M²VAE – Derivation of a Multi-Modal Variational Autoencoder Objective from the Marginal Joint Log-Likelihood

Timo Korthals[†*]

Abstract

This work gives an in-depth derivation of the trainable evidence lower bound (ELBO) obtained from the marginal joint log-Likelihood with the goal of training a multi-modal variational Autoencoder (M²VAE).

I. Introduction

Variational auto encoder (VAE) combine neural networks with variational inference to allow unsupervised learning of complicated distributions according to the graphical model shown in Fig. 1 (left). A D_a-dimensional observation a is modeled in terms of a D_z-dimensional latent vector z using a probabilistic decoder $p_{\theta_\mathrm{a}}(z)$ with parameters θ. To generate the corresponding embedding z from observation a, a probabilistic encoder network with $q_{\phi_\mathrm{a}}(z)$ is being provided which parametrizes the posterior distribution from which z is sampled. The encoder and decoder, given by neural networks, are trained jointly to bring a close to an a' under the constraint that an approximate distribution needs to be close to a prior $p(z)$ and hence inference is basically learned during training.

The specific objective of VAEs is the maximization of the marginal distribution $p(a) = \int p_\theta(a|z)p(z)\,\mathrm{d}a$. Because this distribution is intractable, the model is instead trained via *stochastic gradient variational Bayes* (SGVB) by maximizing the *evidence lower bound* (ELBO) $\mathcal{L}$ of the marginal log-likelihood as described in Sec. II This approach proposed by [1] is used in settings where only a single modality a is present in order to find a latent encoding z (c.f. Fig. 1 (left)).

This work gives an in-depth derivation of the trainable *evidence lower bound* (ELBO) obtained from the marginal joint log-Likelihood, that satisfies all plate models as depicted in Fig. 1, we are with the goal of training a multi-modal variational Autoencoder (M²VAE).

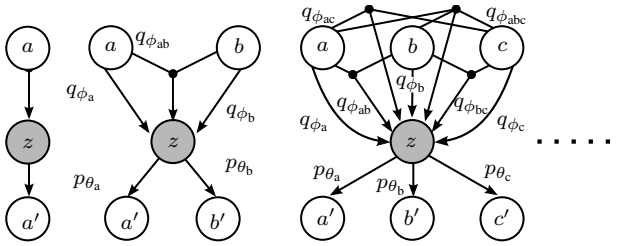

Fig. 1: Evolution of full uni-, bi-, and tri-modal VAEs comprising all modality permutations

[†]Bielefeld University, Cognitronics & Sensor Systems, Inspiration 1, 33619 Bielefeld, Germany
[*]tkorthals@cit-ec.uni-bielefeld.de

Work in progress

II. Variational Autoencoder

First, the derivation of the vanilla *Variational Autoencoder* by [1] is recaped.

A. The Variational Bound

$$
\begin{aligned}
L &= \log(p(a)) && && (1)\\
&= \sum_z q(z|a)\log(p(a)) && \text{Equation 89 w/o conditional} && (2)\\
&= \sum_z q(z|a)\log\left(\frac{p(z,a)}{p(z|a)}\right) && \text{Equation 84} && (3)\\
&= \sum_z q(z|a)\log\left(\frac{p(z,a)}{p(z|a)}\frac{q(z|a)}{q(z|a)}\right) && \text{multiplied by 1} && (4)\\
&= \sum_z q(z|a)\log\left(\frac{p(z,a)}{q(z|a)}\frac{q(z|a)}{p(z|a)}\right) && \text{reordered} && (5)\\
&= \sum_z q(z|a)\log\left(\frac{p(z,a)}{q(z|a)}\right) + \sum_z q(z|a)\log\left(\frac{q(z|a)}{p(z|a)}\right) && \text{Equation 86} && (6)\\
&= \mathcal{L} + \mathrm{D_{KL}}(q(z|a)\|p(z|a)) && \text{Equation 92 \& 88} && (7)\\
&\geq \mathcal{L} && \mathrm{D_{KL}} \geq 0 && (8)
\end{aligned}
$$

$\mathrm{D_{KL}}$ is the Kulbeck-Leibler divergence, with $\mathrm{D_{KL}} \geq 0$, thats depends on how good $q(z|a)$ can approximate $p(z|a)$. $\mathcal{L}$ is the lower variational bound of the marginal log-likelihood, also called the *evidence lower bound* (**ELBO**). If and only if the two distributions q and p are identical, $\mathrm{D_{KL}}$ becomes 0 ($q = p \Leftrightarrow \mathrm{D_{KL}} = 0$). $\mathcal{L} = L$ means on the other hand therefore implicitly, that q perfectly approximates p. It is beacuse $\mathcal{L}$ and $\mathrm{D_{KL}}$ are in equilibrium so that minimizing $\mathrm{D_{KL}}$ is identical to the maximization of $\mathcal{L}$ ($\min \mathrm{D_{KL}} \Leftrightarrow \max \mathcal{L}$). Minimizing $\mathrm{D_{KL}}$ is not feasible, because we don't know the true posterior $p(z|a)$. Therefore, $\mathcal{L}$ is further investigated.

