Multimodal Generative Learning Utilizing Jensen-Shannon-Divergence

Thomas M. Sutter, Imant Daunhawer, Julia E. Vogt
Department of Computer Science
ETH Zurich
{suttetho,dimant,julia.vogt}@inf.ethz.ch

Abstract

Learning from different data types is a long standing goal in machine learning research, as multiple information sources co-occur when describing natural phenomena. Existing generative models that try to approximate a multimodal ELBO rely on difficult training schemes to handle the intermodality dependencies, as well as the approximation of the joint representation in case of missing data. In this work, we propose an ELBO for multimodal data which learns the unimodal and joint multimodal posterior approximation functions directly via a dynamic prior. We show that this ELBO is directly derived from a variational inference setting for multiple data types, resulting in a divergence term which is the Jensen-Shannon divergence for multiple distributions. We compare the proposed multimodal JS-divergence (mmJSD) model to state-of-the-art methods and show promising results using our model in unsupervised, generative learning using a multimodal VAE on two different datasets.

1 Introduction

Humans are able to process and relate information coming from different sources. Replicating this ability is a longstanding goal in machine learning [2]. Multiple information sources offer the potential of learning better and more generalizable representations, but pose challenges at the same time: models have to be aware of complex intra- and inter-modal relationships, and be robust to missing modalities [28, 17].

So far, supervised learning approaches have had the most success at connecting modalities and translating between them [10, 5]. Indeed, multimodal data is expensive and sparse. This leads to a setting in which only a minority of samples provide all possible observations, while for most of the samples only a subset of observations is available. Despite the success of fully-supervised approaches, unsupervised, generative approaches are another promising approach to capture the joint distribution and flexibly support missing modalities.

In this work, we aim to learn probabilistic representations in an unsupervised manner that are able to integrate information from complementary modalities, reduce uncertainty and ambiguity in redundant sources, as well as handle missing modalities while making no assumptions about the nature of the data, especially about the inter-modality relations.

One of the key challenges in a multimodal setting is the balancing of joint and unimodal latent representations. Previous work [27, 12] had to rely on special training schemes with a mixture of multimodal and unimodal objectives to deal with this difficulty. To overcome this limitation, we base our approach directly in the Variational Bayesian framework and propose a new Evidence Lower Bound (ELBO) for the multimodal setting. We introduce the idea of a dynamic prior for

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

multimodal data, which enables the use of the Jensen-Shannon divergence for M distributions [13, 1] and interlinks the unimodal probabilistic representations of the M observation types.

For the experiments, we concentrate on Variational Autoencoders [11, 20] that provide a framework to jointly learn the parameters of the generative and the inference networks. In this setting, our multimodal extension to variational inference implements a scalable method, capable of handling missing observations. This allows for an architecture consisting of only M inference and generator networks. The output of these inference networks are the unimodal posterior approximations to the multimodal joint distribution.

2 Related Work

We focus on methods with the aim of modelling a joint latent distribution, instead of transferring between modalities [29, 9]. [26, 21] concurrently implemented a multimodal VAE which is non-scalable in terms of number of modalities. [21] introduced the idea that the distribution of the unimodal approximation should be close to the multimodal approximation function. [26, 21, 25] have in common that they use labels as a second modality. More recently, [27] and [12] proposed scalable multimodal generative models by using a Product of Experts [6] as a joint distribution. The Product of Experts (PoE) has problems optimizing the individual experts. Hence, they rely on special training objectives to be able to better learn these unimodal distributions. [3] introduced a local-global approach for VAEs. Their setting is weakly-supervised. They use groups, where a group is defined as multiple samples with at least a single common attribute. Hence, they used a combination of sample- and group based KL-divergence terms for training. [24] adapted this idea to the multimodal setting where the local attributes are modality-specific, generative factors whereas the global ones are the modality-independent, discriminative ones. However, they had to rely on supervision for the discriminative factors. [8] also built on the same idea as [3], they created an unsupervised, non-scalable approach for the bimodal setting.

As extension to the standard VAE [11], there is a line of research focusing on improving the prior distribution. [23] introduced a mixture distribution of variational posterior as prior which is learnable from pseudo-data. [15] and [4] use adversarial training schemes to perform variational inference.

