Multimodal Relational VAE

Thomas M. Sutter, Julia E. Vogt
Department of Computer Science
ETH Zurich
{thomas.sutter, julia.vogt}@inf.ethz.ch

Abstract

In this work, we propose a new formulation for multimodal VAEs to model and learn the relationship between data types. Despite their recent progress, current multimodal generative methods are based on simplistic assumptions regarding the relation between data types, which leads to a trade-off between coherence and quality of generated samples - even for simple toy datasets. The proposed method learns the relationship between data types instead of relying on pre-defined and limiting assumptions. Based on the principles of variational inference, we change the posterior approximation to explicitly include information about the relation between data types. We show empirically that the simplified assumption of a single shared latent space leads to inferior performance for a dataset with additional pairwise shared information.

1 Introduction

For weakly-supervised multimodal data, generative models such as Variational Autoencoders (VAE) are successful approaches to learn meaningful representations as well as conditional generation of missing data types [Wu and Goodman, 2018, Shi et al., 2019, Sutter et al., 2021]. While previous works show promising results, they also share a simplistic model based on a single joint latent space for the relation between data types. As a consequence, information is averaged across all dimensions of all modalities in the latent space which results in a conceptually suboptimal and limited model. This becomes particularly problematic if information is shared only pairwise or shared information is unevenly distributed between modalities. Over-restrictive assumptions on these relations may lead to inferior generative performance, i.e. coherence and quality of generative samples. However, enabling less restrictive relations comes with new challenges, as the structure of inter-dependencies between data types is complicated in general and not known a priori. Additionally, the number of possible relations increases exponentially with the number of modalities and ground truth on the true relation between data types is difficult to collect for real-world datasets.

We propose a new framework which is able to learn and automatically model any relation between data types. For a dataset of multiple data types, information can be specific to a single data type, shared between a subset of data types or even shared between all data types. We model these subsets of shared information with additional latent subspaces where the existence and dimensionality of subspaces is not known a priori. Hence, we need to infer the size of every subspace and select latent factors accordingly. The straightforward way would be to marginalize out all combinations of subspace sizes and subset selections of latent variables. However, this is computationally infeasible due to its combinatorial nature.

We introduce a hierarchical VAE to learn the relation between data types, which approximates the additional intractable distributions with a variational inference procedure as well. The additional level of hierarchy not only allows the modelling of any relation between data types but also precise aggregation between latent factors which encode the same information.

Preprint. Under review.

2 Notation

We consider a dataset $\{\boldsymbol{X}^{(i)}\}_{i=1}^{N}$ of N i.i.d. samples, each of which is a set of M data types $\{\boldsymbol{x}_j^{(i)}\}_{j=1}^{M}$[1]. Let $\boldsymbol{X}_A \in \mathcal{P}(\boldsymbol{X})$ define a subset of modalities, where $\mathcal{P}$ denotes the powerset and $A \subseteq \mathcal{M} = \{1, \dots, M\}$ a set of indices defining the respective subset $\boldsymbol{X}_A$. Every subspace S_A corresponds to a latent subspace that contains the set of generative factors shared by modalities $\boldsymbol{x}_j \in \boldsymbol{X}_A$. A latent variable $\boldsymbol{z}_j$ belongs to each modality $\boldsymbol{x}_j$ and every $\boldsymbol{z}_j$ consists of D_j factors. For every subspace S_A, n_A latent factors are selected from each modality $\boldsymbol{x}_j \in \boldsymbol{X}_A$. The number of elements of a subspace n_A takes values in $\{0, \dots, |S_A|\}$ where $|S_A| = \min_{j \in A} D_j$ and is the same for all modalities $\boldsymbol{x}_j \in \boldsymbol{X}_A$. For every modality $\boldsymbol{x}_j$, the selection of latent factors $\boldsymbol{y}_j$ is the assignment of latent factors $z_j^{(k)}, k \in \{1, \dots, D_j\}$ to one of the subspaces $S_A, A \ni j$. We denote the set of all n_A as $\boldsymbol{n}$, i.e. $\boldsymbol{n} = \{n_A\}_{A \subseteq \mathcal{M}}$, and respectively, the set of all $\boldsymbol{y}_j$ as $\boldsymbol{y} = \{\boldsymbol{y}_j\}_{j \in \mathcal{M}}$ and $\boldsymbol{z} = \{\boldsymbol{z}_j\}_{j \in \mathcal{M}}$.

3 Method

Inferring subspaces requires the estimation of the dimensionality per subspace and the selection of the respective amount of latent factors from every modality contributing to this space. Hence, the posterior approximation $q_\phi(\boldsymbol{z})$ becomes $q_\phi(\boldsymbol{z}, \boldsymbol{n}, \boldsymbol{y})$ where $q_\phi(\boldsymbol{n})$ describes the distribution over the dimensionality for all subspaces and $q_\phi(\boldsymbol{y})$ describes the distribution over the selection of latent factors per subspace for all modalities.

3.1 Dimensionality of subspaces

Definition 1 (Probability distribution over subspace dimensions)**.** *For every modality $\boldsymbol{x}_j$, we define the prior distribution $q_\phi(\boldsymbol{n}_j)$, $\boldsymbol{n}_j = \{n_A | A \ni j\}$ over the set of number of elements n_A per subspace S_A contributing to modality $\boldsymbol{x}_j$ as a central multivariate hypergeometric distribution. The posterior distribution $q_\phi(\boldsymbol{n}_j | \boldsymbol{z})$ follows a non-central multivariate hypergeometric distribution with weights $\boldsymbol{\omega}_j = \{\omega_{j,A} | A \ni j\}$.*

The number of elements n to draw is equal to the modality's latent space size, i.e. $n = D_j$. The total number of elements N to be drawn from is equal to the sum of maximum number of elements per subspace S_A, i.e. $N = \sum_{A \ni j} |S_A|$.

