Multimodal Variational Auto-encoder based Audio-Visual Segmentation –Supplementary Materials –

Yuxin Mao[1] Jing Zhang[2] Mochu Xiang[1] Yiran Zhong[3] Yuchao Dai[1†]
[1]Northwestern Polytechnical University & Shaanxi Key Laboratory of Information Acquisition and Processing
[2]Australian National University [3]Shanghai AI Laboratory
https://github.com/OpenNLPLab/MMVAE-AVS https://npucvr.github.io/MMVAE-AVS

Abstract

In this supplementary material, we provide the derivation process of the Conditional multimodal VAE and the Latent Space Factorization. Afterward, we describe the details of the model implemented in the ablation experiments and give a structural diagram. Finally, we show the detailed structure of the three latent encoders to facilitate understanding.

1. Conditional Multimodal VAE

We describe in detail the derivation process of Conditional multimodal VAE [6–9, 11, 13–15] in this section.

1.1. Conditional VAE

For a conditional latent variable model with three variables x (the conditional variable), y (the output) and z (the latent variable), the generative process is as follows:

- Given input x, the latent variable z is drawn from the prior distribution $p_\theta(z|x)$.
- The output is generated via $p_\theta(y|x,z)$.

The inference process is then to infer informative values of the latent variable given the observed data by computing the posterior $p_\theta(z|x,y)$, which is defined as:

$$p_\theta(z|x,y) = \frac{p(x,y,z)}{p(x,y)}, \tag{1}$$

where $p(x,y) = \int p(x,y|z)p(z)dz$, which involves integral over all configurations of latent variables z, leading to intractable computation.

† Corresponding author (daiyuchao@gmail.com).
This work was done when Yuxin Mao was an intern at Shanghai AI Laboratory.

To achieve computationally tractable learning, [7, 11] introduces a recognition model (or the inference model) $q_\phi(z|x,y)$ as an approximation of the intractable true posterior $p_\theta(z|x,y)$. The goal is then finding the variational parameters ϕ that minimize the Kullback–Leibler (KL) divergence between the variational posterior $q_\phi(z|x,y)$ and the true posterior $p_\theta(z|x,y)$ via:

$$\phi^* = \arg\min_\phi D_{\mathrm{KL}}(q_\phi(z|x,y)\|p_\theta(z|x,y)), \tag{2}$$

where the KL-divergence term in Eq. 2 can be decomposed:

$$\begin{aligned} &D_{\mathrm{KL}}(q_\phi(z|x,y)\|p_\theta(z|x,y)) \\ &= \mathbb{E}_{q_\phi(z|x,y)} \log q_\phi(z|x,y) - \mathbb{E}_{q_\phi(z|x,y)} \log p_\theta(z|x,y) \\ &= \mathbb{E}_{q_\phi(z|x,y)} \log q_\phi(z|x,y) - \mathbb{E}_{q_\phi(z|x,y)} \log \frac{p_\theta(x,y,z)}{p_\theta(x,y)}. \end{aligned} \tag{3}$$

Based on Bayes' rule, we have:

$$p_\theta(x,y,z) = p_\theta(y|x,z)p_\theta(z|x)p(x). \tag{4}$$

We can then decompose the second expectation term in Eq. 3 as:

$$\begin{aligned} &\mathbb{E}_{q_\phi(z|x,y)} \log \frac{p_\theta(x,y,z)}{p_\theta(x,y)} \\ &= \mathbb{E}_{q_\phi(z|x,y)} \log \frac{p_\theta(y|x,z)p_\theta(z|x)p_\theta(x)}{p_\theta(x,y)} \\ &= \mathbb{E}_{q_\phi(z|x,y)} \log p_\theta(y|x,z) + \mathbb{E}_{q_\phi(z|x,y)} \log p_\theta(z|x) \\ &+ \mathbb{E}_{q_\phi(z|x,y)} \log \frac{p_\theta(x)}{p_\theta(x,y)} \\ &= \mathbb{E}_{q_\phi(z|x,y)} \log p_\theta(y|x,z) + \mathbb{E}_{q_\phi(z|x,y)} \log p_\theta(z|x) \\ &+ \mathbb{E}_{q_\phi(z|x,y)} \log \frac{p_\theta(x)}{p_\theta(y|x)p_\theta(x)} \\ &= \mathbb{E}_{q_\phi(z|x,y)} \log p_\theta(y|x,z) + \mathbb{E}_{q_\phi(z|x,y)} \log p_\theta(z|x) \\ &- \log p_\theta(y|x). \end{aligned} \tag{5}$$

