|z)] - \mathcal{D}(q_\phi(z|y)||p_\theta(z)) \tag{9}$$

where, input to $\text{FC}_{xy \to z}$ is [$\text{FC}_{y \to x}(z_y)$,z_y].

Note that, for n number of given modalities the retrofit models will require $\sum_{i=1}^{n} {}^nC_i$ networks, while the proposed model will only need $2 * {}^nC_2$ networks, which is very useful when we have large number of modalities and limited computing power. A quantitative comparison on number of trainable parameters needed is given in the experimental section.

4. Experimental Results

Since, it can be difficult to evaluate the performance of generative models, a model performing well on one metric can perform equally worse on another [20, 24]. Therefore, in this section, we will be comparing the proposed model against state-of-the-approaches based on the multiple criterions as follows: (i) Cross-modal generation, (ii) Log-likelihood values and Overfitting, (iii) Image recognition and (iv) Parameter efficiency.

We will be using CelebA [10] and MNIST datasets. For CelebA dataset, similar to [21, 14], we only considered 18 visually distinctive attributes. For all datasets we choose $batch\_size = 128$, $learning\_rate = 10^{-4}$ using Adam optimizer [7]. For CelebA $max\_epcohs = 100$, and for MNIST $max\_epcohs = 500$. We used Batch Normalization [5] and along with LeakyRelu non-linearity during training. The latent space is chosen as 128 dimensional. For CelebA dataset, we used discretized logistic distribution of images [8] as $p_{\theta_\text{x}}(x|z)$ and Bernoulli distribution as $p_{\theta_\text{y}}(y|z)$. And for MNIST dataset, we used Bernoulli distribution as $p_{\theta_\text{x}}(x|z)$ and Categorical distribution as $p_{\theta_\text{y}}(y|z)$.

4.1. Cross-modal generation

Since, our main objective is joint modelling of multiple modalities, therefore in this section we compare the performance of different multimodal models based on cross-modal generation, where for a given value of particular modality we measure how accurate the cross-modal generated results are, and whether they contain enough variations due to stochastic modelling. As pointed out in [21], there is a implicit trade-off between accuracy and variations, therefore it is important that a model should perform well in both of these measures.

For CelebA and MNIST datasets, we show these variations in Fig. 4 and 5 respectively. We can clearly observe that the results generated using proposed model are perfectly accurate and, comparatively, exhibits much more variations both in terms of foreground and background. The TELBO model, though it accurately generates the results, but lacks in the variations compared to the proposed model. The JMVAE model performs even worse in both variations and accuracy. As expected, the BiVCCA model results in generating a mean image only thus lacking the ability to model any variations.

4.2. Log-Likelihood comparison and Overfitting

In this section, we compare these models based on the train and test set marginal log-likelihood values calculated