3 Method: mmJSD

Let us consider again some dataset $\mathcal{X}$ consisting of N i.i.d. sets $\mathcal{X} = \{\boldsymbol{X}^{(i)}\}_{i=1}^{N}$ with every $\boldsymbol{X}^{(i)}$ being a set of M modalities $\boldsymbol{X}^{(i)} = \{\boldsymbol{x}_j^{(i)}\}_{j=1}^{M}$. We assume that the data is generated by some random process involving a hidden random variable $\boldsymbol{z} \in \mathbb{R}^d$, where $d, D \in \mathbb{N}$. The generative process decomposes into two stages: 1) a value $\boldsymbol{z}^{(i)}$ is sampled from a prior distribution $p^*(\boldsymbol{z})$ and 2) a value $\boldsymbol{x}^{(i)}$ is sampled from a conditional distribution $p^*(\boldsymbol{x}|\boldsymbol{z})$. Unfortunately, the true distributions as well as the inter-modality dependencies are unknown. The only assumption we make about the distributions is that they are from a parametric family of distributions $p_\theta(\boldsymbol{z})$ and $p_\theta(\boldsymbol{z}|\boldsymbol{x})$. For the case of data sets $X^{(i)}$, the ELBO can be written as follows:

$$\begin{aligned} ELBO(\boldsymbol{X}) &= E_{q_\phi(\boldsymbol{z}|\boldsymbol{X})}[\log p_\theta(\boldsymbol{X}|\boldsymbol{z})] - KL(q_\phi(\boldsymbol{z}|\boldsymbol{X})||p_\theta(\boldsymbol{z})) \qquad (1) \\ ELBO(\boldsymbol{X}_K) &= E_{q_{\phi_K}(\boldsymbol{z}|\boldsymbol{X}_K)}[\log(p_\theta(\boldsymbol{X}|\boldsymbol{z})] - KL(q_{\phi_K}(\boldsymbol{z}|\boldsymbol{X}_K)||p_\theta(\boldsymbol{z})) \qquad (2) \end{aligned}$$

If one or more data types are missing, we would like to be able to approximate the true multimodal posterior distribution $p_\theta(\boldsymbol{z}|\boldsymbol{X})$. However, we are only able to approximate the posterior by a variational function $q_{\phi_K}(\boldsymbol{z}|\boldsymbol{X}_K)$. $\boldsymbol{X}_K$ denotes a subset of $\boldsymbol{X}$ with K available modalities and $K \leq M$. For simplicity, we always use $\boldsymbol{X}_K$ to symbolize missing data, although there is no information about which or how many modalities are missing. Additionally, different modalities might be missing for different samples. From Equation (2) we see that naively implementing a multimodal VAE for M modalities which is able to handle any combination of missing and available data types $\boldsymbol{X}_K$ would result in 2^M different encoder networks. Nevertheless, this is considered a key property of any multimodal system.

If only a subset of modalities is available as described in Equation (2), we would like the representation of the subset of modalities to be close to the true posterior $p(\boldsymbol{z}|\boldsymbol{X})$ but if we restrict ourselves to a model following Equation (1), it cannot handle missing data. Hence, we need a model that is able to handle missing data, while at the same time still approximating the true posterior distribution.

We propose the following:

Lemma 1. *The multimodal ELBO defined in Equation* (2) *can be rewritten using the Jensen-Shannon-divergence for multiple distributions and a dynamic prior:*

$$\widetilde{ELBO}(\boldsymbol{X}_K) = E_{q_{\phi_K}(\boldsymbol{z}|\boldsymbol{X}_K)}[\log p_\theta(\boldsymbol{X}|\boldsymbol{z})] - JS_{\boldsymbol{\pi}}^{K+1}(\{q_{\phi_{\boldsymbol{z}_j}}(\boldsymbol{z}|\boldsymbol{x}_j)\}_{j=1}^{M}, p_0(\boldsymbol{z})) \tag{3}$$

The proof and a detailed derivation can be found in the appendix.

The JS-divergence for multiple distributions allows us to use a prior which is not only the static, standard Gauss distribution $\mathcal{N}(\mathbf{0},\mathrm{I})$, but also the unimodal latent distributions $q_\phi(\boldsymbol{z}|\boldsymbol{x}_j)$ of the available data types. This defines a dynamic prior as it is learned together with the unimodal distributions during training. Lemma 1 allows us to approximate the true multimodal posterior distribution $p_\theta(\boldsymbol{z}|\boldsymbol{X})$ by using the unimodal posterior variational approximations $q_{\phi_j}(\boldsymbol{z}|\boldsymbol{x}_j)$. Using Equation (2) and Lemma 1, we see that any combination of missing and available data types can approximate the true posterior distribution $p_\theta(\boldsymbol{z}|\boldsymbol{X})$ in a scalable way. Additionally, Lemma 1 allows for the connection of the joint distribution $q_\phi(\boldsymbol{z}|\boldsymbol{X})$ to its unimodal latent approximations $q_{\phi_j}(\boldsymbol{z}|\boldsymbol{x}_j)$, which enables us to directly learn the joint as well as the unimodal latent distributions. In previous works, this direct connection is missing which is one of the reasons [12] and [27] have to rely on difficult training procedures. In this work, this direct connection comes at the cost of a less tight $\widetilde{ELBO}$. According to [19], this is not necessarily a drawback.

