Mixture-of-experts VAEs can disregard variation in surjective multimodal data

Jannik Wolff[*†] TU Berlin
Tassilo Klein, Moin Nabi SAP AI Research
Rahul G. Krishnan[‡] University of Toronto
Shinichi Nakajima TU Berlin

Abstract

Machine learning systems are often deployed in domains that entail data from multiple modalities, for example, phenotypic and genotypic characteristics describe patients in healthcare. Previous works have developed multimodal variational autoencoders (VAEs) that generate several modalities. We consider surjective data, where single datapoints from one modality (such as class labels) describe multiple datapoints from another modality (such as images). We theoretically and empirically demonstrate that multimodal VAEs with a mixture of experts posterior can struggle to capture variability in such surjective data.

1 Introduction

Many datasets entail a surjective mapping between modalities (Fig. 1, *"one-to-many data"*). That is, an instance from one modality may correspond to several instances from another modality. For example, many computer vision datasets contain labels, attributes, or text data that describe sets of images [LeCun, 1998, Nilsback and Zisserman, 2008, Krizhevsky et al., 2009, Deng et al., 2009, Wah et al., 2011, Liu et al., 2015, Xiao et al., 2017]. Note that *"one-to-one data"* such as image/caption pairs can become surjective when using data augmentation, e.g., random horizontal flipping of images. Incorporating further modalities can also invoke surjectivity.

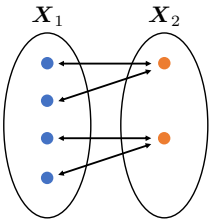

Figure 1: ***Surjective* data.** X_1 and X_2 depict exemplary modalities. The mapping from the second to the first modality is surjective.

Multimodal VAEs maximize a bound on the joint density of several modalities and can thereby learn to generate any modality from any conditioning modality [Suzuki et al., 2016]. For some multimodal VAEs, this bound contains a factor that represents the likelihood of one modality given another modality. We will show that such a factor in the objective function can lead to solutions that disregard heterogeneity within a modality. For example, we demonstrate that samples from models with a mixture of experts posterior such as the MMVAE [Shi et al., 2019] can have a bias towards the class mean of the observed datapoints for a given modality.

2 Method

Let $\boldsymbol{X} = \{\{\boldsymbol{x}_m^{(n)}\}_{m=1}^M\}_{n=1}^N$ be a training set with several modalities, where m and n represent the modality and the sample index, respectively. We consider a multimodal VAE with a generative model

$$\begin{aligned} \boldsymbol{g} &\sim p_\theta(\boldsymbol{g}), \\ \boldsymbol{x}_m &\sim p_\theta(\boldsymbol{x}_m|\boldsymbol{g}) \qquad \text{for } m = 1, \ldots, M, \end{aligned} \tag{1}$$

[*]Correspondence to: wolff.jannik@icloud.com
[†]Part of the work was done at SAP AI Research.
[‡]Part of the work was done at Massachusetts Institute of Technology and Microsoft Research.

Bayesian Deep Learning workshop, NeurIPS 2021.

and an inference model

$$\boldsymbol{g} \sim q_\phi(\boldsymbol{g}|\{\boldsymbol{x}_m\}_{m=1}^M). \tag{2}$$

Assume that the generative model (1) is a parametric model, e.g., Gaussian,

$$p_\theta(\boldsymbol{x}_m|\boldsymbol{g}) = f_m(\boldsymbol{x}_m|\boldsymbol{\tau}_m(\boldsymbol{g};\boldsymbol{\theta})), \tag{3}$$

with the parameters $\{\boldsymbol{\tau}_m\}$, e.g., means and covariances, defined as a function of $\boldsymbol{g}$ and (typically) neural networks weights $\boldsymbol{\theta}$. Assume that the inference model (2) is defined as a finite mixture with parameters $\boldsymbol{\kappa}_m$ indicating mean and covariance for mixture component r_m (as in the MMVAE [Shi et al., 2019], for example):

$$q_\phi(\boldsymbol{g}|\{\boldsymbol{x}_m\}_{m=1}^M) = \frac{1}{M}\sum_{m=1}^M q_\phi(\boldsymbol{g}|\boldsymbol{x}_m) = \frac{1}{M}\sum_{m=1}^M r_m(\boldsymbol{g}|\boldsymbol{\kappa}_m(\boldsymbol{x}_m;\boldsymbol{\phi})).$$