B. Approximate Inference (i.e. rewriting $\mathcal{L}$)

$$
\begin{aligned}
\mathcal{L} &= \sum_z q(z|a)\log\left(\frac{p(z,a)}{q(z|a)}\right) && && (9)\\
&= \sum_z q(z|a)\log\left(\frac{p(a|z)p(z)}{q(z|a)}\right) && \text{Equation 84} && (10)\\
&= \sum_z q(z|a)\log\left(\frac{p(z)}{q(z|a)}\right) + \sum_z q(z|a)\log(p(a|z)) && \text{Equation 86} && (11)\\
&= -\mathrm{D_{KL}}(q(z|a)\|p(z)) + \mathbb{E}_{q(z|a)}\log(p(a|z)) && \text{Equation 92} && (12)
\end{aligned}
$$

If the variable a is replaced by some real valued sample $a^{(i)}$ (e.g. image or LiDAR scan), two terms can be identified:

$$
\mathcal{L} = \underbrace{-\mathrm{D_{KL}}\Big(q_\phi\Big(z|a^{(i)}\Big)\|p(z)\Big)}_{\text{Regularization}} + \underbrace{\mathbb{E}_{q_\phi\left(z|a^{(i)}\right)}\log\Big(p_\theta\Big(a^{(i)}|z\Big)\Big)}_{\text{Reconstruction}} \tag{13}
$$

The first term is just a regularize that punishes the variational distribution q, that is the approximatior of the posterior distribution, if it deviates from some prior $p(z)$. The reconstruction term on the other hand compares the difference between the data $a^{(i)}$ of $q_\phi\big(z|a^{(i)}\big)$ wrt. the sampled data $a^{(i)}$ from the likelihood function $p_\theta\big(a^{(i)}|z\big)$.

That means if $\mathcal{L}$ is going to be maximized, the posterior function has to be equal to some prior and the data that is used to sample the latent feature z from the posterior should be equal to the data from the likelihood function p_θ. The objective is now, to find a function q_ϕ and p_θ which own these properties. Luckily, if some parametrized functions q_ϕ and p_θ (with parameters ϕ and θ) are applied, every increase of Equation 13 means that the variational approximator $q_\phi(z|a)$ comes closer to the real posterior functions $p(z|a)$. Therefore, numerical optimization techniques like gradient descent can be applied to this issue. Commonly, two neuronal network, where each tries to find the optimal parameters ϕ and θ, are applied to approximate the functions q_ϕ (i.e. the encoder) and p_θ (i.e. the decoder). However, since the true value of L remains unknown, maximization of $\mathcal{L}$ can only be done until convergence. Thus, the overall procedure has the property of finding local optima.

III. Joint Variational Autoencoder

Second, we expand the VAE from Sec. II to the marginal joint log-likelihood and derive the variational bound as follows:

$$
\begin{aligned}
L_{\text{J}} &= \log(p(a,b)) && && (14)\\
&= \sum_z q(z|a,b)\log(p(a,b)) && \text{Equation 89 w/o conditional} && (15)\\
&= \sum_z q(z|a,b)\log\left(\frac{p(z,a,b)}{p(z|a,b)}\right) && \text{Equation 85} && (16)\\
&= \sum_z q(z|a,b)\log\left(\frac{p(z,a,b)}{p(z|a,b)}\frac{q(z|a,b)}{q(z|a,b)}\right) && \text{multiplied by 1} && (17)\\
&= \sum_z q(z|a,b)\log\left(\frac{p(z,a,b)}{q(z|a,b)}\frac{q(z|a,b)}{p(z|a,b)}\right) && \text{reordered} && (18)\\
&= \sum_z q(z|a,b)\log\left(\frac{p(z,a,b)}{q(z|a,b)}\right) + \sum_z q(z|a,b)\log\left(\frac{q(z|a,b)}{p(z|a,b)}\right) && \text{Equation 86} && (19)\\
&= \mathcal{L}_{\text{J}} + \mathrm{D}_{\mathrm{KL}}(q(z|a,b)\|p(z|a,b)) && \text{Equation 92 \& 88} && (20)\\
&\geq \mathcal{L}_{\text{J}} && && (21)
\end{aligned}
$$

Approximate Inference (i.e. rewriting $\mathcal{L}_{\text{J}}$):

$$
\begin{aligned}
\mathcal{L}_{\text{J}} &= \sum_z q(z|a,b)\log\left(\frac{p(z,a,b)}{q(z|a,b)}\right) && && (22)\\
&= \sum_z q(z|a,b)\log\left(\frac{p(a,b|z)p(z)}{q(z|a,b)}\right) && \text{Equation 84} && (23)\\
&= \sum_z q(z|a,b)\log\left(\frac{p(z)}{q(z|a,b)}\right) + \sum_z q(z|a,b)\log(p(a,b|z)) && \text{Equation 86} && (24)\\
&= \sum_z q(z|a,b)\log\left(\frac{p(z)}{q(z|a,b)}\right) + \sum_z q(z|a,b)\log(p(a|z)) + \sum_z q(z|a,b)\log(p(b|z)) && \text{Equation 90} && (25)\\
&= -\mathrm{D}_{\mathrm{KL}}(q(z|a,b)\|p(z)) + \mathbb{E}_{q(z|a,b)}\log(p(a|z)) + \mathbb{E}_{q(z|a,b)}\log(p(b|z)) && \text{Equation 92} && (26)
\end{aligned}
$$

Three different terms can be identified:

$$
\mathcal{L}_{\text{J}} = \underbrace{-\mathrm{D}_{\mathrm{KL}}(q_{\phi_{\text{ab}}}(z|a,b)\|p(z))}_{\text{Regularization}} + \underbrace{\mathbb{E}_{q_{\phi_{\text{ab}}}(z|a,b)}\log(p_{\theta_{\text{a}}}(a|z))}_{\text{Reconstruction wrt. } a} + \underbrace{\mathbb{E}_{q_{\phi_{\text{ab}}}(z|a,b)}\log(p_{\theta_{\text{b}}}(b|z))}_{\text{Reconstruction wrt. } b} \tag{27}
$$

A regularization for the joint encoder $q_{\phi_{\text{ab}}}$ and two reconstruction terms, one for each decoder $p_{\theta_{\text{a}}}$ and $p_{\theta_{\text{b}}}$.

IV. Joint Multi-Modal Variational Autoencoder via Variation of Information

The issue with the joint VAE is the lacking possibility of encoding just one modality a or b. Thus, Suzuki et al. [2] exploit the *Variation of Information* (VI) and derive the evidence lower bound wrt. the VI.

First, the conditional probability is investigated

$$
\begin{aligned}
p(a|b) &= \frac{p(z,a|b)}{p(z|a,b)} && \text{Equation 85} && (28)\\
&= \frac{1}{p(z|a,b)} p(z,a|b) && && (29)\\
&= \frac{1}{p(z|a,b)} \frac{p(z,a,b)}{p(b)} && \text{Equation 84} && (30)\\
&= \frac{1}{p(z|a,b)} \frac{p(a,b|z)p(z)}{p(b)} && \text{Equation 84} && (31)\\
&= \frac{1}{p(z|a,b)} \frac{p(a|z)p(b|z)p(z)}{p(b)} && \text{Equation 90} && (32)\\
&= \frac{1}{p(z|a,b)} \frac{p(a|z)p(z|b)\frac{p(b)}{p(z)}p(z)}{p(b)} && \text{Equation 84} && (33)\\
&= \frac{p(a|z)p(z|b)}{p(z|a,b)} && && (34)
\end{aligned}
$$

Further, the marginal log-likelihood of a conditional distribution can be written as:

$$
\begin{aligned}
L_{\mathrm{M_a}} &= \log(p(a|b)) && && (35)\\
&= \sum_z q(z|a,b) \log(p(a|b)) && \text{Equation 89 w/o conditional} && (36)\\
&= \sum_z q(z|a,b) \log\left(\frac{p(z,a|b)}{p(z|a,b)}\right) && \text{Equation 85} && (37)\\
&= \sum_z q(z|a,b) \log\left(\frac{p(z,a|b)}{p(z|a,b)} \frac{q(z|a,b)}{q(z|a,b)}\right) && \text{multiplied by 1} && (38)\\
&= \sum_z q(z|a,b) \log\left(\frac{p(z,a|b)}{q(z|a,b)} \frac{q(z|a,b)}{p(z|a,b)}\right) && \text{reordered} && (39)\\
&= \sum_z q(z|a,b) \log\left(\frac{p(z,a|b)}{q(z|a,b)}\right) + \sum_z q(z|a,b) \log\left(\frac{q(z|a,b)}{p(z|a,b)}\right) && \text{Equation 86} && (40)\\
&= \mathcal{L}_{\mathrm{M_a}} + \mathrm{D_{KL}}(q(z|a,b) \| p(z|a,b)) && \text{Equation 92 \& 88} && (41)\\
&\geq \mathcal{L}_{\mathrm{M_a}} && && (42)
\end{aligned}
$$

Further, the log-likelihood of the VI can be written as

$$
\begin{aligned}
L_{\mathrm{M}} &= L_{\mathrm{M_a}} + L_{\mathrm{M_b}} && && (43)\\
&= \log(p(a|b)) + \log(p(b|a)) && && (44)\\
&= \mathcal{L}_{\mathrm{M_a}} + \mathcal{L}_{\mathrm{M_b}} + 2\,\mathrm{D_{KL}}(q(z|a,b) \| p(z|a,b)) && \text{Equation 41} && (45)\\
&\geq \mathcal{L}_{\mathrm{M_a}} + \mathcal{L}_{\mathrm{M_b}} && && (46)
\end{aligned}
$$

$$\mathcal{L}_{\mathrm{M_a}} + \mathcal{L}_{\mathrm{M_b}} = \sum_z q(z|a,b) \log\left(\frac{p(z,a|b)}{q(z|a,b)}\right) + \sum_z q(z|a,b) \log\left(\frac{p(z,b|a)}{q(z|a,b)}\right) \tag{47}$$

$$= \sum_z q(z|a,b) \log\left(\frac{p(a|z)p(z|b)}{q(z|a,b)}\right) + \sum_z q(z|a,b) \log\left(\frac{p(b|z)p(z|a)}{q(z|a,b)}\right) \quad \text{Equation 34} \tag{48}$$