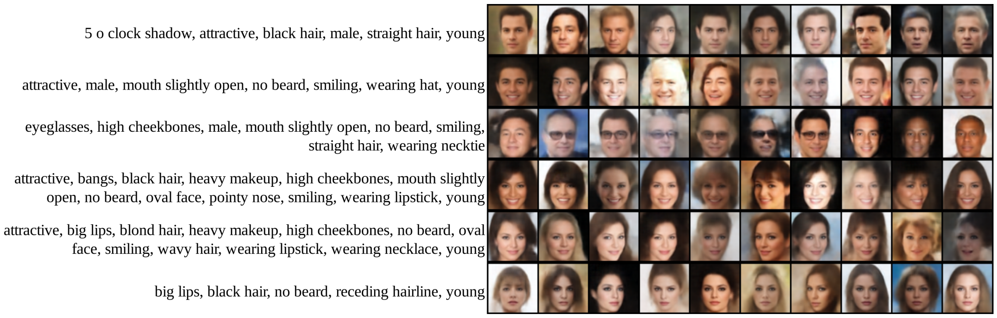
5 o clock shadow, attractive, black hair, male, straight hair, young
attractive, male, mouth slightly open, no beard, smiling, wearing hat, young
eyeglasses, high cheekbones, male, mouth slightly open, no beard, smiling, straight hair, wearing necktie
attractive, bangs, black hair, heavy makeup, high cheekbones, mouth slightly open, no beard, oval face, pointy nose, smiling, wearing lipstick, young
attractive, big lips, blond hair, heavy makeup, high cheekbones, no beard, oval face, smiling, wavy hair, wearing lipstick, wearing necklace, young
big lips, black hair, no beard, receding hairline, young

Figure 1: Examples of generated faces based on text strings using mmJSD. We are able to generate faces of high quality, while keeping some of the attributes like the gender: female faces are sampled if 'male' is not part of the string. The same is true for the smiling attribute. More specific and less dominant attributes like eyeglasses are more difficult to learn, as seen in the third row, where not all faces show a face with eyeglasses.

4 Experiments & Results

We carried out experiments on the dSprites dataset [16] and the CelebA faces dataset [14]. For the two datasets, which both are image-based datasets, we create a second modality by generating text from the labels. Different to previous research, we are using actual text and not just "0-1"-based strings representing the labels. We compare the proposed method to a PoE-based model [27, 12]. For more details about the PoE-model used, we refer the reader to the Appendix.

4.1 dSprites

dSprites, a dataset mostly used for disentangling tasks, offers control over all generating factors. We are interested in the quality of the generated samples if a data type is missing. Evaluating the quality of generated images is not straightforward. [22] report difficulties in evaluating images in terms of their likelihood. Inspired by [25], we hence evaluate the quality by training an additional classifier on the original training split of the dataset which classifies the generated samples according to their attributes. We train a classifier for each modality. Hence, we classify text samples which were generated based on the modality-invariant information inferred by images, and vice versa. The text

Table 1: dSprites Evaluation: Classification accuracy of generated samples and latent represenations. In Table 1a the classification accuracy of conditionally generated samples can be seen, in Table 1b the classification accuracy of a classifier trained on a single batch of latent representations.

(a) Conditionally Generated Samples

Model	Shape		Scale		Orientation		Pos X		Pos Y	
	Img	Text	Img	Text	Img	Text	Img	Text	Img	Text
mmPoE	0.95	0.34	0.17	0.44	0.48	0.25	0.33	0.97	0.33	0.93
mmJSD	**0.96**	**0.97**	**0.92**	**0.89**	**0.50**	0.26	0.33	**0.98**	0.33	**0.99**

(b) Latent Representations

Model		Shape	Scale	Orientation	Pos X	Pos Y
mmPoE	Img	0.38	0.70	0.25	0.991	0.97
	Text	0.34	0.23	**0.81**	0.34	0.43
	Joint	0.38	0.68	**0.79**	0.97	0.96
mmJSD	Img	**1.0**	**1.0**	**0.443**	**0.998**	**0.999**
	Text	**1.0**	**1.0**	0.449	**1.0**	**1.0**
	Joint	**1.0**	**1.0**	0.447	**0.999**	**0.999**

samples are a discrete version of the original values in text form. For details on the generative factors we refer to [16]. The classification results for all attributes can be seen in Table 1a. Our proposed model generates samples which better allow to re-infer the original values for most of the generative factors.