We take Eq. 5 back to Eq. 3 and obtain:

$$
\begin{aligned}
&D_{\mathrm{KL}}(q_\phi(z|x,y)\|p_\theta(z|x,y)) \\
&= \mathbb{E}_{q_\phi(z|x,y)} \log q_\phi(z|x,y) - \mathbb{E}_{q_\phi(z|x,y)} \log p_\theta(z|x) \\
&- \mathbb{E}_{q_\phi(z|x,y)} \log p_\theta(y|x,z) + \log p_\theta(y|x) \\
&= \underbrace{D_{KL}(q_\phi(z|x,y)\|p_\theta(z|x)) - \mathbb{E}_{q_\phi(z|x,y)} \log p_\theta(y|x,z)}_{\text{-ELBO}(x,y,\theta,\phi)} \\
&+ \log p_\theta(y|x).
\end{aligned} \tag{6}
$$

We simplify Eq. 6, and obtain:

$$
\begin{aligned}
&\log p_\theta(y|x) \\
&= \text{ELBO}(x,y,\theta,\phi) + D_{\mathrm{KL}}(q_\phi(z|x,y)\|p_\theta(z|x,y)).
\end{aligned} \tag{7}
$$

By Jensen's inequality, $D_{\mathrm{KL}}(q_\phi(z|x,y)\|p_\theta(z|x,y))$ in Eq. 7 is always greater or equal to zero. In this case, minimizing it can be achieved by maximizing $\text{ELBO}(x,y,\theta,\phi)$, which is the evidence lower bound (ELBO). With the reparameterization trick [7], the KL-divergence term in Eq. 7 can be solved in the closed form given that both the prior and posterior are Gaussian.

Following the maximum likelihood training pipeline, a conditional VAE (CVAE) is then trained to maximize the conditional log-likelihood of individual data points $\log p_\theta(y|x)$ via:

$$
\begin{aligned}
\theta^*, \phi^* &= \arg\max_{\theta,\phi} \log p_\theta(y|x) \\
&= \arg\max_{\theta,\phi} \text{ELBO}(x,y,\theta,\phi).
\end{aligned} \tag{8}
$$

1.2. Conditional Multimodal VAE

The unimodal variational auto-encoders (VAEs) [7, 11] are optimized by maximizing the evidence lower bound (ELBO), which includes a reconstruction term and the Kullback-Leibler (KL) divergence term to measure the divergence from the variational posterior to the prior distribution of the latent variable. With the reparameterization trick [7], KL-Divergence within the unimodal VAEs can be solved in closed form.

In our multimodal setting [12, 15], we obtain the same derivation as in Eq. 7, except that we change the unimodal data x to the multimodal data $X = \{\{x_t^v\}_{t=1}^T, x^a\}$, *i.e.* the visual $\{x_t^v\}_{t=1}^T$ for T non-overlapping yet continuous frames, audio x^a of the current clip. The ELBO of conditional multimodal VAE is then obtained as:

$$
\begin{aligned}
&\text{ELBO}(X,y,\theta,\phi) \\
&= D_{KL}(q_\phi(z|X,y)\|p_\theta(z|X)) - \mathbb{E}_{q_\phi(z|X,y)} \log p_\theta(y|X,z),
\end{aligned} \tag{9}
$$

where $q_\phi(z|X,y)$ and $p_\theta(z|X)$ represent the joint posterior and prior respectively. Product of experts (POE) [3, 5] is widely studied for estimation of the joint prior/posterior distributions. Specifically, for input multimodal data X and latent variable z, [15] obtains joint prior $p_\theta(z|X)$ and posterior $p_\theta(z|X,y)$ via product of Gaussian experts across the modalities. In this case, the KL divergence within ELBO is computed between two Gaussian distributions, leading to a closed form solution.

One main disadvantage of PoE based latent space factorization is that one miscalibrated expert will dominate the prediction, which can be detrimental to the whole model. Alternatively, Mixture of Expert (MoE) [10] is introduced to factorize the joint prior/posterior distribution as a combination of unimodal priors/posteriors, facilitates the optimization of individual expert. However, MoE [10] is computationally less efficient as the joint prior and posterior are not Gaussian anymore, thus the KL-Divergence within ELBO cannot be solved in closed form. Although importance sampling [2] is adopted to achieve tight ELBO, computational efficiency is reduced. The Jensen-Shannon (JS) divergence [12] based multimodal VAE in this paper is to achieve a trade-off between computational efficiency and prediction quality.