$$= \sum_z q(z|a,b) \log\left(\frac{p(a|z)}{q(z|a,b)}\right) + \sum_z q(z|a,b) \log\left(\frac{p(z|b)}{q(z|a,b)}\right) \tag{49}$$

$$+ \sum_z q(z|a,b) \log\left(\frac{p(b|z)}{q(z|a,b)}\right) + \sum_z q(z|a,b) \log\left(\frac{p(z|a)}{q(z|a,b)}\right) \quad \text{reordering} \tag{50}$$

$$= \mathbb{E}_{q(z|a,b)} \log(p(a|z)) - \mathrm{D_{KL}}(q(z|a,b) \| p(z|b)) \tag{51}$$

$$+ \mathbb{E}_{q(z|a,b)} \log(p(b|z)) - \mathrm{D_{KL}}(q(z|a,b) \| p(z|a)) \quad \text{Equation 92} \tag{52}$$

$$= \mathbb{E}_{q(z|a,b)} \log(p(a|z)) - \mathrm{D_{KL}}(q(z|a,b) \| p(z|b)) \tag{53}$$

$$+ \mathbb{E}_{q(z|a,b)} \log(p(b|z)) - \mathrm{D_{KL}}(q(z|a,b) \| p(z|a)) \tag{54}$$

$$+ \mathrm{D_{KL}}(q(z|a,b) \| p(z|a,b)) - \mathrm{D_{KL}}(q(z|a,b) \| p(z|a,b)) \quad \text{added 0} \tag{55}$$

$$= \mathcal{L}_{\mathrm{J}} - \mathrm{D_{KL}}(q(z|a,b) \| p(z|b)) - \mathrm{D_{KL}}(q(z|a,b) \| p(z|a)) + \mathrm{D_{KL}}(q(z|a,b) \| p(z|a,b)) \quad \text{substitute eq. 26} \tag{56}$$

$$\geq \mathcal{L}_{\mathrm{J}} - \mathrm{D_{KL}}(q(z|a,b) \| p(z|b)) - \mathrm{D_{KL}}(q(z|a,b) \| p(z|a)) \tag{57}$$

$$=: \mathcal{L}_{\mathrm{M}} \tag{58}$$

With respect to Equation 27, the following regularization terms can be identified:

$$\mathcal{L}_{\mathrm{M}} = \mathcal{L}_{\mathrm{J}} - \underbrace{\mathrm{D_{KL}}(q_{\phi_{\mathrm{ab}}}(z|a,b) \| q_{\phi_{\mathrm{b}}}(z|b))}_{\text{Unimodal PDF fitting of encoder b}} - \underbrace{\mathrm{D_{KL}}(q_{\phi_{\mathrm{ab}}}(z|a,b) \| q_{\phi_{\mathrm{a}}}(z|a))}_{\text{Unimodal PDF fitting of encoder a}} \tag{59}$$

A. Conclusion

The introduced KL regularization by Suzuki et al. [2] tries to find a mean representative of all parameters between clusters in the latent space. This is absolutely correct for the mean values but is insufficient for all other statistics of the distribution that are estimated by q_{ϕ^*}. The approach would be sufficient, iff the true latent joint probability would be known, because only then the KL divergence is able to adapt to it in a correct manner. Unfortunately, this is not the case and thus all the estimated statistic values of the uni-modal encoders q_{ϕ^*}, except the mean, are questionable.

V. Proposed Joint Multi-Modal Autoencoder

A. The Variational Bound

$$
\begin{aligned}
L_{\text{PJ}} &= \log(p(a,b)) && && (60)\\
&= \frac{2}{2}\log(p(a,b)) && \text{multiplied by 1} && (61)\\
&= \frac{1}{2}\log\big(p(a,b)^2\big) && \text{Equation 87} && (62)\\
&= \frac{1}{2}\log(p(a,b)p(a,b)) && && (63)\\
&= \frac{1}{2}\log(p(b)p(a|b)p(b|a)p(a)) && \text{Equation 84} && (64)\\
&= \frac{1}{2}(\log(p(a)) + \log(p(b|a)) + \log(p(a|b)) + \log(p(b))) && \text{Equation 86} && (65)\\
&\geq \frac{1}{2}(\mathcal{L}_{\text{a}} + \mathcal{L}_{\text{M}_\text{a}} + \mathcal{L}_{\text{M}_\text{b}} + \mathcal{L}_{\text{b}}) && \text{Equation 8 \& 46} && (66)\\
&\geq \frac{1}{2}(\mathcal{L}_{\text{a}} + \mathcal{L}_{\text{M}} + \mathcal{L}_{\text{b}}) && \text{Equation 59} && (67)\\
&:= \mathcal{L}_{\text{PJ}} && && (68)
\end{aligned}
$$