To evaluate the quality of the learnt approximation functions, we train a classifier on their latent representations: on the unimodal as well as the joint latent representations. For training and testing we are using the the test set part of dSprites, which we split again into training and testing set. This way, we ensure that images and text used for this downstream task did not influence the weights of our network during training. We compare the proposed approach (see section 3) again to the PoE approach. The results can be seen in Table 1b: The proposed method achieves similar or better performance than the PoE-based method.

4.2 CelebA

We generate 64x64 images and text of length 256. More details on the architectures and dataset can be found in the Appendix. Figure 1 shows qualitative results for images generated based on text samples. Every row of images was generated based on the information inferred from the modality-invariant information from the text left of each row. From this information, 10 images with randomly sampled image-specific information were generated. Some attributes are more difficult to learn than others. At first, rare attributes like eyeglasses are difficult to learn. Additionally, eyeglasses are a subtle attribute which do not contribute much to the loss during training which makes it even more difficult. A bit surprising, hair-colors seem difficult to learn as well while gender and smiling are well learnt.

5 Conclusion

In this work, we propose a novel generative model for learning with multimodal data. Our contributions are threefold: (i) we formulate a new multimodal ELBO using a dynamic prior, (ii) we propose to use the JS-divergence for multiple distributions as a divergence measure for multimodal datasets. This measure enables direct optimization of the unimodal as well as the joint latent approximation functions, and (iii) we demonstrate that the proposed method does not need any additional training objectives while reaching state-of-the-art or superior performance compared to two recently proposed multimodal generative models.

References

[1] Javed A Aslam and Virgil Pavlu. Query hardness estimation using Jensen-Shannon divergence among multiple scoring functions. In *European conference on information retrieval*, pages 198–209. Springer, 2007.

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.

[3] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[4] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially Learned Inference. pages 1–18, 2016.

[5] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, and Others. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015.

[6] Geoffrey E Hinton. Products of experts. 1999.

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

[8] Wei-Ning Hsu and James Glass. Disentangling by Partitioning: A Representation Learning Framework for Multimodal Sensory Data. 2018.

[9] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[10] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[11] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. (Ml):1–14, 2013.

[12] Richard Kurle, Stephan Günnemann, and Patrick van der Smagt. Multi-Source Neural Variational Inference. 2018.

[13] Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[14] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[15] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[16] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. *URL https://github. com/deepmind/dsprites-dataset/.[Accessed on: 2018-05-08]*, 2017.

[17] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[18] Frank Nielsen. On the Jensen-Shannon symmetrization of distances relying on abstract means. *Entropy*, 2019.

[19] Tom Rainforth, Adam R Kosiorek, Tuan Anh Le, Chris J Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better. *arXiv preprint arXiv:1802.04537*, 2018.

[20] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

[21] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint Multimodal Learning with Deep Generative Models. pages 1–12, 2016.

[22] Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. pages 1–10, 2015.

[23] Jakub M Tomczak and Max Welling. VAE with a VampPrior. *arXiv preprint arXiv:1705.07120*, 2017.

[24] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*, 2018.

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

[26] Weiran Wang, Xinchen Yan, Honglak Lee, and Karen Livescu. Deep Variational Canonical Correlation Analysis. 1, 2016.

[27] Mike Wu and Noah Goodman. Multimodal Generative Models for Scalable Weakly-Supervised Learning. (Nips), 2018.

[28] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

[29] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

A Derivation of proposed method

Lemma 2 (Multimodal Dynamic Prior)**.** *The prior of the multimodal joint latent representation is defined as follows:*

$$p_f(\boldsymbol{z}) = f(q_{\phi_1}(\boldsymbol{z}|\boldsymbol{x}_1), \ldots, q_{\phi_K}(\boldsymbol{z}|\boldsymbol{x}_K), p_0(\boldsymbol{z})) \tag{4}$$

where f is a function of probability distribution, $p_0(z)$ is a standard Gauss distribution of the form $\mathcal{N}(\boldsymbol{0},I)$. $q_\phi(\boldsymbol{z}|\boldsymbol{x}_j)$ are unimodal posterior approximations of the available data types.