2. Latent Space Factorization

Conventionally, the objective of the JS divergence based conditional multimodal VAE is defined as:

$$
\begin{aligned}
&\mathcal{L}(X,y,\theta,\phi) \\
&= \mathbb{E}_{q_\phi(z|X,y)} \left[\log p_\theta(y|X,z)\right] - \text{JSD}(q_\phi(z|X,y), p_\theta(z|X)),
\end{aligned} \tag{10}
$$

which is maximized to obtain parameter estimation. As discussed in the manuscript, the audio data provides category information for the localization of the object in the video, whereas the visual data provides an object pool with precise structure information. In this case, we claim the shared information of both modalities can be reliable in localizing the target object(s). The latent space factorization is then used to factorize the latent code of each modality, achieving both shared representation learning c and modality-specific (s) representation learning.

The posterior latent code of each modal is then defined as $q(z) = q(s,c) = q_{\phi_s}(s|x,y)q_{\phi_c}(c|x,y)$ following [1]. Similarly, the prior latent code is obtained as: $p(s,c) = p_\theta(s|x)p_\theta(c|x)$, where we use x here to represent one modality of data. Based on this, the reconstruction part in Eq. 14 can be obtained as:

$$
\begin{aligned}
&\mathbb{E}_{q_\phi(z|X,y)} \left[\log p_\theta(y|X,z)\right] \\
&= \sum_{k=1}^{K} \mathbb{E}_{q_{\phi_c}(c|X,y)} \left[\gamma \mathbb{E}_{q_{\phi_{s^k}}(s^k|x^k,y)} \left[\log p_\theta(y|x^k,s^k,c)\right]\right],
\end{aligned} \tag{11}
$$

which is exactly the reconstruction part in Eq. (5) of the manuscript.

For our multimodal setting, the shared representation (c) models the multimodal representation, while the modality-

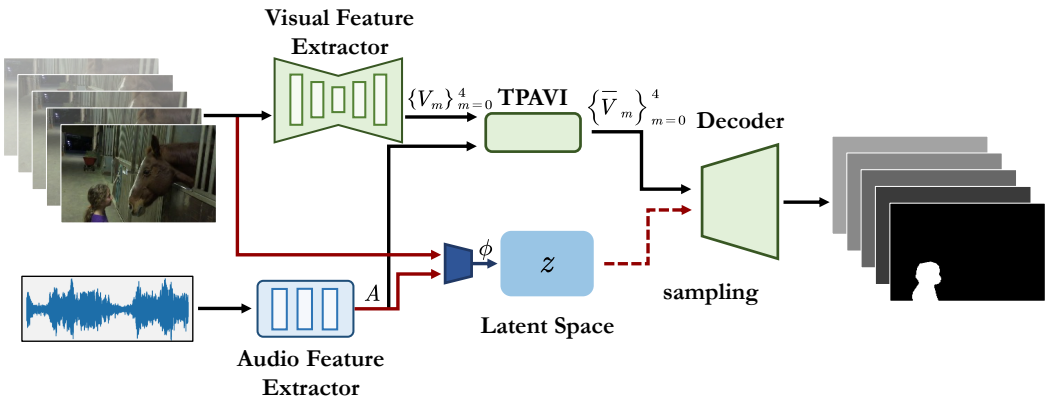

Figure 1. Overview of the model without the latent space factorization

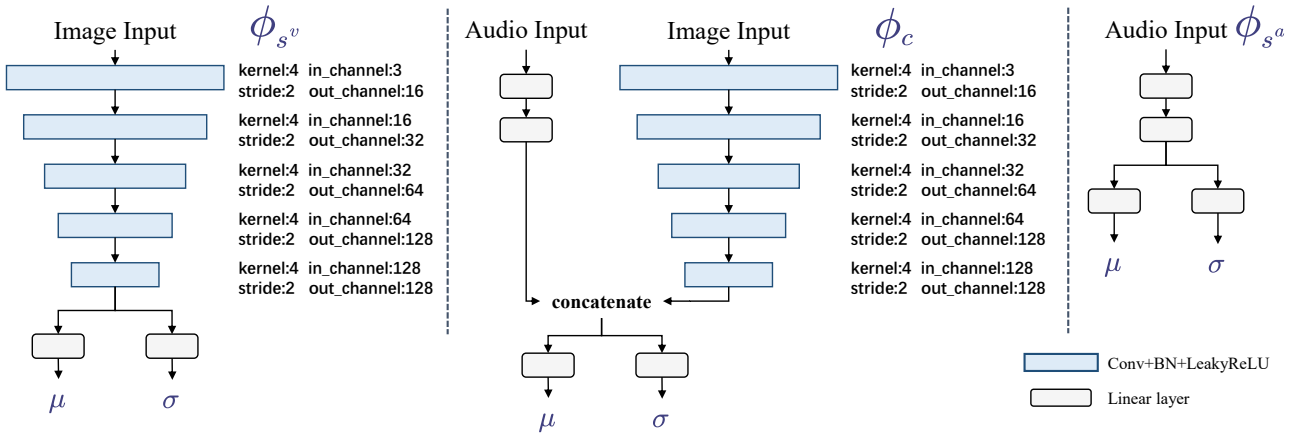

Figure 2. The detail structure of the three latent encoders $\phi_{s^v}, \phi_c, \phi_{s^a}$.