B. Approximating Inference (i.e. rewriting $\mathcal{L}_{PJ}$)

$$
\begin{aligned}
2\mathcal{L}_{\text{PJ}} &= \mathcal{L}_{\text{a}} + \mathcal{L}_{\text{M}} + \mathcal{L}_{\text{b}} && && (69)\\
&= -\,\mathrm{D}_{\mathrm{KL}}(q(z|a)\|p(z)) + \mathbb{E}_{q(z|a)}\log(p(a|z)) && \text{Equation 12} && (70)\\
&\quad -\,\mathrm{D}_{\mathrm{KL}}(q(z|a,b)\|p(z)) + \mathbb{E}_{q(z|a,b)}\log(p(a|z)) + \mathbb{E}_{q(z|a,b)}\log(p(b|z)) && \text{Equation 26} && (71)\\
&\quad -\,\mathrm{D}_{\mathrm{KL}}(q(z|a,b)\|p(z|b)) - \mathrm{D}_{\mathrm{KL}}(q(z|a,b)\|p(z|a)) && \text{Equation 113} && (72)\\
&\quad -\,\mathrm{D}_{\mathrm{KL}}(q(z|b)\|p(z)) + \mathbb{E}_{q(z|b)}\log(p(b|z)) && \text{Equation 12} && (73)
\end{aligned}
$$

Applying the corresponding function approximators, the formula can be written as:

$$
\begin{aligned}
2\mathcal{L}_{\text{PJ}} &= \mathcal{L}_{\text{a}} + \mathcal{L}_{\text{M}} + \mathcal{L}_{\text{b}} && && (74)\\
&= -\,\mathrm{D}_{\mathrm{KL}}(q_{\phi_\text{a}}(z|a)\|p(z)) + \mathbb{E}_{q_{\phi_\text{a}}(z|a)}\log(p_{\theta_\text{a}}(a|z)) && \text{Equation 13} && (75)\\
&\quad -\,\mathrm{D}_{\mathrm{KL}}(q_{\phi_\text{ab}}(z|a,b)\|p(z)) + \mathbb{E}_{q_{\phi_\text{ab}}(z|a,b)}\log(p_{\theta_\text{a}}(a|z)) + \mathbb{E}_{q_{\phi_\text{ab}}(z|a,b)}\log(p_{\theta_\text{b}}(b|z)) && \text{Equation 27} && (76)\\
&\quad -\,\mathrm{D}_{\mathrm{KL}}(q_{\phi_\text{ab}}(z|a,b)\|q_{\phi_b}(z|b)) - \mathrm{D}_{\mathrm{KL}}(q_{\phi_\text{ab}}(z|a,b)\|q_{\phi_\text{a}}(z|a)) && \text{Equation 59} && (77)\\
&\quad -\,\mathrm{D}_{\mathrm{KL}}(q_{\phi_\text{b}}(z|b)\|p(z)) + \mathbb{E}_{q_{\phi_b}(z|b)}\log(p_{\theta_\text{b}}(b|z)) && \text{Equation 13} && (78)
\end{aligned}
$$

Investigating every line of the formula, the following properties can be identified: Equation 76 is the common multi-modal VAE loss derived from the joint probability, while Equation 77 adds the feature introduced by Suzuki et al. [2]. It introduces the KL regularization that brings the posterior distribution of an uni-modal encoder close to the distribution of the multi-modal case. The drawback of this approach is discussed in Section IV-A. The new lines, i.e. Equation 75 and 78, introduce the regularization of the uni-modal encoders wrt. the common prior and the reconstruction loss. The regularizer cares about the fact, that the uni-modal distribution does not deviate to much from the common prior while the reconstruction term shapes the remaining statistics including the mean. However, the last fact is very important, while the mean value in latent space might not be the best representative of the likelihood (i.e. the decoded data). This property cannot be respected by the KL divergence, but by the introduced reconstruction term.

1) of General Expression for Arbitrary Number of Modalities: Comprising the applied steps to derive $\mathcal{L}_{\text{PJ}}$ from the former section, we can identify that by successively applying logarithm and Bayes rules, we derive the ELBO for the proposed multi-modal VAE as follows: First, given the independent set of observable modalities $\mathcal{M} = \{\text{a,b,c,}\ldots\}$, its marginal log-likelihood $\log p(\mathcal{M}) =: L_{\text{M}^2}$ is multiplied by the cardinality of the set as the neutral element $1 = \frac{|\mathcal{M}|}{|\mathcal{M}|}$. Second, applying logarithm

multiplication rule, the nominator is written as the argument's exponent. Third, Bayes rule is applied to each term wrt. the remaining observable modalities to derive their conditionals. Therefore, we can write

$$L_{\mathrm{M}^2{}_{\mathcal{M}}} = \log p(\mathcal{M}) \overset{\text{mul. 1}}{=} \frac{|\mathcal{M}|}{|\mathcal{M}|} \log p(\mathcal{M}) \overset{\text{log. mul.}}{=} \frac{1}{|\mathcal{M}|} \log p(\mathcal{M})^{|\mathcal{M}|} \tag{79}$$

$$\overset{\text{Bayes}}{=} \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \log p(\mathcal{M} \setminus m) p(m | \mathcal{M} \setminus m) \tag{80}$$

$$\overset{\text{log. add}}{=} \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \log p(\mathcal{M} \setminus m) + \log p(m | \mathcal{M} \setminus m). \tag{81}$$