Proof. Any probability distribution can be a suitable choice for a prior distribution. The components of the function f are themselves probability distributions. Hence, for $p_f(\boldsymbol{z})$ to be a proper probability distribution, the only additional condition that needs to be satisfied by f is the integration to 1: $\int p_f(\boldsymbol{z})d\boldsymbol{z} = 1$. □

Lemma 3. *If we define $q_\phi(\boldsymbol{z}|\{\boldsymbol{x}_j\}_{j=1}^M)$ as a mixture model of the unimodal variational posterior approximations $q_{\phi_j}(\boldsymbol{z}|\boldsymbol{x}_j)$, the KL-divergence of the multimodal variational posterior approximation $q_\phi(\boldsymbol{z}|\{\boldsymbol{x}_j\}_{j=1}^M)$ is an upper bound for the weighted sum of the KL-divergences of the unimodal variational approximation functions $q_{\phi_j}(\boldsymbol{z}|\boldsymbol{x}_j)$.*

$$\begin{aligned} ELBO(\boldsymbol{X}) &\geq \widetilde{ELBO}(\boldsymbol{X}) \\ &= E_{q_\phi(\boldsymbol{z}|\boldsymbol{X})}[\log p_\theta(\boldsymbol{X}|\boldsymbol{z})] - \sum_{j=1}^{M} \pi_j KL(q_{\phi_j}(\boldsymbol{z}|\boldsymbol{x}_j)||p_f(\boldsymbol{z})) \end{aligned} \tag{5}$$

Proof. Under the assumption

$$q_\phi(\boldsymbol{z}|\{\boldsymbol{x}_j\}_{j=1}^M) := \sum_{j=1}^{M} \pi_j q_{\phi_j}(\boldsymbol{z}|\boldsymbol{x}_j) \tag{6}$$

the KL-divergence in the equation for the multimodal dynamic prior can be written as the sum of KL-divergences, which are less than or equal to the mixture distribution:

$$
\begin{aligned}
KL(q_\phi(\boldsymbol{z}|\{\boldsymbol{x}_j\}_{j=1}^M)||p_f(\boldsymbol{z})) &= KL(\sum_{j=1}^M \pi_j q_{\phi_j}(\boldsymbol{z}|\boldsymbol{x}_j)||p_f(\boldsymbol{z})) \\
&= \int_{\boldsymbol{z}} \left(\sum_{j=1}^M \pi_j q_{\phi_j}(\boldsymbol{z}|\boldsymbol{x}_j) \right) \log \left(\frac{\sum_{k=1}^M \pi_k q_{\phi_k}(\boldsymbol{z}|\boldsymbol{x}_k)}{p_f(\boldsymbol{z})} \right) dz \\
&\geq \sum_{j=1}^M \pi_j \left(\int_{\boldsymbol{z}} q_{\phi_j}(\boldsymbol{z}|\boldsymbol{x}_j) \sum_{k=1}^M \pi_k \log \left(\frac{q_{\phi_k}(\boldsymbol{z}|\boldsymbol{x}_k)}{p_f(\boldsymbol{z})} \right) dz \right) \\
&\geq \sum_{j=1}^M \pi_j \left(\int_{\boldsymbol{z}} q_{\phi_j}(\boldsymbol{z}|\boldsymbol{x}_j) \log \left(\frac{q_{\phi_j}(\boldsymbol{z}|\boldsymbol{x}_j)}{p_f(\boldsymbol{z})} \right) dz \right) \\
&= \sum_{j=1}^M \pi_j KL(q_{\phi_j}(\boldsymbol{z}|\boldsymbol{x}_j)||p_f(\boldsymbol{z})) \qquad (7)
\end{aligned}
$$

Equation (5) follows directly from Equation (7). □

Lemma 4. *The ELBO formulation* $\widetilde{L}(\boldsymbol{X})$ *in Equation* (5) *can be written as a Jensen-Shannon divergence, if f defines a function according to the specification of abstract means [18]:*

$$
\widetilde{ELBO}(\boldsymbol{X}) \quad = \quad E_{q_\phi(\boldsymbol{z}|\boldsymbol{X})}[\log p_\theta(\boldsymbol{X}|\boldsymbol{z})] - JS_{\boldsymbol{\pi}}^{M+1}(\{q_{\phi_{\boldsymbol{z}_j}}(\boldsymbol{z}|\boldsymbol{x}_j)\}_{j=1}^M, p_0(\boldsymbol{z}))
$$

where $JS_{\boldsymbol{\pi}}^{M+1}$ is the Jensen-Shannon divergence for $M+1$ distributions with distribution weights $\boldsymbol{\pi} = [\pi_1, \dots, \pi_{M+1}]$ and $\sum \pi_i = 1$ [13, 18].

Proof. Follows directly from Lemma 3 and the definition of the general JS-divergence for N distributions [13] and the definition of abstract means [18]. □

A.1 Practical Implementations

A.1.1 Generalized Jensen-Shannon Divergence

[**?**] generalized the Jensen-Shannon divergence to the definition of abstract means. Abstract means are a suitable class of functions for aggregating information from different distributions while being able to handle missing distributions. In the special case of a geometric mean, there exists a closed form solution for $JS_{\boldsymbol{\pi}}^N$, if all involved distributions are Gaussian. For more details see [**?**].

Following the common line of VAE research, the unimodal latent distributions q_{ϕ_j} are assumed to be Gaussian distributed $q_{\phi_j}(\boldsymbol{z}|\boldsymbol{x}_j) = \mathcal{N}(\boldsymbol{\mu}_j(\boldsymbol{x}_j), \boldsymbol{\Sigma}_j(\boldsymbol{x}_j))$. The mean distribution can be written as:

$$
\begin{aligned}
\boldsymbol{\Sigma}_K = &\quad (\textstyle\sum_{k=1}^K \pi_k \boldsymbol{\Sigma}_k)^{-1} \qquad (8) \\
\boldsymbol{\mu}_K = &\quad \boldsymbol{\Sigma}_K \textstyle\sum_{k=1}^K \pi_k \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k \qquad (9)
\end{aligned}
$$

For Gaussian functions, the geometric mean distribution in Equation (8) can be seen as a weighted PoE [6, 7]. Different to previous work using the PoE in a multimodal setting [**? ?** 24], in this work the PoE is part of the dynamic prior distribution p_f. Our formulation of the ELBO in Equation (8) allows the optimization of the individual representations $q_j(\boldsymbol{z}|\boldsymbol{x}_j)$, as well as the joint representation which is equal to the dynamic prior.

A.1.2 Factorization of Representations

We mostly base our derivation on the paper of [3]. [24] and [8] used a similar idea. We define a set $\boldsymbol{X}^{(i)}$ of modalities as group and analogous every modality as a sample of a group. We model every $\boldsymbol{x}_j$ to have its own modality-specific latent code $\boldsymbol{s}_j \in \boldsymbol{S}$.

$$\boldsymbol{S} = (\boldsymbol{s}_j, \forall \boldsymbol{x}_j \in \boldsymbol{X}^{(i)}) \tag{10}$$

From Equation (10), we see that $\boldsymbol{S}$ is the collection of all modality-specific latent variables for the set $\boldsymbol{X}^{(i)}$. Contrary to this, the modality-invariant latent code $\boldsymbol{c}$ is shared between all modalities $\boldsymbol{x}_j$ of the set $\boldsymbol{X}^{(i)}$. $\boldsymbol{c}$ takes the role of the multimodal latent code $\boldsymbol{z}$ in the Theory section in the main paper. Also like [3], we model the variational approximation function $q_\phi(\boldsymbol{S}, \boldsymbol{c})$ such that it decomposes into a modality-dependent and -invariant part:

$$q_\phi(\boldsymbol{S}, \boldsymbol{c}) = q_{\phi_{\boldsymbol{S}}}(\boldsymbol{S}|\boldsymbol{X}) q_{\phi_{\boldsymbol{c}}}(\boldsymbol{c}|\boldsymbol{X}) \tag{11}$$

Additionally, we model q_{ϕ_S} such that it factorizes among the different modalities:

$$q_{\phi_{\boldsymbol{S}}}(\boldsymbol{S}|\boldsymbol{X}) = \prod_{j=1}^{M} q_{\phi_{\boldsymbol{s}_j}}(\boldsymbol{s}_j|\boldsymbol{x}_j) \tag{12}$$

From Equation (12) and from the fact that the multimodal relationships are only modelled by the latent factor $\boldsymbol{c}$, it is reasonable to only apply the multimodal modelling from the Theory section to $\boldsymbol{c}$. It follows:

$$\begin{aligned} ELBO(\boldsymbol{X}) =& E_{q_\phi(\boldsymbol{z}|\boldsymbol{X})}[\log p_\theta(\boldsymbol{X}|\boldsymbol{z})] - KL(q_\phi(\boldsymbol{z}|\boldsymbol{X})||p_\theta(\boldsymbol{z})) \\ =& E_{q_\phi(\boldsymbol{S},\boldsymbol{c}|\boldsymbol{X})}[\log p_\theta(\boldsymbol{X}|\boldsymbol{S},\boldsymbol{c})] - KL(q_\phi(\boldsymbol{S},\boldsymbol{c}|\boldsymbol{X})||p_\theta(\boldsymbol{S},\boldsymbol{c})) \\ =& E_{q_\phi(\boldsymbol{S},\boldsymbol{c}|\boldsymbol{X})}[\log p_\theta(\boldsymbol{X}|\boldsymbol{S},\boldsymbol{c})] - KL(q_{\phi_{\boldsymbol{S}}}(\boldsymbol{S}|\boldsymbol{X})||p_\theta(\boldsymbol{S})) - KL(q_{\phi_{\boldsymbol{c}}}(\boldsymbol{c}|\boldsymbol{X})||p_\theta(\boldsymbol{c})) \\ =& E_{q_\phi(\boldsymbol{S},\boldsymbol{c}|\boldsymbol{X})}[\log p_\theta(\boldsymbol{X}|\boldsymbol{S},\boldsymbol{c})] - \sum_{j=1}^{M} KL(q_{\phi_{\boldsymbol{s}_j}}(\boldsymbol{s}_j|\boldsymbol{x}_j)||p_\theta(\boldsymbol{s}_j)) - KL(q_{\phi_{\boldsymbol{c}}}(\boldsymbol{c}|\boldsymbol{X})||p_\theta(\boldsymbol{c})) \end{aligned} \tag{13}$$

