Learning multi-modal generative models with permutation-invariant encoders and tighter variational objectives

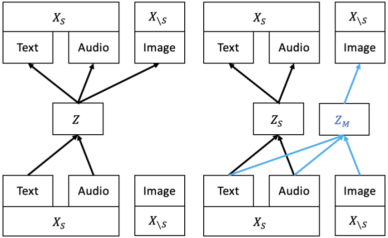
Marcel Hirt [1] Domenico Campolo [1] Victoria Leong [1] Juan-Pablo Ortega [1]

[1]Nanyang Technological University, Singapore

Overview

- Multi-modal generative models such as Variational Autoencoders (VAEs) aim to learn representations that capture shared content across multiple modalities in addition to modality-specific information.
- Various objective functions for such models have been suggested, often motivated as lower bounds on the multi-modal data log-likelihood or from information-theoretic considerations.
- Fixed aggregation schemes such as Product-of-Experts (PoE) or Mixture-of-Experts (MoE) are commonly used to encode latent variables from different modality subsets.
- In contrast to previous works, we consider a variational objective that can tightly approximate the multi-modal data log-likelihood (LLH).
- We develop more flexible aggregation schemes that avoid inductive biases in PoE or MoE approaches by combining encoded features from different modalities based on permutation-invariant neural networks.

Reconstruction and cross-modal prediction for multi-modal models



Previous bound Our objective

Previous mixture-based bound:

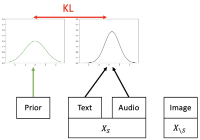
$$\mathbb{E}_{q_\phi(z|x_\mathcal{S})}\left[\log p_\theta(x|z)\right] - \beta \mathrm{KL}(q_\phi(z|x_\mathcal{S})|p_\theta(z)).$$

Our objective:

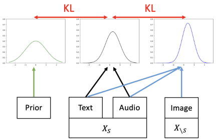
$$\mathbb{E}_{q_\phi(z|x_\mathcal{S})}\left[\log p_\theta(x_\mathcal{S}|z)\right] - \beta \mathrm{KL}(q_\phi(z|x_\mathcal{S})|p_\theta(z))$$
$$+\mathbb{E}_{q_\phi(z|x)}\left[\log \log p_\theta(x_{\setminus\mathcal{S}}|z)\right] - \beta \mathrm{KL}(q_\phi(z|x)|q_\phi(z|x_\mathcal{S})).$$

The mixture-based bound resorts to a single latent variable $Z \sim q_\phi(\cdot|x_\mathcal{S})$ that encodes information from a modality subset $x_\mathcal{S}$ and is trained to reconstruct conditioning modalities $x_\mathcal{S}$, and predict masked modalities $x_{\setminus\mathcal{S}}$. Our objective relies on two latent variables $Z_\mathcal{S} \sim q_\phi(\cdot|x_\mathcal{S})$, and $Z_\mathcal{M} \sim q_\phi(\cdot|x_\mathcal{S}, x_{\setminus\mathcal{S}})$, where $Z_\mathcal{S}$ learns to reconstruct all its conditioning modalities, and $Z_\mathcal{M}$ learns to reconstruct the remaining modalities.

KL-Regularization for multi-modal models and log-likelihood approximations



Previous bound



Our objective

Previous mixture-based bound $\mathcal{L}^{\text{Mix}}_{\theta,\phi}$:

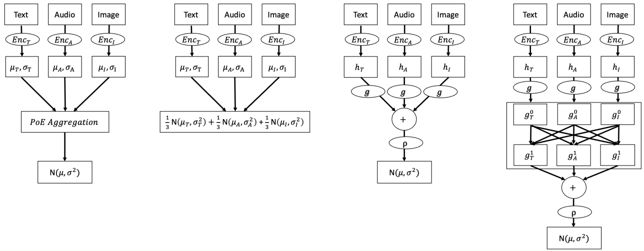
$$\mathbb{E}_{p_d(x)}[\log p_\theta(x)]\mathrm{d}x \geq \mathbb{E}_{p_d(x)}[\mathcal{L}^{\text{Mix}}_{\theta,\phi}(x)] + \mathcal{H}(p_d(X_{\setminus\mathcal{S}}|X_{\mathcal{S}})).$$