Without loss of generality, we assume that $\boldsymbol{x}_M$ is the label modality, and let $\boldsymbol{S}_c = \{n \mid \boldsymbol{x}_M^{(n)} = c\}$ be the set of indices of the samples belonging to the label $c \in \{1, \ldots, C\}$. We consider a maximization problem given the following objective function:

$$L_m(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{X}) \equiv \sum_{n=1}^N \int r_M(\boldsymbol{g}|\boldsymbol{\kappa}_M(\boldsymbol{x}_M^{(n)};\boldsymbol{\phi})) \log f_m(\boldsymbol{x}_m^{(n)}|\boldsymbol{\tau}_m(\boldsymbol{g};\boldsymbol{\theta}))d\boldsymbol{g}, \tag{4}$$

which is an ELBO for

$$\log p(\boldsymbol{x}_m|\boldsymbol{x}_M) = \log \int q_\phi(\boldsymbol{g}|\boldsymbol{x}_M)p_\theta(\boldsymbol{x}_m|\boldsymbol{g})d\boldsymbol{g} \geq \int q_\phi(\boldsymbol{g}|\boldsymbol{x}_M)\log p_\theta(\boldsymbol{x}_m|\boldsymbol{g})d\boldsymbol{g} = L_m(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{X}).$$

Importantly, the MMVAE [Shi et al., 2019] relies on term (4) for learning data translation ability from $\boldsymbol{x}_M$ to $\boldsymbol{x}_m$. Specifically, the authors used stratified sampling for training[4], which implies that Eq. 4 and term ① from Eq. 5 are related:

$$\begin{aligned}
&\log p_\theta(\{\boldsymbol{x}_m\}_{m=1}^M) \geq \frac{1}{M}\sum_{m=1}^M \mathbb{E}_{q_\phi(\boldsymbol{g}|\boldsymbol{x}_m)}\big[\log \frac{p_\theta(\boldsymbol{g}, \{\boldsymbol{x}_m\}_{m=1}^M)}{q_\phi(\boldsymbol{g}|\{\boldsymbol{x}_m\}_{m=1}^M)}\big] \\
=&\frac{1}{M}\Bigg(\sum_{m=1}^{M-1}\Bigg(\mathbb{E}_{q_\phi(\boldsymbol{g}|\boldsymbol{x}_m)}\big[\log \frac{p_\theta(\boldsymbol{g}, \{\boldsymbol{x}_m\}_{m=1}^M)}{q_\phi(\boldsymbol{g}|\{\boldsymbol{x}_m\}_{m=1}^M))}\big]\Bigg) + \mathbb{E}_{q_\phi(\boldsymbol{g}|\boldsymbol{x}_M)}\big[\log \frac{p_\theta(\boldsymbol{g})}{q_\phi(\boldsymbol{g}|\{\boldsymbol{x}_m\}_{m=1}^M)}\big] \\
&+\sum_{i=1}^M \underbrace{\mathbb{E}_{q_\phi(\boldsymbol{g}|\boldsymbol{x}_M)}\big[\log p_\theta(\boldsymbol{x}_i|\boldsymbol{g})\big]}_{①}\Bigg)
\end{aligned} \tag{5}$$

The following theorem holds:

Theorem 1. *Assume a training set $X = \{\boldsymbol{x}_m^{(n)}\}_{n\in S_c}$ which belong to the same label, i.e., $\boldsymbol{x}_M^{(n)} = c, \forall n \in S_c$, and there exists $\widehat{\boldsymbol{\theta}}$ such that $\boldsymbol{\tau}_m(\boldsymbol{g};\widehat{\boldsymbol{\theta}})$ is a constant with respect to $\boldsymbol{g}$ and the maximum likelihood estimator of the parametric model $f_m(\boldsymbol{x}_m|\boldsymbol{\tau}_m(\boldsymbol{g};\boldsymbol{\theta}))$ for the training data. Then, for any $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, it holds that*

$$L_m(\widehat{\boldsymbol{\theta}}, \boldsymbol{\phi}; \boldsymbol{X}) \geq L_m(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{X}). \tag{6}$$

(Proof) Since we assume that $\boldsymbol{x}_M^{(n)} = c$ for all $n \in \boldsymbol{S}_c$, the inferred distribution for $\boldsymbol{g}$ is the same for all n, i.e., $\widetilde{r}_M(\boldsymbol{g}) = r_M(\boldsymbol{g}|\boldsymbol{\kappa}_M(\boldsymbol{x}_M^{(n)};\boldsymbol{\phi}))$. For any such inference model $\widetilde{r}_M(\boldsymbol{g})$, the objective is

[4]Moving Σ_m into the log in Eq. 5 would imply a tighter bound. However, the model may then weigh the experts differently w.r.t. to their gradients, which can disproportionally favor the representation of single modalities at the expense of learning structure across all modalities.