In Equation (16), we can rewrite the KL-divergence which includes $\boldsymbol{c}$ using the multimodal dynamic prior and the JS-divergence for multiple distributions:

$$\begin{aligned} \widetilde{ELBO}(\boldsymbol{X}) =& E_{q_\phi(\boldsymbol{S},\boldsymbol{c}|\boldsymbol{X})}[\log p_\theta(\boldsymbol{X}|\boldsymbol{S},\boldsymbol{c})] \\ & - \sum_{j=1}^{M} KL(q_\phi(\boldsymbol{s}_j|\boldsymbol{x}_j)||p_\theta(\boldsymbol{s}_j)) - JS_{\boldsymbol{\pi}}^{M+1}(\{q_{\phi_{\boldsymbol{c}_j}}(\boldsymbol{c}|\boldsymbol{x}_j)\}_{j=1}^{M}, p_0(\boldsymbol{c})) \end{aligned} \tag{14}$$

The expectation over $q_\phi(\boldsymbol{S}, \boldsymbol{c}|\boldsymbol{X})$ can be rewritten as an expectation over $q_{\phi_{\boldsymbol{c}}}(\boldsymbol{X}|\boldsymbol{c})$ as an expectation over $q_{\phi_{\boldsymbol{s}_j}}$:

$$
\begin{aligned}
E_{q_{\phi}(\boldsymbol{S},\boldsymbol{c}|\boldsymbol{X})}[\log p_{\theta}(\boldsymbol{X}|\boldsymbol{S},\boldsymbol{c})] &= \int_{\boldsymbol{c}}\int_{\boldsymbol{S}} q_{\phi}(\boldsymbol{S},\boldsymbol{c}|\boldsymbol{X}) \log p_{\theta}(\boldsymbol{X}|\boldsymbol{S},\boldsymbol{c}) d\boldsymbol{S} d\boldsymbol{c} \\
&= \int_{\boldsymbol{c}} q_{\phi_{\boldsymbol{c}}}(\boldsymbol{c}|\boldsymbol{X}) \int_{\boldsymbol{S}} q_{\phi_{\boldsymbol{S}}}(\boldsymbol{S}|\boldsymbol{X}) \log p_{\theta}(\boldsymbol{X}|\boldsymbol{S},\boldsymbol{c}) d\boldsymbol{S} d\boldsymbol{c} \\
&= \int_{\boldsymbol{c}} q_{\phi_{\boldsymbol{c}}}(\boldsymbol{c}|\boldsymbol{X}) \sum_{j=1}^{M} \int_{\boldsymbol{s}_j} q_{\phi_{\boldsymbol{s}_j}}(\boldsymbol{s}_j|\boldsymbol{x}_j) \log p_{\theta}(\boldsymbol{x}_j|\boldsymbol{s}_j,\boldsymbol{c}) d\boldsymbol{s}_j d\boldsymbol{c} \\
&= \sum_{j=1}^{M} \int_{\boldsymbol{c}} q_{\phi_{\boldsymbol{c}}}(\boldsymbol{c}|\boldsymbol{X}) \int_{\boldsymbol{s}_j} q_{\phi_{\boldsymbol{s}_j}}(\boldsymbol{s}_j|\boldsymbol{x}_j) \log p_{\theta}(\boldsymbol{x}_j|\boldsymbol{s}_j,\boldsymbol{c}) d\boldsymbol{s}_j d\boldsymbol{c} \\
&= \sum_{j=1}^{M} E_{q_{\phi_{\boldsymbol{c}}}(\boldsymbol{c}|\boldsymbol{X})}[E_{q_{\phi_{\boldsymbol{s}_j}}(\boldsymbol{s}_j|\boldsymbol{x}_j)}[\log p_{\theta}(\boldsymbol{x}_j|\boldsymbol{s}_j,\boldsymbol{c})]] \qquad (15)
\end{aligned}
$$