The expression $\sum_{m \in \mathcal{M}} \log p(m | \mathcal{M} \setminus m)$ is the general form of the marginal log-likelihood for the *variation of information* (VI), as introduced by [2] for the JMVAE, for any set $\mathcal{M}$. Thus, it can be directly substituted with $L_{\mathrm{M}_{\mathcal{M}}}$. The expression $\sum_{m \in \mathcal{M}} \log p(\mathcal{M} \setminus m)$ is the combination of all joint log-likelihoods of the subsets of $\mathcal{M}$ which have one less element. Therefore, this term can be rewritten as

$$\sum_{m \in \mathcal{M}} \log p(\mathcal{M} \setminus m) = \sum_{\widetilde{m} \in \widetilde{\mathcal{M}}} \log p(\widetilde{m}) \tag{82}$$

with $\widetilde{\mathcal{M}} = \{m | m \in \mathcal{P}(\mathcal{M}), |m| = |\mathcal{M}| - 1\}$ Finally, $\log p(\widetilde{m})$ can be substituted by $L_{\mathrm{M}^2{}_{\widetilde{m}}}$ without loss of generality. However, it is worth noticing that substitution stops at the end of recursion and therefore, all final expressions $\log p(\widetilde{m}) \; \forall \; |\widetilde{m}| \equiv 1$ remain. □

This results in the final recursive log-likelihood expression from which the ELBO can be directly derived as follows:

$$L_{\mathrm{M}^2{}_{\mathcal{M}}} = \frac{1}{|\mathcal{M}|} \left(L_{\mathrm{M}_{\mathcal{M}}} + \sum_{\widetilde{m} \in \widetilde{\mathcal{M}}} L_{\mathrm{M}^2{}_{\widetilde{m}}} \right) \geq \frac{1}{|\mathcal{M}|} \left(\mathcal{L}_{\mathrm{M}_{\mathcal{M}}} + \sum_{\widetilde{m} \in \widetilde{\mathcal{M}}} \mathcal{L}_{\mathrm{M}^2{}_{\widetilde{m}}} \right) =: \mathcal{L}_{\mathrm{M}^2{}_{\mathcal{M}}}. \tag{83}$$

VI. Appendix

Variants of Bayes equation:

$$p(a) = \frac{p(z,a)}{p(z|a)}, \qquad p(z|a) = \frac{p(z,a)}{p(a)}, \qquad p(z,a) = \frac{p(a)}{p(z,a)} \tag{84}$$

$$p(a|b,c) \overset{\text{eq. 84}}{=} \frac{p(a,b|c)}{p(b|c)} \overset{\text{eq. 84}}{=} \frac{p(a,b,c)}{p(b|c)p(c)} \overset{\text{eq. 84}}{=} \frac{p(a,b,c)}{p(b,c)} \tag{85}$$

Logarithm rules:

$$\log(ab) = \log(a) + \log(b) \tag{86}$$

$$a\log(b) = \log(b^a) \tag{87}$$

Evidence lower bound:

$$\mathcal{L} = \sum_z q(z|a) \log\left(\frac{p(z,a)}{q(z|a)}\right) \tag{88}$$

Marginal likelihood:

$$p(a|b) = \sum_z p(a|z)p(z|b) \tag{89}$$

Independent and identically distributed random variables (i.i.d. or iid or IID):

$$p(a,b,c) = p(a)p(b)p(c) \tag{90}$$

A. Kulbeck-Leibler Divergence

$$\mathrm{D}_{\mathrm{KL}}(q(z|a)\|p(z|a)) = \sum_z q(z|a) \log\left(\frac{q(z|a)}{p(z|a)}\right) \tag{91}$$

$$\mathrm{D}_{\mathrm{KL}}(\mathcal{N}_1(\mu_1,\sigma_1)\|\mathcal{N}_2(\mu_2,\sigma_2)) = \log(\sigma_2) - \log(\sigma_1) + \frac{\sigma_1^2}{2\sigma_2^2} + \frac{(\mu_1-\mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \tag{92}$$

more tbd.

B. Variation of Information

1) Operator Names:

- $\mathrm{VI}(A,B)$: Variation of Information between some properties A and B
- $\mathrm{I}(A)$: Information of A (or mutual information)
- $\mathrm{I}(A,B)$: *Mutual Information* (MI) of A and B
- $\mathrm{I}(A,B|C)$: *Mutual Conditional Information* (MCI) of A and B given C
- $\mathrm{H}(A)$: Entropy of A
- $\mathrm{H}(A,B)$: *Joint Entropy* (JE) of A and B
- $\mathrm{H}(A|B)$: *Conditional Entropy* (CE) of A given B

The *Variation of Information* (VI) between some random variables can be written as

$$\mathrm{VI}(A,B) = \mathrm{H}(A) + \mathrm{H}(B) - 2\,\mathrm{I}(A,B) = \mathrm{H}(A|B) + \mathrm{H}(B|A) \tag{93}$$

and

$$\mathrm{VI}(A,B,C) = \mathrm{H}(A) + \mathrm{H}(B) + \mathrm{H}(C) - 3\,\mathrm{I}(A,B) = \mathrm{H}(A|B,C) + \mathrm{H}(B|A,C) + \mathrm{H}(C|A,B) \tag{94}$$

and so on ....