The hypergeometric distribution [Upton and Cook, 2014] offers a way to formalize the inference of subspace dimensionality constrained to limited resources. Different to the non-central case, the prior probability on the number of factors to be chosen per subspace only depends on the maximum dimensionality of every subspace, i.e. $n_A \propto |S_A|$. The non-central hypergeometric distribution introduces an additional property $\omega_{j,A}$ for every subspace S_A denoting its relative importance. For identical $\omega_{j,A}, A \ni j$, the non-central distribution reduces to the central hypergeometric distribution. Depending on the application and prior knowledge, we can also choose a non-central version of the distribution as prior.

We denote the approximate distribution for $q_\phi(\boldsymbol{n}|\boldsymbol{z})$ as $r_\psi(\boldsymbol{n}|\boldsymbol{z})$. As different modalities share different subspaces, the number of elements assigned to shared subspaces must be equal for all modalities contributing to this subspace. Hence, it follows

$$r_\psi(\boldsymbol{n}|\boldsymbol{z}) = r_\psi(\{n_A | \boldsymbol{z}_A\}_{A \subseteq \mathcal{M}}) \tag{1}$$

where we define every $r_\psi(n_A | \boldsymbol{z}_A)$ as categorical distribution taking values in $\{0, \dots, |S_A|\}$. For differentiability, we use its continuous relaxation [Jang et al., 2016, Maddison et al., 2016] (see Appendix E.1 for details).

3.2 Selection of subspace elements

Independent of the inferred distribution over dimensionalities n_A, each $z_j^{(k)}, k \in \{1, \dots, D_j\}$ will be assigned to one of the subspaces $\{S_A | A \ni j\}$, which describes a categorical distribution over subspaces S_A.

[1] from now on we drop the superscript (i) to reduce clutter

Definition 2 (Prior distribution $q_\phi(y_j^{(k)})$ and posterior distribution $q_\phi(y_j^{(k)}|\boldsymbol{z})$)**.** *For every modality $\boldsymbol{x}_j$, we define the prior distribution $q_\phi(y_j^{(k)})$ for every latent factor $z_j^{(k)}$ as a categorical distribution over subspaces $\{S_A | A \ni j\}$ with uniform weights $s_{j,A}^{(k)} = \frac{1}{2^{M-1}}$. The posterior distribution $q_\phi(y_j^{(k)}|\boldsymbol{z})$ is defined as a categorical distribution with weights $\boldsymbol{s}_j^{(k)}$ where $\sum_{A \ni j} s_{j,A}^{(k)} = 1$ and $s_{j,A}^{(k)} \geq 0$*

For the prior distribution $q_\phi(y_j^{(k)})$, the value of the category weights $s_{j,A}^{(k)}$ follows the inverse of the number of subspaces S_A every modality $\boldsymbol{x}_j$ contributes to, which is 2^{M-1} for M modalities. Depending on application and prior knownledge, the weights for the prior distribution can be skewed as well.

The number of factors n_A per subspace S_A needs to be the same across all modalities $\boldsymbol{x}_j$ contributing to this subspace, which is difficult to achieve and expensive to compute using the formulation in Definition 2. Therefore, we model the assignment of latent factors $z_j^{(k)}$ to a subspace S_A as a subset sampling procedure. The inference of the number of latent factors n_A per subspace S_A enables the formulation as sampling of n_A out of $|S_A|$ latent factors. Xie and Ermon [2019] state that any top-k relaxation can be used as subset sampling. Hence, we approximate the true posterior with the following approximative distribution

$$r_\psi(\boldsymbol{y}_j|\boldsymbol{n}, \boldsymbol{z}) = r_\psi(\{\boldsymbol{y}_{j,A}|n_A, \boldsymbol{z}_A\}_{A \ni j}), \forall j \in \mathcal{M} \tag{2}$$

where every $r_\psi(\boldsymbol{y}_{j,A}|n_A, \boldsymbol{z}_A)$ is the output of a top-k relaxation scheme which allows to sample the top n_A values given $\boldsymbol{z}_A$. We use the framework by Grover et al. [2019], a relaxed subset sampling method, which allows the integration into a fully differentiable model (see Appendix E.2 for details).

3.3 ELBO Formulation

To approximate the intractable $q_\phi(\boldsymbol{z}, \boldsymbol{n}, \boldsymbol{y})$, we introduce an additional posterior approximation $r_\psi(\boldsymbol{n}, \boldsymbol{y} \mid \boldsymbol{z})$.

$$-\log q_\phi(\boldsymbol{z} \mid \boldsymbol{X}) \geq -E_{q_\phi(\boldsymbol{n},\boldsymbol{y}|\boldsymbol{z})}[\log q_\phi(\boldsymbol{z}, \boldsymbol{n}, \boldsymbol{y}) - \log r_\psi(\boldsymbol{n}, \boldsymbol{y} \mid \boldsymbol{z})] \tag{3}$$

The approximation of $q_\phi(\boldsymbol{z} \mid \boldsymbol{X})$ using Equation (3) leads to the new multimodal ELBO formulation presented in Definition 3.

Definition 3. *Let $r_\psi(\boldsymbol{n}, \boldsymbol{y}|\boldsymbol{z})$ be the posterior approximation to $q_\phi(\boldsymbol{n}, \boldsymbol{y}|\boldsymbol{z})$. Then, the objective $\mathcal{L}(\theta, \phi, \psi; \boldsymbol{X})$ for learning a joint distribution of multiple data types $\boldsymbol{X}$ and the relation $(\boldsymbol{n}, \boldsymbol{y})$ between them is defined as*

$$\mathcal{L}(\theta, \phi, \psi; \boldsymbol{X}) = E_{q_\phi(\boldsymbol{z},\boldsymbol{n},\boldsymbol{y}|\boldsymbol{X})}\left[\log p_\theta(\boldsymbol{X} \mid \boldsymbol{z}) - \log \frac{q_\phi(\boldsymbol{z}|\boldsymbol{n}, \boldsymbol{y})}{p_\theta(\boldsymbol{z})} - \log \frac{q_\phi(\boldsymbol{n}, \boldsymbol{y})}{r_\psi(\boldsymbol{n}, \boldsymbol{y}|\boldsymbol{z})}\right] \tag{4}$$

Lemma 1. *The objective $\mathcal{L}(\theta, \phi, \psi; \boldsymbol{X})$, defined in Definition 3 is a valid multimodal ELBO, i.e. $\log p_\theta(\boldsymbol{X}) \geq \mathcal{L}(\theta, \phi, \psi; \boldsymbol{X})$.*

The proof to Lemma 1 can be found in Appendix B. The proposed ELBO $\mathcal{L}(\theta, \phi, \psi; \boldsymbol{X})$ draws inspiration from Ranganath et al. [2016]. Nonetheless, the proposed model significantly differs from the original work as we are interested in learning the relation between multiple sets of latent factors compared to learning more expressive distributions in the original paper.

4 Related Work

Previous work on multimodal VAEs put focus on how to aggregate posterior approximation in a single joint latent space [Wu and Goodman, 2018, Shi et al., 2019, Sutter et al., 2021]. Every modality or data type consists of shared as well as modality-specific information. The connection between different data types can be seen as how much and what information is shared between two modalities. Aggregating over all dimensions in the latent space is only desirable if the amount of shared information is relatively high compared to the amount of modality-specific information. Another line of research explicitly split the latent space into a shared and a modality-specific part [Hsu and Glass, 2018, Daunhawer et al., 2020, Sutter et al., 2020]. Besides introducing a lot of hyper-parameters, these works are only able to model a star-like relation between modalities.

5 Experiments

We evaluate the proposed multimodal relational VAE (mrVAE) using the recently proposed PolyMNIST dataset [Sutter et al., 2021] in this work. Additionally, we propose an extension to the vanilla version that adds pairwise shared information between images. In our extension, we add digit information to two corners of every modality. Every corner digit is shared between two modalities only. See Figure 2 in the appendix for examples and differences of the two datasets.

Table 1: Comparison between mrVAE and MoPoE-VAE on two different versions of the PolyMNIST dataset, vanilla and extended, regarding their coherence of generated samples. The coherence of generated samples is assessed using pre-trained classifiers, see Sutter et al. [2021] for details. $\boldsymbol{x}_j$ denote input or output modality and $\boldsymbol{X}_A$ input sets.

		Accuracy	
Input	Output	mrVAE	MoPoE
$\boldsymbol{x}_0$	$\boldsymbol{x}_0$	0.50	0.47
$\boldsymbol{X}_{\{0,1\}}$	$\boldsymbol{x}_0$	0.51	0.51
$\boldsymbol{X}_{\{0,1,2\}}$	$\boldsymbol{x}_0$	0.51	0.53

(a) PolyMNIST vanilla

		Accuracy	
Input	Output	mrVAE	MoPoE
$\boldsymbol{x}_0$	$\boldsymbol{x}_0$	0.40	0.31
$\boldsymbol{X}_{\{0,1\}}$	$\boldsymbol{x}_0$	0.44	0.32
$\boldsymbol{X}_{\{0,1,2\}}$	$\boldsymbol{x}_0$	0.44	0.32

(b) PolyMNIST extended

Tables 1 and 2 show the results of the proposed mrVAE in comparison to the MoPoE-VAE [Sutter et al., 2021] with respect to their generative sample quality and coherence. For the vanilla version of the dataset, mrVAE is able to reach the performance of the MoPoE-VAE regarding the coherence of samples (see Table 1a). If we evaluate the two models on the extended version of PolyMNIST, we see a drastic decrease of coherence for the MoPoE-VAE, while mrVAE-s performance remains comparable(Table 1b). Notice that we only evaluate with respect to the main digit. The proposed mrVAE reaching the same performance as the MoPoE-VAE on the vanilla PolyMNIST dataset is even more remarkable, if we consider that the vanilla version of the dataset applies to the restrictive assumptions for relations between modalities.

Regarding the quality of generated samples, we see another effect of the limiting assumptions of previous work. The averaging across all dimensions may lead to a decrease in quality of samples if we generate samples of a modality which is not given as input (see Table 2a). Even more surprising, the additional relations between modalities lead to a further decrease in the quality of generated samples for the MoPoE-VAE (see Table 2b). The proposed mrVAE on the other side is able to consistently generate high quality samples for the vanilla and extended version of the PolyMNIST dataset (see Tables 2a and 2b).

Table 2: Comparison between mrVAE and MoPoE-VAE on two different versions of the PolyMNIST dataset, vanilla and extended, regarding their quality of generated samples. The quality of samples is assessed using the FID [Heusel et al., 2017]. $\boldsymbol{x}_j$ denote input or output modality and $\boldsymbol{X}_A$ input sets.

		FID	
Input	Output	mrVAE	MoPoE
$\boldsymbol{x}_0$	$\boldsymbol{x}_0$	105.8	108.7
$\boldsymbol{x}_1$	$\boldsymbol{x}_0$	129.9	216.6
$\boldsymbol{X}_{\{1,2\}}$	$\boldsymbol{x}_0$	139.6	221.5

(a) PolyMNIST vanilla

		FID	
Input	Output	mrVAE	MoPoE
$\boldsymbol{x}_0$	$\boldsymbol{x}_0$	95.1	157.9
$\boldsymbol{x}_1$	$\boldsymbol{x}_0$	182.4	265.5
$\boldsymbol{X}_{\{1,2\}}$	$\boldsymbol{x}_0$	189.0	311.6

(b) PolyMNIST extended

6 Conclusion

In this work, we propose a new formulation to learn from multimodal datasets. Our formulation does not rely on simplistic assumptions between data types and is able to learn relations between data types. In our empirical evaluation, we show the limitations of previous work if the simple relations between data types are made slightly more complicated. The proposed method shows promising results towards overcoming these limitations. Nevertheless, this is still work in progress and there is additional research needed to fully understand the dynamics of such relations. Also, the formulations for the approximation distributions are not ideal yet and will be worked on in future steps.

References

Y. Bengio, N. Léonard, and A. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

I. Daunhawer, T. M. Sutter, R. Marcinkevics, and J. E. Vogt. Self-supervised Disentanglement of Modality-specific and Shared Factors Improves Multimodal Generative Models. In *German Conference on Pattern Recognition*. Springer, 2020.

A. Grover, E. Wang, A. Zweig, and S. Ermon. Stochastic Optimization of Sorting Networks via Continuous Relaxations. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=H1eSS3CcKX`.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

W.-N. Hsu and J. Glass. Disentangling by Partitioning: A Representation Learning Framework for Multimodal Sensory Data. 2018. URL `http://arxiv.org/abs/1805.11264`.

E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

W. Kool, H. van Hoof, and M. Welling. Ancestral Gumbel-Top-k Sampling for Sampling Without Replacement. *Journal of Machine Learning Research*, 21(47):1–36, 2020. URL `http://jmlr.org/papers/v21/19-985.html`.

C. J. Maddison, A. Mnih, and Y. W. Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

R. Ranganath, D. Tran, and D. Blei. Hierarchical variational models. In *International Conference on Machine Learning*, pages 324–333, 2016.

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.

T. M. Sutter, I. Daunhawer, and J. E. Vogt. Multimodal Generative Learning Utilizing Jensen-Shannon-Divergence. 2020. URL `https://arxiv.org/abs/2006.08242`.

T. M. Sutter, I. Daunhawer, and J. E. Vogt. Generalized Multimodal ELBO. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=5Y21V0RDBV`.

G. Upton and I. Cook. *A dictionary of statistics 3e*. Oxford university press, 2014.

M. Wu and N. Goodman. Multimodal Generative Models for Scalable Weakly-Supervised Learning. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montreal, Canada*, pages 5580–5590, 2 2018. URL `http://arxiv.org/abs/1802.05335`.

S. M. Xie and S. Ermon. Reparameterizable subset sampling via continuous relaxations. *arXiv preprint arXiv:1901.10517*, 2019.

A Hierarchical Posterior Approximation

Ranganath et al. [2016] propose to use a hierarchical posterior approximation instead of a "flat" approximation. They use the additional level of hierarchy to increase the expressivity of their posterior approximation. A hierarchical model allows for breaking of the mean-field assumption (we refer to their paper for more details). We propose to use the additional level of hierarchy to learn the structure of the latent space. For multimodal learning problems, we are interested in knowing which modalities share information between each other. Such that only information is aggregated which is connected.

For this, we introduce two additional latent variables, $\boldsymbol{n}$ and $\boldsymbol{y}$ which define how many and which latent factors are shared between the different modalities.

$$
\begin{aligned}
q_\phi(\boldsymbol{z}) &= \int q_\phi(\boldsymbol{n}, \boldsymbol{y}) \cdot q(\boldsymbol{z} \mid \boldsymbol{n}, \boldsymbol{y}) d\boldsymbol{n} d\boldsymbol{y} && (5) \\
&= \sum q_\phi(\boldsymbol{n}, \boldsymbol{y}) \cdot q(\boldsymbol{z} \mid \boldsymbol{n}, \boldsymbol{y}) && (6)
\end{aligned}
$$

where Equation (6) follows from $\boldsymbol{n}$ and $\boldsymbol{y}$ being discrete random variables. In general, this summation is computationally intractable. Hence, we use the variational inference procedure twice to create a computationally feasible objective which approximates the joint probability of all our data types $\log p_\theta(\boldsymbol{X})$.

A.1 Variational Inference for q

To approximate the intractable $q_\phi(\boldsymbol{z}, \boldsymbol{n}, \boldsymbol{y})$, we introduce an additional posterior approximation $r_\psi(\boldsymbol{n}, \boldsymbol{y} \mid \boldsymbol{z})$. We can approximate $q_\phi(\boldsymbol{z})$ either by taking the expectation over $r_\psi(\cdot)$ or by taking the expectation over $q_\phi(\cdot)$. The first case leads to

$$
\log q_\phi(\boldsymbol{z} \mid \boldsymbol{X}) \geq E_{r_\psi(\boldsymbol{n}, \boldsymbol{y} \mid \boldsymbol{z})}[\log q_\phi(\boldsymbol{z}, \boldsymbol{n}, \boldsymbol{y}) - \log r_\psi(\boldsymbol{n}, \boldsymbol{y} \mid \boldsymbol{z})] \quad (7)
$$

This would lead to an Expectation-Maximization-type like learning procedure.