From Equation (15), the final form from the paper follows directly:

$$
\begin{aligned}
\widetilde{ELBO}(\boldsymbol{X}) = &\sum_{j=1}^{M} E_{q_{\phi_{\boldsymbol{c}}}(\boldsymbol{c}|\boldsymbol{X})}[E_{q_{\phi_{\boldsymbol{s}_j}}(\boldsymbol{s}_j|\boldsymbol{x}_j)}[\log p_{\theta}(\boldsymbol{x}_j|\boldsymbol{s}_j,\boldsymbol{c})]] \\
&- \sum_{j=1}^{M} KL(q_{\phi_{\boldsymbol{s}_j}}(\boldsymbol{s}_j|\boldsymbol{x}_j)||p_{\theta}(\boldsymbol{s}_j)) \qquad (16) \\
&- JS_{\boldsymbol{\pi}}^{M+1}(\{q_{\phi_{\boldsymbol{c}_j}}(\boldsymbol{c}|\boldsymbol{x}_j)\}_{j=1}^{M}, p_0(\boldsymbol{c}))
\end{aligned}
$$

B Data

*square,
zero point
eight, left, upper
mid*************
****

(a) dSprites image

(b) dSprites text

Figure 2: Example of a data pair for the MNIST data set. From the labels, we generate strings of length 32 consisting of the number as word and blank spaces. The starting position of the word in this 32 length string is chosen randomly. For dSprites, we generate strings of length 64 consisting of the discretized generative attributes and fill the remaining space with $*$.

B.1 dSprites

To have a bimodal dataset, we created text from the generative factors. The shape attribute is just converted to a string (square, ellipse, heart). The scale factor is digit-wise converted into words. E.g. a scale factor of 0.5 becomes zero point five. For the orientation attribute, we create 4 text realizations from the 40 values between 0 and 2π: upright, left, upside down, right. The same applies to the position in both directions which originally is a value between 0 and 1. We divide the x-direction into left, mid, right and the y-direction into lower, mid, upper. We concatenate the strings from the attributes into a comma-separated list. We always create strings of length 64. As the atribute string is not of length 64, we fill the remaining space by $*$ and choose a random starting position. A sample pair from this dataset can be seen in Figure 2.

B.2 CelebA

Every face in the dataset is labelled with 40 attributes. To generate the second modality, we create text strings from these attributes. All the attribute names are concatenated in a comma-separated list if the attribute is present in a face. Underline characters are replaced by a blank space. We create strings of length 256. If a given face has only a small number of attributes which would result in a short string, we fill the remaining space with the asterix character $*$.

C Experiments

C.1 PoE-based Model

The PoE-based model used for the MNIST and dSprites experiments is adopted from [27, 12]. They both use a combination of unimodal and joint divergence terms where the joint representations is the output of a PoE. [27] uses a subsampling of all datatypes if more than two datatypes are available. The resulting training objective can be written as follows:

$$
\begin{align}
L(\boldsymbol{X}) =& \sum_{j=1}^{M} E_{q_{\phi}(\boldsymbol{z}|\boldsymbol{X})}[\log p_{\theta}(\boldsymbol{x}_j|\boldsymbol{z})] \notag \\
&- \sum_{j=1}^{M} KL(q_{\phi_j}(\boldsymbol{z}|\boldsymbol{x}_j)||p_{\theta}(\boldsymbol{z})) - KL(PoE(\{q_{\phi_j}(\boldsymbol{z}|\boldsymbol{x}_j)\}_{j=1}^{M})||p_{\theta}(\boldsymbol{z})) \tag{17}
\end{align}
$$

In case of a factorized representation like we use it, Equation (18) changes as follows:

$$
\begin{align}
L(\boldsymbol{X}) =& \sum_{j=1}^{M} E_{q_{\phi_{\boldsymbol{c}}}(\boldsymbol{c}|\boldsymbol{X})}[E_{q_{\phi_{\boldsymbol{s}_j}}(\boldsymbol{s}_j|\boldsymbol{x}_j)}[\log p_{\theta}(\boldsymbol{x}_j|\boldsymbol{s}_j, \boldsymbol{c})]] \notag \\
&- \sum_{j=1}^{M} KL(q_{\phi_{\boldsymbol{s}_j}}(\boldsymbol{s}_j|\boldsymbol{x}_j)||p_{\theta}(\boldsymbol{s}_j)) \tag{18} \\
&- \sum_{j=1}^{M} KL(q_{\phi_{\boldsymbol{c}}}(\boldsymbol{c}|\boldsymbol{x}_j)||p_{\theta}(\boldsymbol{c})) \tag{19} \\
&- KL(PoE(\{q_{\phi_j}(\boldsymbol{z}|\boldsymbol{x}_j)\}_{j=1}^{M})||p_{\theta}(\boldsymbol{z})) \notag
\end{align}
$$