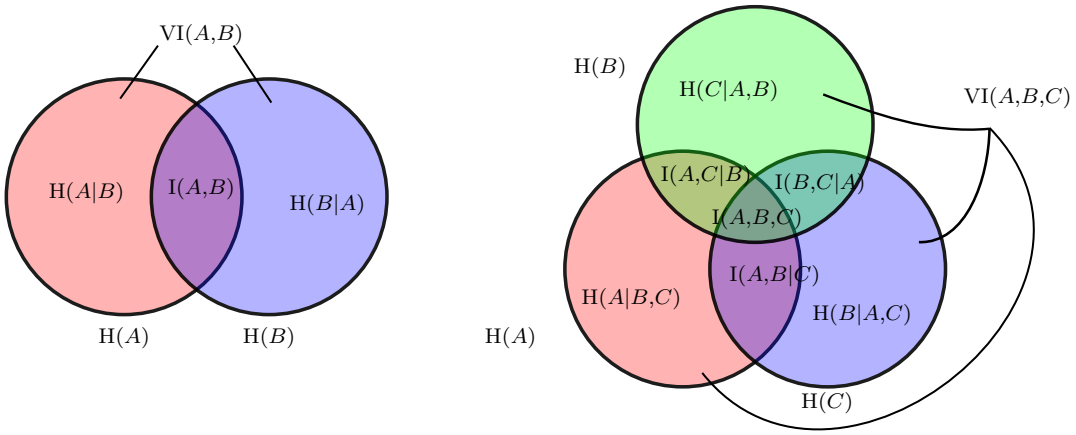

Fig. 2: Visualization of VI as Venn digram.

C. Extension to three Modalities

It should be clear that both approaches, proposed and Suzuki's [2], can be extended to multiple modalities. In the following, an example for three modalities is given.

First, the conditional probability is investigated:

$$
\begin{aligned}
p(a|b,c) &= \frac{p(a,b,c,z)}{p(a,b,c,z)}\frac{p(a,b,c)}{p(b,c)} && \text{multiplied by 1 \& Equation 85} && (95)\\
&= \frac{p(z,a|b,c)}{p(z|a,b,c)} && \text{Equation 85} && (96)\\
&= \frac{1}{p(z|a,b,c)}p(z,a|b,c) && \text{reorder} && (97)\\
&= \frac{1}{p(z|a,b,c)}\frac{p(z,a,b,c)}{p(b,c)} && \text{Equation 85} && (98)\\
&= \frac{1}{p(z|a,b,c)}\frac{p(a,b,c|z)p(z)}{p(b,c)} && \text{Equation 90} && (99)\\
&= \frac{1}{p(z|a,b,c)}\frac{p(a|z)p(b,c|z)p(z)}{p(b,c)} && \text{Equation 85} && (100)\\
&= \frac{1}{p(z|a,b,c)}\frac{p(a|z)p(z|c,b)\frac{p(c,b)}{p(z)}p(z)}{p(b,c)} && \text{Equation 85} && (101)\\
&= \frac{p(a|z)p(z|c,b)}{p(z|a,b,c)} && && (102)
\end{aligned}
$$

The log-likelihood of a single joint distribution can be written as:

$$
\begin{aligned}
\log(p(a|b,c)) &= \sum_z q(z|a,b,c)\log\left(\frac{p(z,a|b,c)}{q(z|a,b,c)}\right) + \sum_z q(z|a,b,c)\log\left(\frac{q(z|a,b,c)}{p(z|a,b,c)}\right) && (103)\\
&= \mathcal{L}_{\widetilde{\mathrm{M}}_\mathrm{a}} + \mathrm{D}_{\mathrm{KL}}(q(z|a,b,c)\|p(z|a,b,c)) && (104)\\
&\geq \mathcal{L}_{\widetilde{\mathrm{M}}_\mathrm{a}} && (105)
\end{aligned}
$$

1) JMVAE for three Modalities: The log-likelihood of the VI between three distributions can be written as:

$$
\begin{aligned}
L_{3\mathrm{M}} &= \log(p(a|b,c)) + \log(p(b|a,c)) + \log(p(c|b,c)) && (106)\\
&= \mathcal{L}_{\widetilde{\mathrm{M}}_\mathrm{a}} + \mathcal{L}_{\widetilde{\mathrm{M}}_\mathrm{b}} + \mathcal{L}_{\widetilde{\mathrm{M}}_\mathrm{c}} + 3\,\mathrm{D}_{\mathrm{KL}}(q(z|a,b,c)\|p(z|a,b,c)) && \text{Equation 41} && (107)\\
&\geq \mathcal{L}_{\widetilde{\mathrm{M}}_\mathrm{a}} + \mathcal{L}_{\widetilde{\mathrm{M}}_\mathrm{b}} + \mathcal{L}_{\widetilde{\mathrm{M}}_\mathrm{c}} && && (108)
\end{aligned}
$$