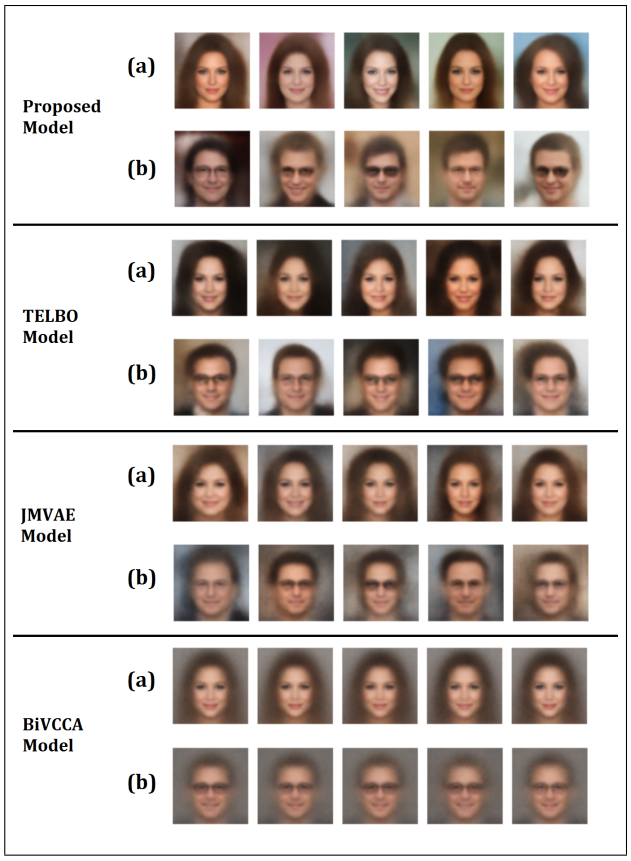

Figure 4. CelebA dataset: Images generated by different models for attributes: (a) {Female, HeavyMakeUp, Smiling, Wavyhair} and (b) {Male, Smiling, Eyeglasses}. Best viewed in color.

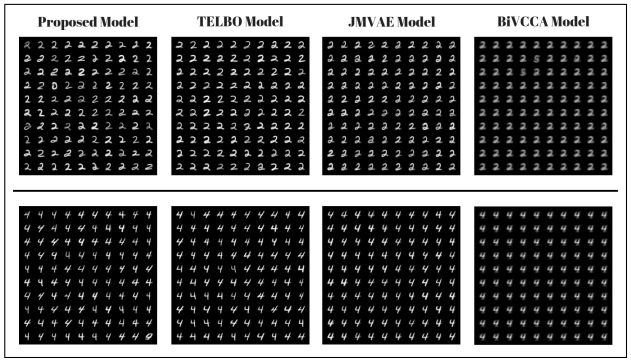

Figure 5. MNIST dataset: Images generated for label value of 2 and 4 by the different models.

for image modality (x). The likelihood values are estimated using the Importance Sampling with 5,000 sample points [1]. Since, modality y is quite a low-dimensional vector, therefore, likelihood values for it are not suited to give useful information.

In the multimodal scenario, to compute likelihood there arise two cases, that is, when we use both x and y modalities and thus sample from posterior distribution $q_\phi(z|x,y)$, and second when we only use x modality to sample from $q_\phi(z|x)$. The calculated values for both the cases are shown in Table 1 and 3. As expected, the joint model gives higher log-likelihood values than the vanilla VAE model which only takes image modality as input, proving the fact that the multimodal models captures the underlying data distribution better than the unimodal models.

Table 1. Test marginal log-likelihood values for MNIST dataset.

Model	$q_\phi(z\|x,y)$	$q_\phi(z\|x)$	Difference
VAE	—	-100.89	
Joint Model	-98.93	—	
JMVAE	-102.16	-103.45	1.29
TELBO	-98.40	-99.64	1.24
MVAE	-96.82	-100.52	3.7
Our Model	-98.55	**-98.76**	**0.21**

Table 2. Train marginal log-likelihood value for MNIST dataset.

Model	$q_\phi(z\|x,y)$	$q_\phi(z\|x)$	Difference
Our Model	-98.59	-98.77	0.18

Table 3. Test marginal log-likelihood values for CelebA dataset.

Model	$q_\phi(z\|x,y)$	$q_\phi(z\|x)$	Difference
VAE	—	-53425.07	
Joint Model	-53283.15	—	
JMVAE	-54333.20	-54853.05	519.85
TELBO	-53293.64	-53601.94	308.3
MVAE	-53017.65	-53743.72	726.07
Our Model	-52170.34	**-52369.66**	**199.32**

As our objective is cross-modal generation, therefore there are two main points we should focus on: (i) Magnitude of log-likelihood values when we only have image modality as input i.e. $q_\phi(z|x)$, and (ii) difference between the log-likelihoods found using both modalities as input and other using only image modality as input. This difference basically signifies how better unimodal encoders are able to learn from the joint model, acting as a quantitative measure of transfer-learning. The proposed model and TELBO model give almost similar results as joint model when these models are given both modalities. The JMVAE model performs worse compared to the two models.