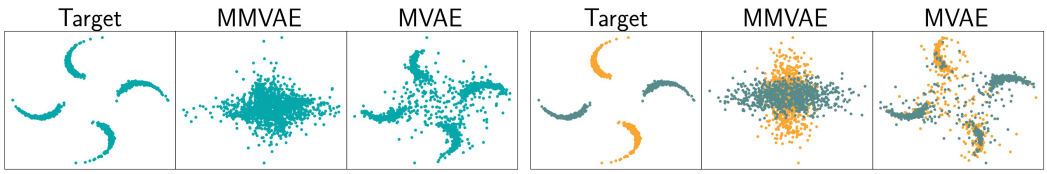

Figure 2: **Generated samples for the first modality.** Left: using samples from $p(\boldsymbol{g})$. Right: using samples from $q(\boldsymbol{g}|\boldsymbol{x}_2)$, where $\boldsymbol{x}_2$ are class labels (yellow or green).

upper-bounded by

$$L_m(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{X}) = \int \widetilde{r}_M(\boldsymbol{g}; \boldsymbol{\phi}) \left(\sum_{n=1}^{N} \log f_m(\boldsymbol{x}_m^{(n)} | \boldsymbol{\tau}_m(\boldsymbol{g}; \boldsymbol{\theta})) \right) d\boldsymbol{g} \tag{7}$$
$$\leq \int \widetilde{r}_M(\boldsymbol{g}; \boldsymbol{\phi}) \left(\sum_{n=1}^{N} \log f_m(\boldsymbol{x}_m^{(n)} | \widehat{\boldsymbol{\tau}}_m) \right) d\boldsymbol{g}$$

with the maximum likelihood estimator $\widehat{\boldsymbol{\tau}}_m$ for the parametric model f_m given the training set $\{\boldsymbol{x}_m^{(n)}\}_{n=S_c}$. The assumed existence of $\widehat{\boldsymbol{\theta}}$ such that $\boldsymbol{\tau}_m(\boldsymbol{g}; \widehat{\boldsymbol{\theta}}) = \widehat{\boldsymbol{\tau}}_m$ leads to Eq. (6). □

Intuitively, consider a single class: $c \in \{1\}$. Let $p_\theta(\boldsymbol{x}_m|\boldsymbol{g})$ be Gaussian with diagonal covariance, where $\boldsymbol{g} \sim q_\phi(\boldsymbol{g}|\boldsymbol{x}_M)$. Theorem 1 implies the existence of an upper bound where the mean parameter from $p_\theta(\boldsymbol{x}_m|\boldsymbol{g})$ always coincides with the mean from $\{\boldsymbol{x}_m^n\}_{n \in S_c}$ for any $\boldsymbol{g}$. This solution is invariant to $\boldsymbol{g}$ because $\boldsymbol{x}_M$ does not carry information about across-datapoint variability in $\boldsymbol{x}_m$. In other words, the solution maximizes the likelihood of the training data $\{\boldsymbol{x}_m^{(n)}\}_{n=S_c}$ with a single Gaussian distribution. That is, the mean parameter minimizes the distance to all datapoints from modality m simultaneously: the model captures the mean of the target distribution – not its variability.

3 Experiments

We create a synthetic dataset (inspired by Johnson et al. [2016]) with modality $\boldsymbol{x}_1 \in \mathbb{R}^2$ and label modality $\boldsymbol{x}_2 \in \{0, 1\}$. We implement the MVAE [Wu and Goodman, 2018] and MMVAE [Shi et al., 2019]. The latent distributions are isotropic Gaussian. The generative distributions are isotropic Gaussian for the first modality and categorical for the second modality.

For the MMVAE, Fig. 2 supports our argument that samples for the first modality tend towards the mean of the observed datapoints (for the same class). The MVAE does not suffer from this problem, possibly because the MVAE's objective function does not contain the factor $p(\boldsymbol{x}_1|\boldsymbol{x}_2)$ (App. A). App. B visualizes the latent spaces, which are two-dimensional to avoid possible obfuscation from dimensionality-reduction techniques.