The second case leads to the following approximation

$$
\log q_\phi(\boldsymbol{z} \mid \boldsymbol{X}) \leq E_{q_\phi(\boldsymbol{n}, \boldsymbol{y} \mid \boldsymbol{z})}[\log q_\phi(\boldsymbol{z}, \boldsymbol{n}, \boldsymbol{y}) - \log r_\psi(\boldsymbol{n}, \boldsymbol{y} \mid \boldsymbol{z})] \quad (8)
$$

We will focus on this second case.

B Proof to Lemma 1

Proof. We start with the standard ELBO-formulation

$$
\log p_\theta(\boldsymbol{X}) \geq E_{q(\boldsymbol{z}|\boldsymbol{X})} \left[\log p_\theta(\boldsymbol{X}, \boldsymbol{z}) - \log q_\phi(\boldsymbol{z}|\boldsymbol{X})\right] \quad (9)
$$

We add an additional level of hierarchy in the ELBO formulation [Ranganath et al., 2016] to introduce latent variables for the number of factors per subspace $\boldsymbol{n}$ and the subset selection $\boldsymbol{y}$:

$$
\log q_\phi(\boldsymbol{z} \mid \boldsymbol{X}) \leq E_{q_\phi(\boldsymbol{n}, \boldsymbol{y} \mid \boldsymbol{z})}[\log q_\phi(\boldsymbol{z}, \boldsymbol{n}, \boldsymbol{y}) - \log r_\psi(\boldsymbol{n}, \boldsymbol{y} \mid \boldsymbol{z})] \quad (10)
$$

from where it directly follows:

$$
-\log q_\phi(\boldsymbol{z} \mid \boldsymbol{X}) \geq -E_{q_\phi(\boldsymbol{n}, \boldsymbol{y} \mid \boldsymbol{z})}[\log q_\phi(\boldsymbol{z}, \boldsymbol{n}, \boldsymbol{y}) - \log r_\psi(\boldsymbol{n}, \boldsymbol{y} \mid \boldsymbol{z})] \quad (11)
$$

Combining Equations (9) and (11), we get

$$
\begin{aligned}
\log p_\theta(\boldsymbol{X}) \geq & E_{q_\phi(\boldsymbol{z}|\boldsymbol{X})} \left[\log p_\theta(\boldsymbol{X}, \boldsymbol{z}) - E_{q_\phi(\boldsymbol{n}, \boldsymbol{y}|\boldsymbol{z})}[\log q_\phi(\boldsymbol{n}, \boldsymbol{y}, \boldsymbol{z}) - \log r_\psi(\boldsymbol{n}, \boldsymbol{y}|\boldsymbol{z})]\right] && (12) \\
= & E_{q_\phi(\boldsymbol{z}|\boldsymbol{X})} \left[E_{q_\phi(\boldsymbol{n}, \boldsymbol{y}|\boldsymbol{z})}[\log p_\theta(\boldsymbol{X}, \boldsymbol{z}) - \log q_\phi(\boldsymbol{n}, \boldsymbol{y}, \boldsymbol{z}) + \log r_\psi(\boldsymbol{n}, \boldsymbol{y}|\boldsymbol{z})]\right] && (13) \\
= & E_{q_\phi(\boldsymbol{z}, \boldsymbol{n}, \boldsymbol{y}|\boldsymbol{X})} \left[\log p_\theta(\boldsymbol{X} \mid \boldsymbol{z}) - \log \frac{q_\phi(\boldsymbol{z}|\boldsymbol{n}, \boldsymbol{y})}{p_\theta(\boldsymbol{z})} - \log \frac{q_\phi(\boldsymbol{n}, \boldsymbol{y})}{r_\psi(\boldsymbol{n}, \boldsymbol{y}|\boldsymbol{z})}\right] && (14)
\end{aligned}
$$