However, an important observation here is that the proposed model performs equally well even when it is fed with only image modality. On the other hand, the performance of TELBO and JMVAE models reduces under single modality

Table 4. Number of training parameters and test marginal negative log-likelihoods for trimodal dataset.

Model	Fashion MNIST		MNIST		Number of Model Parameters	
	$q_\phi(z\|x,y,w)$	$q_\phi(z\|x)$	$q_\phi(z\|x,y,w)$	$q_\phi(z\|y)$	bimodal	trimodal
JMVAE	**-253.71**	-249.74	-107.17	-105.83	60794122	226958346
TELBO	-256.23	-263.48	-107.17	-106.26	60794122	226958346
MVAE	-537.71	-531.31	-527.23	-517.49	31484362	61550346
Our model	-257.28	**-249.35**	**-105.74**	**-102.57**	36697290	72141450

input scenario. This verifies our statement, that in multimodal networks rather than learning a new set of encoders it would be better if we can have same encoders for both scenarios. This is the case when all modalities are available to us on training and some modalities are missing during test. The MVAE model gives the least likelihood values for unimodal cases. The difference between the two likelihood values is also comparatively large, implying the fact that the product-of-expert approach does train the individual sub-networks well (as authors have also observed in their experiments).

In Table 2, we show the train set log-likelihood values for the proposed model for MNIST dataset. Comparing the test and train log-likelihood values in Table1 and 3 shows that the proposed model haven't overfitted to the training data.

4.3. Image recognition application

Since, we are modelling images along with the attribute vector, therefore we can use the trained models for image recognition purposes. The Celeb dataset can be used for this task where model only looks at the image and predicts the corresponding attributes of the image. In table 5, we show the performance of various models in terms of classification accuracy on 2000 test images. We can see that for most of the attributes our model gives best result (11 out of 18 total attributes), only for few cases it gives slightly lesser result (but in those cases also the margin between our model and model performing best is very less). The possible reason for this improved performance is may be due to retaining the weights of the joint model to initialize bottom layers of the unimodal encoders.

4.4. Performance on Trimodal dataset and Number of training parameters

One of the main advantage of the proposed bridged model, in comparison to the retrofit models, is in terms of the number of training parameters. To show the parameter efficiency of the model, as well as, to illustrate that the proposed model allows extension to datasets having more than two modalities, we show the log-likelihood of the models on an artificially generated trimodal dataset. This trimodal dataset is constructed by combining together the Fashion MNIST [25] and MNIST datasets based on the tag value. Using the tag value as a bridge allows a artificial sync to occur between the images of the two datasets. From Table 4, we can see that in bridged model, we require far lesser number of learnable parameters while achieving better log-likelihood than retrofit models in most cases.

Among all the models, the MVAE model require the least number of parameters, but as we can notice the product-of-expert training approach severely deteriorates the learning capability of individual encoders as we increase the number of modalities from two to three (comparing table1 and table4). Our model, on the other hand, require almost same number of training parameters, while achieving much better likelihood values than the MVAE model.

Examples on this cross-modal generation for trimodal dataset are shown in Fig. 6 and Fig. 7. In fig 6, we show the results when models are fed with only the MNIST images, shown in first row, while the second and third rows are the output of the two decoders. Similarly, in fig 7, the models

Table 5. Image to Attributes: Number of errors (out of 2000 test images). Smaller is better.

Attribute	Our model	TELBO	JMVAE	MVAE
Bald	43	51	**42**	50
Bangs	**134**	192	179	310
Black Hair	**226**	369	246	506
Blond Hair	113	159	**107**	289
Brown Hair	**286**	365	352	388
Bushy Eyebrows	**189**	256	222	282
Eye glasses	59	84	**56**	130
Gray Hair	53	78	**50**	80
Heavy Makeup	**230**	337	242	320
Male	**88**	253	98	177
Mouth Slightly Open	296	463	**268**	567
Mustache	82	82	**79**	97
Pale Skin	**74**	84	81	90
Receding Hairline	**142**	162	181	156
Smiling	**193**	316	222	470
Straight Hair	427	479	448	**415**
Wavy Hair	**466**	518	489	603
Wearing Hat	**30**	55	54	92