The combined ELBO can then be rewritten as

$$\mathcal{L}_{\widetilde{\mathrm{M}}_\mathrm{a}} + \mathcal{L}_{\widetilde{\mathrm{M}}_\mathrm{b}} + \mathcal{L}_{\widetilde{\mathrm{M}}_\mathrm{c}} = \mathbb{E}_{q(z|a,b,c)} \log(p(a|z)) - \mathrm{D}_{\mathrm{KL}}(q(z|a,b,c) \| p(z|c,b)) \tag{109}$$

$$+ \mathbb{E}_{q(z|a,b,c)} \log(p(b|z)) - \mathrm{D}_{\mathrm{KL}}(q(z|a,b,c) \| p(z|a,c)) \tag{110}$$

$$+ \mathbb{E}_{q(z|a,b,c)} \log(p(c|z)) - \mathrm{D}_{\mathrm{KL}}(q(z|a,b,c) \| p(z|a,b)) \qquad \text{Equation 92} \tag{111}$$

$$\geq \mathcal{L}_{\widetilde{\mathrm{J}}} - \mathrm{D}_{\mathrm{KL}}(q(z|a,b,c) \| p(z|b,c)) \tag{112}$$

$$- \mathrm{D}_{\mathrm{KL}}(q(z|a,b,c) \| p(z|a,c)) - \mathrm{D}_{\mathrm{KL}}(q(z|a,b,c) \| p(z|b,c)) \tag{113}$$

$$:= \mathcal{L}_{\widetilde{\mathrm{M}}} \tag{114}$$

$\mathcal{L}_{\widetilde{\mathrm{J}}}$ is the joint ELBO of a joint probability distribution having three arguments (i.e. a, b, c). The derivation is analog to Section III. The next steps are the application of encoders and decoders for this network which is straight forward and should be clear to the reader.

However, if we investigate the last equations, the following properties can be identified: There are the common reconstruction terms ($\mathbb{E}$) for each decoder $p(*|z)$ wrt. the full multi-modal decoder $q(z|a,b,c)$. The KL terms show the **drawback** of the VI approach. As before, these regularizer tend to bring the encoders' distribution to match each other. But only pairwise encoders (e.g. $p(z|a,b)$) remain and thus, uni-modal encoders are neglected.

This means from a practical point of view, that when we have N modalities in a setup, we only can build derived setups having $N-1$ modalities.

2) Proposed JMVAE for three Modalities: The derivation from the joint log likelihood can be written analogously:

$$\log(p(a,b,c)) = \frac{3}{3} \log(p(a,b,c)) \tag{115}$$

$$= \frac{1}{3} \log\big(p(a,b,c)^3\big) \tag{116}$$

$$= \frac{1}{3} \log(p(a,b,c)p(a,b,c)p(a,b,c)) \tag{117}$$

$$= \frac{1}{3} \log(p(a,b)p(b,c)p(a,c)p(a|b,c)p(b|a,c)p(c|a,b)) \tag{118}$$

$$= \frac{1}{3} (\log(p(a,b)) + \log(p(b,c)) + \log(p(a,c)) + \log(p(a|b,c)) + \log(p(b|a,c)) + \log(p(c|a,b))) \tag{119}$$

$$= \frac{1}{3} \left(\frac{2}{2} (\log(p(a,b)) + \log(p(b,c)) + \log(p(a,c))) + \log(p(a|b,c)) + \log(p(b|a,c)) + \log(p(c|a,b)) \right) \tag{120}$$

$$= \frac{1}{6} \big(\log\big(p(a,b)^2\big) + \log\big(p(b,c)^2\big) + \log\big(p(a,c)^2\big) \big) \tag{121}$$

$$+ \frac{1}{3} (\log(p(a|b,c)) + \log(p(b|a,c)) + \log(p(c|a,b))) \tag{122}$$

$$= \frac{1}{6} (\log(p(a)p(b)p(a|b)p(b|a)) + \log(p(c)p(b)p(c|b)p(b|c)) + \log(p(a)p(c)p(a|c)p(c|a))) \tag{123}$$

$$+ \frac{1}{3} (\log(p(a|b,c)) + \log(p(b|a,c)) + \log(p(c|a,b))) \tag{124}$$

$$= \frac{1}{6} (\log(p(a|b)) + \log(p(b|a)) + \log(p(c|b)) + \log(p(b|c)) + \log(p(a|c)) + \log(p(c|a))) \tag{125}$$

$$+ \frac{1}{3} (\log(p(a)) + \log(p(b)) + \log(p(c)) + \log(p(a|b,c)) + \log(p(b|a,c)) + \log(p(c|a,b))) \tag{126}$$

It is now straight forward, by applying all former mentioned equations, to derive the ELBO for the above marginal log-likelihood. As one can imagine, the above equation results in a pretty heavy loss term but with the big advantage of respecting all permutations of modalities.

This means again from a practical point of view, in comparison to the approach by Suzuki et al. [2], that we can build arbitrary sensor setups having 1 to N modalities.

REFERENCES

[1] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *CoRR*, vol. abs/1312.6, 2013.

[2] M. Suzuki, K. Nakayama, and Y. Matsuo, "Joint multimodal learning with deep generative models," pp. 1–12, 2017.