□

C Detailed Model

C.1 ELBO

$$\begin{align}
\log p_\theta(\boldsymbol{X}) \geq & E_{q_\phi(\boldsymbol{z}|\boldsymbol{X})}\left[\log p_\theta(\boldsymbol{X},\boldsymbol{z}) - E_{q_\phi(\boldsymbol{n},\boldsymbol{y}|\boldsymbol{z})}[\log q_\phi(\boldsymbol{n},\boldsymbol{y},\boldsymbol{z}) - \log r_\psi(\boldsymbol{n},\boldsymbol{y}|\boldsymbol{z})]\right] \tag{15} \\
= & E_{q_\phi(\boldsymbol{z}|\boldsymbol{X})}\left[E_{q_\phi(\boldsymbol{n},\boldsymbol{y}|\boldsymbol{z})}[\log p_\theta(\boldsymbol{X},\boldsymbol{z}) - \log q_\phi(\boldsymbol{n},\boldsymbol{y},\boldsymbol{z}) + \log r_\psi(\boldsymbol{n},\boldsymbol{y}|\boldsymbol{z})]\right] \tag{16} \\
= & E_{q_\phi(\boldsymbol{z}|\boldsymbol{X})}[E_{q_\phi(\boldsymbol{n},\boldsymbol{y}|\boldsymbol{z})}[\log p_\theta(\boldsymbol{X}|\boldsymbol{z}) + \log p_\theta(\boldsymbol{z}) - \log q_\phi(\boldsymbol{z}|\boldsymbol{n},\boldsymbol{y}) \nonumber \\
& - \log q_\phi(\boldsymbol{n},\boldsymbol{y}) + \log r_\psi(\boldsymbol{n},\boldsymbol{y}|\boldsymbol{z})]] \tag{17} \\
= & E_{q_\phi(\boldsymbol{z},\boldsymbol{n},\boldsymbol{y}|\boldsymbol{X})}\left[\log p_\theta(\boldsymbol{X} \mid \boldsymbol{z}) - \log\frac{q_\phi(\boldsymbol{z}|\boldsymbol{n},\boldsymbol{y})}{p_\theta(\boldsymbol{z})} - \log\frac{q_\phi(\boldsymbol{n},\boldsymbol{y})}{r_\psi(\boldsymbol{n},\boldsymbol{y}|\boldsymbol{z})}\right] \tag{18} \\
= & E_{q_\phi(\boldsymbol{z},\boldsymbol{n},\boldsymbol{y}|\boldsymbol{X})}\left[\log p_\theta(\boldsymbol{X} \mid \boldsymbol{z}) - \log\frac{q_\phi(\boldsymbol{z}|\boldsymbol{n},\boldsymbol{y})}{p_\theta(\boldsymbol{z})}\right] \nonumber \\
& - E_{q_\phi(\boldsymbol{z},\boldsymbol{n},\boldsymbol{y}|\boldsymbol{X})}\left[\log\frac{q_\phi(\boldsymbol{n},\boldsymbol{y})}{r_\psi(\boldsymbol{n},\boldsymbol{y}|\boldsymbol{z})}\right] \tag{19} \\
= & E_{q_\phi(\boldsymbol{n},\boldsymbol{y})q_\phi(\boldsymbol{z}|\boldsymbol{n},\boldsymbol{y})}\left[\log p_\theta(\boldsymbol{X} \mid \boldsymbol{z}) - \log\frac{q_\phi(\boldsymbol{z}|\boldsymbol{n},\boldsymbol{y})}{p_\theta(\boldsymbol{z})}\right] \nonumber \\
& - E_{q_\phi(\boldsymbol{n},\boldsymbol{y})q_\phi(\boldsymbol{z}|\boldsymbol{n},\boldsymbol{y})}\left[\log\frac{q_\phi(\boldsymbol{n},\boldsymbol{y})}{r_\psi(\boldsymbol{n},\boldsymbol{y}|\boldsymbol{z})}\right] \tag{20} \\
= & E_{q_\phi(\boldsymbol{n},\boldsymbol{y})}\left[E_{q_\phi(\boldsymbol{z}|\boldsymbol{n},\boldsymbol{y})}\left[\log p_\theta(\boldsymbol{X} \mid \boldsymbol{z}) - \log\frac{q_\phi(\boldsymbol{z}|\boldsymbol{n},\boldsymbol{y})}{p_\theta(\boldsymbol{z})}\right]\right] \nonumber \\
& - E_{q_\phi(\boldsymbol{z}|\boldsymbol{n},\boldsymbol{y})q_\phi(\boldsymbol{n},\boldsymbol{y})}\left[\log\frac{q_\phi(\boldsymbol{n},\boldsymbol{y})}{r_\psi(\boldsymbol{n},\boldsymbol{y}|\boldsymbol{z})}\right] \tag{21} \\
= & E_{q_\phi(\boldsymbol{n},\boldsymbol{y})}\left[E_{q_\phi(\boldsymbol{z}|\boldsymbol{n},\boldsymbol{y})}\left[\log p_\theta(\boldsymbol{X} \mid \boldsymbol{z}) - \log\frac{q_\phi(\boldsymbol{z}|\boldsymbol{n},\boldsymbol{y})}{p_\theta(\boldsymbol{z})}\right]\right] \nonumber \\
& - E_{q_\phi(\boldsymbol{z}|\boldsymbol{n},\boldsymbol{y})}\left[E_{q_\phi(\boldsymbol{n},\boldsymbol{y})}\left[\log\frac{q_\phi(\boldsymbol{n},\boldsymbol{y})}{r_\psi(\boldsymbol{n},\boldsymbol{y}|\boldsymbol{z})}\right]\right] \tag{22}
\end{align}$$

C.2 Graphical Models

Figure 1 shows graphical models of the generative process for previous works (Figure 1a) as well as the proposed work (Figure 1b). They highlight the differences and extensions from a modelling perspective.

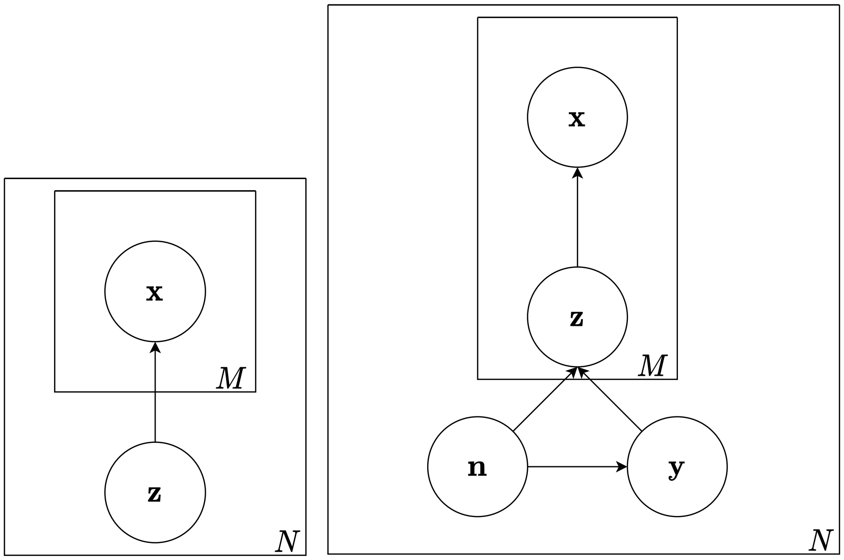
D Example for M=3

Let us write the terms above in a more detailed form as we constrain (without loss of generality) the number of modalities to $M = 3$ for now. For ease of notation we write the set of indices for a subset without the curly brackets, e.g. $\boldsymbol{z}_{123}$ instead of $\boldsymbol{z}_{\{123\}}$. We start with $p_\theta(\boldsymbol{X}|\boldsymbol{z})$:

$$\begin{align}
p_\theta(\boldsymbol{X}|\boldsymbol{z}) &= p_\theta(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3|\boldsymbol{z}) \tag{23} \\
&= p_\theta(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3|\boldsymbol{z}_{123},\boldsymbol{z}_{12},\boldsymbol{z}_{13},\boldsymbol{z}_{23},\boldsymbol{z}_1,\boldsymbol{z}_2,\boldsymbol{z}_3) \tag{24} \\
&= p_\theta(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3|\boldsymbol{z}_{123},\boldsymbol{z}_{12},\boldsymbol{z}_{13},\boldsymbol{z}_{23},\boldsymbol{z}_1,\boldsymbol{z}_2,\boldsymbol{z}_3) \tag{25} \\
&= p_\theta(\boldsymbol{x}_1|\boldsymbol{z}_{123},\boldsymbol{z}_{12},\boldsymbol{z}_{13},\boldsymbol{z}_1)\cdot p_\theta(\boldsymbol{x}_2|\boldsymbol{z}_{123},\boldsymbol{z}_{12},\boldsymbol{z}_{23},\boldsymbol{z}_2)\cdot p_\theta(\boldsymbol{x}_3|\boldsymbol{z}_{123},\boldsymbol{z}_{13},\boldsymbol{z}_{23},\boldsymbol{z}_3) \tag{26}
\end{align}$$

The step from Equation (25) to Equation (26) follows from the conditional indepence between modalities given the latents. Additionally, by definition modalities are independent of subspaces they do not contribute to. From the Bayes' rule, it follows:

$$r_\psi(\boldsymbol{n},\boldsymbol{y}|\boldsymbol{z},\boldsymbol{X}) = r_\psi(\boldsymbol{y}|\boldsymbol{n},\boldsymbol{z},\boldsymbol{X})\cdot r_\psi(\boldsymbol{n}|\boldsymbol{z},\boldsymbol{X}) \tag{27}$$



(a) Previous Works

(b) Proposed Work

Figure 1: Graphical Models of previous works and the proposed model.

Following Definition 1 and Definition 2, we can write the distribution over subspaces more explicit for $M = 3$:

$$\begin{aligned}
r_{\psi}(\boldsymbol{n}|\boldsymbol{z}, \boldsymbol{X}) =& r_{\psi}(n_{123}, n_{12}, n_{13}, n_{23}, n_1, n_2, n_3|\boldsymbol{z}, \boldsymbol{X}) \qquad (28)\\
& 0. \\
r_{\psi}(\boldsymbol{y}|\boldsymbol{n}, \boldsymbol{z}, \boldsymbol{X}) =& r_{\psi}(\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3|\boldsymbol{n}, \boldsymbol{z}, \boldsymbol{X}) \qquad (29)\\
=& r_{\psi}(\boldsymbol{y}_1|\boldsymbol{n}, \boldsymbol{z}, \boldsymbol{X}) \cdot r_{\psi}(\boldsymbol{y}_2|\boldsymbol{n}, \boldsymbol{z}, \boldsymbol{X}) \cdot r_{\psi}(\boldsymbol{y}_3|\boldsymbol{n}, \boldsymbol{z}, \boldsymbol{X}) \qquad (30)\\
=& r_{\psi}(\boldsymbol{y}_1|n_{123}, n_{12}, n_{13}, n_1, \boldsymbol{z}, \boldsymbol{X}) \cdot r_{\psi}(\boldsymbol{y}_2|n_{123}, n_{12}, n_{23}, n_2, \boldsymbol{z}, \boldsymbol{X}) \\
& \cdot r_{\psi}(\boldsymbol{y}_3|n_{123}, n_{13}, n_{23}, n_3, \boldsymbol{z}, \boldsymbol{X}) \qquad (31)\\
=& r_{\psi}(\boldsymbol{y}_{1,123}, \boldsymbol{y}_{1,12}, \boldsymbol{y}_{1,13}, \boldsymbol{y}_{1,1}|n_{123}, n_{12}, n_{13}, n_1, \boldsymbol{z}, \boldsymbol{X}) \\
& \cdot r_{\psi}(\boldsymbol{y}_{2,123}, \boldsymbol{y}_{2,12}, \boldsymbol{y}_{2,13}, \boldsymbol{y}_{2,2}|n_{123}, n_{12}, n_{23}, n_2, \boldsymbol{z}, \boldsymbol{X}) \\
& \cdot r_{\psi}(\boldsymbol{y}_{3,123}, \boldsymbol{y}_{3,13}, \boldsymbol{y}_{3,23}, \boldsymbol{y}_{3,3}|n_{123}, n_{13}, n_{23}, n_3, \boldsymbol{z}, \boldsymbol{X}) \qquad (32)
\end{aligned}$$

E Implementation Details, Experiments and Evaluation

E.1 Number of elements per latent subspace

All distributions are implemented as neural networks with the conditioning set being the input to the neural net and the output being the parameters of the respective distribution. This is similar to the unimodal VAE where the posterior approximation $q_{\phi}(\boldsymbol{z}|\boldsymbol{x})$ is a neural network which outputs $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$: $\boldsymbol{\mu}, \boldsymbol{\Sigma} = f_{\phi}(\boldsymbol{x})$.