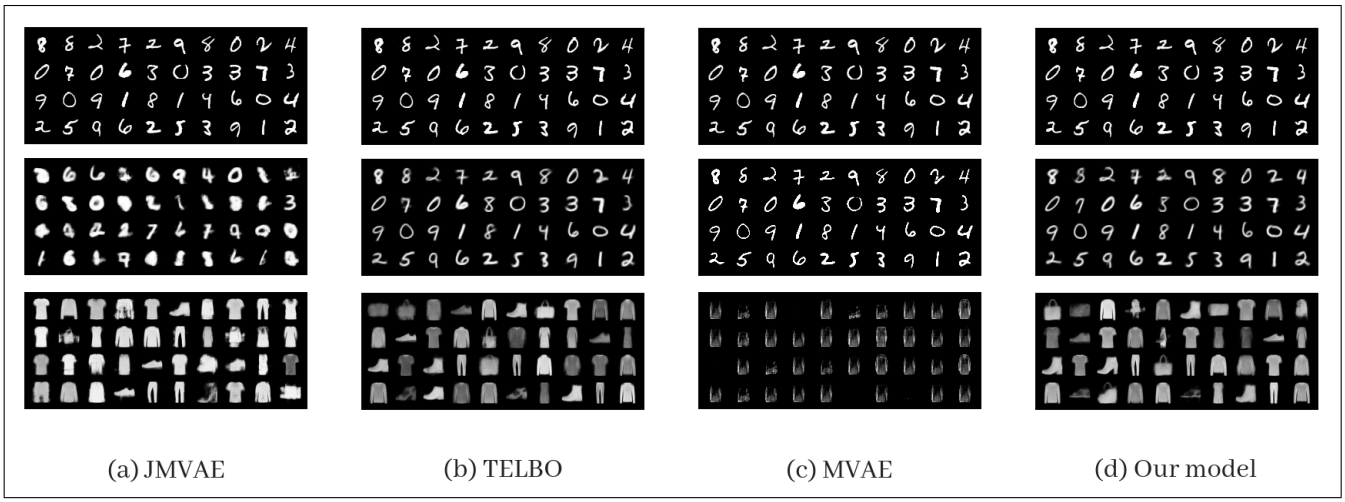

Figure 6. MNSIT to Fashion-MNIST: The top row shows the input to the various models, i.e, the images from MNIST dataset, while the output of decoder1 and decoder2 are shown in row2 and row3 respectively.