4 Conclusion

We show that multimodal VAEs with a mixture posterior can struggle to capture heterogeneity in surjective data. This finding implies that practitioners should closely consider the type of data when training such models: for example, data augmentation may not be beneficial since this procedure often promotes surjectivity. Future work may investigate possible solutions, e.g., by considering models that do not maximize $p(\boldsymbol{x}_m|\boldsymbol{x}_{M \neq m})$ explicitly. It would be interesting to analyze how such a solution affects robustness.

Acknowledgements

SN is supported by the German Ministry for Education and Research as BIFOLD - Berlin Institute for the Foundations of Learning and Data (ref. 01IS18025A and ref. 01IS18037A). RGK was supported by a grant from SAP Corporation.

References

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

M. J. Johnson, D. K. Duvenaud, A. Wiltschko, R. P. Adams, and S. R. Datta. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.

A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Y. LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

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

M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

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

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

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

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

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

A Theorem 1 does not apply to the MVAE

The MVAE [Wu and Goodman, 2018] employs a product posterior inspired by the true posterior:

$$q_\phi(\boldsymbol{g}|\{\boldsymbol{x}_m\}_{m=1}^M) \propto p_\theta(\boldsymbol{g}) \prod_{m=1}^{M} q_\phi(\boldsymbol{g}|\boldsymbol{x}_m). \tag{8}$$

In our experiments from § 3, we follow Wu and Goodman [2018] and maximize the following three ELBOs:

$$L(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{X}) \coloneqq ELBO(\boldsymbol{x}_1, \boldsymbol{x}_2) + ELBO(\boldsymbol{x}_1) + ELBO(\boldsymbol{x}_2) \tag{9}$$

The ELBO for M modalities is defined as:

$$\begin{aligned} ELBO(\{\boldsymbol{x}_m\}_{m=1}^M) &\coloneqq \mathbb{E}_{q_\phi(\boldsymbol{g}|\{\boldsymbol{x}_m\}_{m=1}^M)} \left[\log \frac{p_\theta(\boldsymbol{g})}{q_\phi(\boldsymbol{g}|\{\boldsymbol{x}_m\}_{m=1}^M)}\right] + \sum_{m=1}^{M} \mathbb{E}_{q_\phi(\boldsymbol{g}|\{\boldsymbol{x}_m\}_{m=1}^M)} \left[\log p_\theta(\boldsymbol{x}_m|\boldsymbol{g})\right] \\ &\leq \log p_\theta(\{\boldsymbol{x}_m\}_{m=1}^M), \end{aligned} \tag{10}$$

Therefore, $p_\theta(\boldsymbol{x}_m|\boldsymbol{g})$ is always conditioned on $\boldsymbol{x}_m$ via the importance distribution, i.e., the model learns $p(\boldsymbol{x}_m|\{\boldsymbol{x}_i\}_{i=1}^M)$ or $p(\boldsymbol{x}_m|\boldsymbol{x}_m)$. This implies that the MVAE does not explicitly optimize $p(\boldsymbol{x}_{m\neq M}|\boldsymbol{x}_M)$ for any $m \neq M$, i.e., Theorem 1 does not apply to the MVAE.

B Additional experimental results

The solution $q(\boldsymbol{g}|\boldsymbol{x}_1) = q(\boldsymbol{g}|\boldsymbol{x}_2)$ can be helpful because it implies that samples from either posterior produce the same generative distribution for any modality. Figure 3 indicates that the MVAE aligns these marginal posteriors better than the MMVAE, which possibly explains the MVAE's better generative capability in Fig. 2. Figure 2 further exposes that even the MVAE struggles to represent the data perfectly. Its latent representations from Fig. 3 reveal that the model produces some overlap between the class manifolds of the marginal posteriors for the second modality – possibly in an attempt to fit the isotropic Gaussian prior $p(\boldsymbol{g})$. We assume that this struggle is caused by the fact that there are just two unique label datapoints.

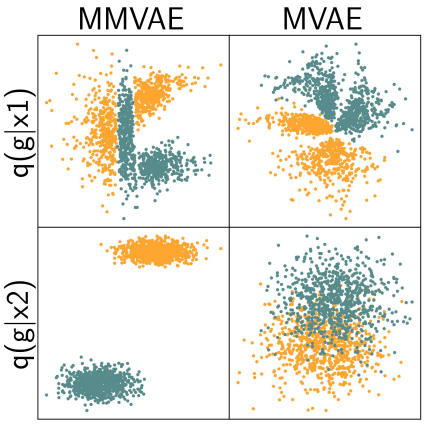

Figure 3: **Marginal posteriors over the latent variable $\boldsymbol{g}$.**