specific representation (s) is unrelated to the multimodal representation. In this case, the KL divergence in Eq. 9 can be derived as:

$$\begin{aligned} &D_{KL}(q_\phi(z|X,y)\|p_\theta(z|X)) \\ &= D_{KL}(q_\phi(s,c|X,y)\|p_\theta(s,c|X)) \\ &= D_{KL}(q_{\phi_s}(s|X,y)\|p_\theta(s|X)) + D_{KL}(q_{\phi_c}(c|X,y)\|p_f(c|X)), \end{aligned} \tag{12}$$

where $p_f(c|X)$ is the dynamic prior, which can be the geometric mean of the unimodal shared representations, or the $p_\theta(c|X)$ in Eq. (5) of the manuscript. Thus, $D_{KL}(q_{\phi_c}(c|X,y)\|p_f(c|X))$ computes the KL divergence of $q_{\phi_c}(c|X,y)$ from the geometric mean of the prior, which can be rewritten as JS divergence:

$$D_{KL}(q_{\phi_c}(c|X,y)\|p_f(c|X)) = \text{JSD}(q_{\phi_c}(c|X,y), p_\theta(c|X)), \tag{13}$$

which is exactly the JSD term in Eq. (5) of the manuscript.

Let's take Eq. 11 and Eq. 12 back to Eq. 9, we obtain the objective of the JS divergence based multimodal VAE with latent space factorization as:

$$\begin{aligned} &\widehat{\text{ELBO}}(X,y,\theta,\phi) \\ &= \sum_{k=1}^{K} \mathbb{E}_{q_{\phi_c}(c|X,y)} \left[\gamma \mathbb{E}_{q_{\phi_{s^k}}(s^k|x^k,y)} \left[\log p_\theta(y|x^k,s^k,c)\right]\right] \\ &- \beta \sum_{k=1}^{K} D_{KL}(q_{\phi_{s^k}}(s^k|x^k,y)||p_\theta(s^k|x^k)) \\ &- \beta \text{JSD}(q_{\phi_c}(c|X,y), p_\theta(c|X)), \end{aligned} \tag{14}$$

where hyper-parameters γ and β are used to achieve stable training as in [4].

3. More details of Ablation Studies

We have performed an ablation study of the effectiveness of the latent space factorization in the manuscript by training a model without the latent space factorization. We implement a model with one joint latent code to model the joint distribution of audio and visual. For a better comparison with our proposed model, we show a detailed diagram of the model structure without the latent space factoriza-

tion in Fig. 1. And results shown in Table 2 (*b*)-(*c*) in the manuscript demonstrate the effectiveness of the latent space factorization.

4. Detail structure of the latent encoders

We describe in detail the structure of the three latent encoders $\phi_{s^v}, \phi_{s^a}, \phi_c$ to facilitate understanding, as shown in Fig. 2.

References

[1] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 2

[2] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *International Conference on Learning Representations (ICLR)*, 2016. 2

[3] Yanshuai Cao and David J Fleet. Generalized product of experts for automatic and principled fusion of gaussian process predictions. *arXiv preprint arXiv:1410.7827*, 2014. 2

[4] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017. 3

[5] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800, 2002. 2

[6] Tom Joy, Yuge Shi, Philip Torr, Tom Rainforth, Sebastian M Schmon, and Siddharth N. Learning multimodal VAEs through mutual supervision. In *International Conference on Learning Representations (ICLR)*, 2022. 1

[7] Diederik Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014. 1, 2

[8] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 1

[9] Emile Mathieu, Tom Rainforth, Nana Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning (ICML)*, 2019. 1

[10] Yuge Shi, Brooks Paige, Philip Torr, et al. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

[11] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 1, 2

[12] Thomas M. Sutter, Imant Daunhawer, and Julia E. Vogt. Multimodal generative learning utilizing jensen-shannon-divergence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[13] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016. 1

[14] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 1

[15] Mike Wu and Noah D. Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 1, 2