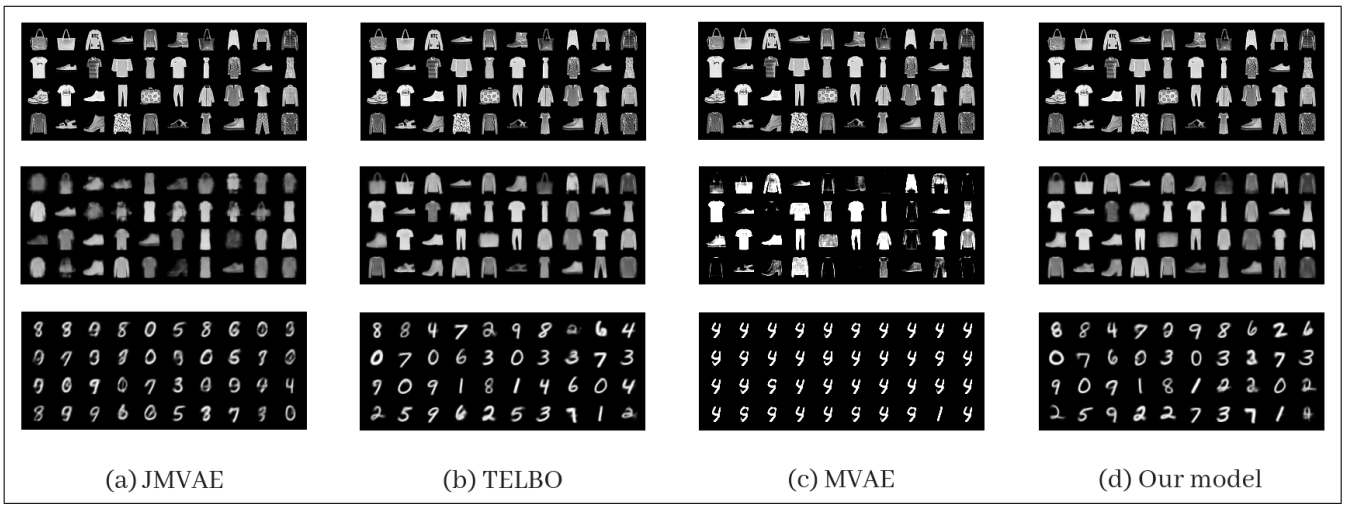

Figure 7. Fashion-MNIST to MNSIT: In this case, only Fashion-MNIST images (shown in top row) are given to the models, the output of the two decoders are shown in row2 and row3 respectively.

are fed with only the Fashion-MNIST images shown in top row of the figure.

In both cases, both TELBO and proposed model can generate the missing modaliteis correctly, while the JMVAE model tend to make mistake very often. As we have found, MVAE model, on the other hand, completely fails to learn the correct data distribution for the present trimodal scenario. This is mainly due to product-of-expert approach used for training, because of which model could not learn correct correlation between different modalities, and therefore collapses to producing average result (producing same cross-modal output irrespective of input), as shown in last row of column three.

5. Conclusion

In this paper we have proposed a bridged variational autoencoder for learning the joint distribution of images and attributes. By incorporating hallucination loss in latent space we have proposed a parameter-efficient network for multimodal datasets, that outperforms state-of-the-art models both quantitatively in terms of log-likelihood values, as well as, qualitatively based of quality of generated images. Furthermore, the results on application of various models for image recognition task have been reported, for which also our model outperforms the state-of-the-art models by a large margin.

References

[1] Y. Burda, R. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

[2] L. Dinh, J. S. Dickstein, and S. Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, pages 2672–2680, 2014.

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

[5] S. Ioffe and S. Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[6] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976. IEEE, 2017.

[7] D. P. Kingma and B. Jimmy. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[8] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improved variational inference with inverse autoregressive flow. *In Advances in Neural Information Processing Systems*, pages 4743–4751, 2016.

[9] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[10] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[11] J. Ngiam, A. Khosla, M. Kim, J. Nam, and A. Y. Ng. Multimodal deep learning. *In Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[12] A. V. D. Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

[13] A. Pal and V. Balasubramanian. Adversarial data programming: Using gans to relax the bottleneck of curated labeled data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[14] G. Perarnau, J. V. D. Weijer, B. Raducanu, and J. M. lvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016.

[15] A. Relja and A. Zisserman. Look, listen and learn. *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[16] A. Relja and A. Zisserman. Objects that sound. *arXiv preprint arXiv:1712.06651*, 2017.

[17] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

[18] N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems*, 2012.

[19] M. Suzuki, K. Nakayama, and Y. Matsuo. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016.

[20] L. Theis, A. V. D. Oord, and M. Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.

[21] R. Vedantam, I. Fischer, J. Huang, and K. Murphy. Generative models of visually grounded imagination. *arXiv preprint arXiv:1705.10762*, 2017.

[22] W. Wang, X. Yan, H. Lee, and K. Livescu. Deep variational canonical correlation analysis. *arXiv preprint arXiv:1610.03454*, 2016.

[23] M. Wu and N. Goodman. Multimodal generative models for scalable weakly-supervised learning. *In Advances in Neural Information Processing Systems (NIPS)*, pages 5575–5585.

[24] Y. Wu, Y. Burda, R. Salakhutdinov, and R. Grosse. On the quantitative analysis of decoder-based generative models. *arXiv preprint arXiv:1611.04273*, 2016.

[25] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. In *arXiv preprint arXiv:1708.07747*, 2017.

[26] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.