Here, we have individual networks for all $r_{\psi}(n_A|\boldsymbol{z}_A)$ which output the weights of the respective categorical distribution $\boldsymbol{c}_A$:

$$\boldsymbol{c}_A = f_{\psi,A}(\boldsymbol{z}_A), \text{ with } \boldsymbol{z}_A = \{\boldsymbol{z}_j\}_{j \in A} \qquad (33)$$

To handle missing modalities, the function f_{ψ_A} consists itself of $|A|$ subfunctions $g_{\psi,A,j}$ which take the modalitities $\boldsymbol{z}_j$ as input and are blocks of one fully-connected layer followed by a ReLU activation function. The outputs of $g_{\psi,A,j}$ are again averaged and fed into a linear layer followed by a ReLU activation function:

$$f_{\psi,A}(\boldsymbol{z}_A) = \text{ReLU}\left(\text{FC}\left(\frac{1}{|A|}\sum_{j\in A} f_{\psi,A,j}(\boldsymbol{z}_j)\right)\right) \tag{34}$$

$$= \text{ReLU}\left(\text{FC}\left(\frac{1}{|A|}\sum_{j\in A} \text{ReLU}(\text{FC}(\boldsymbol{z}_j))\right)\right) \tag{35}$$

where FC denotes a fully-connected layer. The values $\boldsymbol{c}_A$ can be used as the input for a Gumbel-Softmax model [Maddison et al., 2016, Jang et al., 2016]. The output will be a one-hot vector denoting the number of elements n_A per subspace (see Bengio et al. [2013]).

E.2 Selection of latent factors

And similarly for the selection of latent factors $\boldsymbol{y}_j$. As stated in Xie and Ermon [2019], the subset sampling procedure is equivalent to a ranking scheme where the ranking determines the probability of being selected for a subset. The random variable $\boldsymbol{y}_{j,A}$ is a matrix where the first n_A rows are one-hot-vectors with element to be sampled being one. For more details on the relation between subset sampling, top-k relaxation and continuous sorting and ranking, we refer to [Grover et al., 2019, Xie and Ermon, 2019, Kool et al., 2020]. The implementation of the respective building blocks follows a similar principle as in Appendix E.1 to enable missingness of modalities.

$$\boldsymbol{w}_{j,A} = g_{\psi,A,j}(\boldsymbol{z}_A) \text{ with } \boldsymbol{z}_A = \{\boldsymbol{z}_j\}_{j\in A} \tag{36}$$

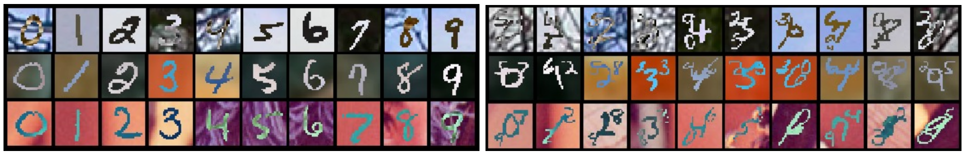
and more detailed, similar as in Appendix E.1

$$g_{\psi,A,j}(\boldsymbol{z}_A) = \text{ReLU}\left(\text{FC}\left(\frac{1}{|A|}\sum_{i\in A} g_{\psi,A,j,i}(\boldsymbol{z}_i)\right)\right) \tag{37}$$

$$= \text{ReLU}\left(\text{FC}\left(\frac{1}{|A|}\sum_{i\in A} \text{ReLU}(\text{FC}(\boldsymbol{z}_i))\right)\right) \tag{38}$$

The weights $\boldsymbol{w}_{j,A}$ are reparameterized using Gumbel-noise and then fed to the `neuralsort` building block [Grover et al., 2019]. `neuralsort` outputs a permutation matrix where every row is a one-hot vector (or a continuous relaxation version of it, see Bengio et al. [2013]). Using n_A it is possible to use the top-n_A elements in a fully-differentiable pipeline.

E.3 Further Details on the Experiments and their Evaluation



(a) PolyMNIST vanilla

(b) PolyMNIST extended

Figure 2: Examples of the two PolyMNIST dataset variations we use in this work, vanilla and extended.

We use the same architectures and capacities for enccoder and decoder in all experiments, as well as the latent space dimensionalities. This lays the ground for a fair comparison between models. Figure 2 shows samples of the used versions of the datasets.

We also looked into model selection criteria. We strongly believe that for real-world multimodal datasets it is difficult to collect ground-truth labels with respect to the relation between datasets, but also all possible downstream tasks. Hence, we selected the models the evaluation based on unsupervised metrics only. For this submission, we evaluated the model at the point in training which leads to the lowest test loss. This does not necessarily lead to the best scores with respect to metrics like generation accuracy or FID-scores, but - in our opinion - is a fair way to evaluate models based on their optimization objective. Also, this model selection criteria would work in case there is no access to expert labels, which is an important direction for future work. Therefore the performance numbers of the MoPoE-VAE differ compared to the experiments in previous work.

Nevertheless, model selection in multimodal VAEs is still an open research question which requires more work.