Our objective $\mathcal{L}_{\theta,\phi}$ for idealized encoders:

$$\mathbb{E}_{p_d(x)}[\log p_\theta(x)]\mathrm{d}x = \mathbb{E}_{p_d(x)}[\mathcal{L}_{\theta,\phi^\star}(x)].$$

In addition to minimizing the KL divergence between the encoding distribution given a modality subset $x_{\mathcal{S}}$ and a prior distribution in the mixture-based bound, our objective aims to minimize the KL divergence between the encoding distribution given all modalities relative to the encoding distribution of a modality subset $x_{\mathcal{S}}$. The conditional part of our objective is an approximation to the conditional log-likelihood $\log p_\theta(x_{\setminus\mathcal{S}}|x_{\mathcal{S}})$, and becomes a true lower bound only if $q_\phi(z|x_{\mathcal{S}}) = p_\theta(z|x_{\mathcal{S}})$.

Permutation-invariant modality encoding

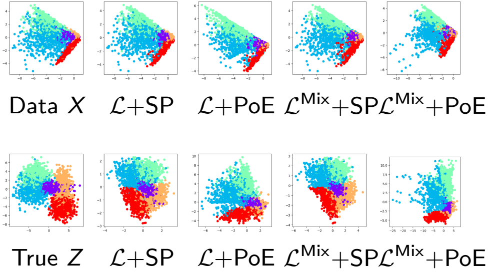


PoE MoE SumPooling SelfAttention

We want to learn encoding distributions for all modality subsets. The common scalable approach is to first encode each modality onto some feature H_s. These features are then aggregated using permutation-invariant functions.

We learn these aggregation schemes instead of previous approaches where these are given implictly via PoEs or MoEs. A Sum-Pooling (SP) model sums up the encoded features before projecting them onto the variational parameters. Pairwise interactions between the encoded modalities can be accounted for by adding a self-attention layer.

Weak identifiability of conditional or multi-modal models



Identifiability can be achieved in conditional models for the posterior distribution which is however not the optimal encoding distribution for the mixture based bound or satisfies a MoE/PoE aggregation scheme.

Auxiliary labels as modalities. The true latent variables are Gaussian distributed with means and variances modulated by the label modality (color-coded). The continuous data modality with reconstructions in the top row. True and inferred latent variables are shown in the bottom row with a linear transformation indeterminacy.

Table 1: Test LLH estimates for the joint data (M+S+T) and marginal data.

	Proposed objective				Mixture bound			
Aggregation	M+S+T	M	S	T	M+S+T	M	S	T
PoE+	6872 (9.62)	**2599 (5.6)**	4317 (1.1)	-9 (0.2)	5900 (10)	2449 (10.4)	3443 (11.7)	-19 (0.4)
PoE	6775 (54.9)	2585 (18.7)	4250 (8.1)	-10 (2.2)	5813 (1.2)	2432 (11.6)	3390 (17.5)	-19 (0.1)
MoE+	5428 (73.5)	2391 (104)	3378 (92.9)	-74 (88.7)	5420 (60.1)	2364 (33.5)	3350 (58.1)	-112 (133.4)
MoE	5597 (26.7)	2449 (7.6)	3557 (26.4)	-11 (0.1)	5485 (4.6)	2343 (1.8)	3415 (5.0)	-17 (0.4)
SumPooling	**7056 (124)**	2478 (9.3)	**4640 (114)**	**-6 (0.0)**	6130 (4.4)	2470 (10.3)	3660 (1.5)	-16 (1.6)
SelfAttention	**7011 (57.9)**	2508 (18.2)	**4555 (38.1)**	-7 (0.5)	6127 (26.1)	2510 (12.7)	3621 (8.5)	-13 (0.2)

For our numerical experiments, we find that our variational objective and more flexible aggregation models achieve higher log-likelihoods. Although the mixture-based bound can lead to inexact and average predictions or reconstructions, it can lead to improved cross-modal predictions.

Acknowledgments

This work is supported by funding from the Wellcome Leap 1kD Program and by the RIE2025 Human Potential Programme Prenatal/Early Childhood Grant (H22P0M0002), administered by A*STAR. The computational work for this article was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg).