Variational Mixture-of-Experts Autoencoders for Multi-modal Deep Generative Models
Yuge Shi*, N. Siddharth*, Brooks Paige†, Philip H.S. Torr
University of Oxford†, Alan Turing institute‡
{yshi, nsid, phst}@robots.ox.ac.uk, bpaige@turing.ac.uk

Overview
Our work focuses on learning generative models that capture common information between multiple modalities such as language and vision. We propose the following criteria for such model:
(a) Latent factorisation: latent space decomposes into shared and private subparts;
(b) Joint generation: generations in different modalities from the same latent coding are coherent;
(c) Cross generation: generating data in one modality conditioned on data from another modality while keeping the commonality;
(d) Synergy: the model’s performance is improved from observing multiple modalities compared to single modality.
In this work, we develop a generative model that satisfies the above criteria named MoE-Multimodal VAE (MMVAE).

Objective
MoE-ELBO We employ a VAE [1] to learn a multi-modal generative model over modalities \( m = 1, \ldots, M \) of the form \( p(\mathbf{x}; \mathbf{z}, \mathbf{m}) \). The objective is to maximise the marginal likelihood of the data \( p(\mathbf{x}; \mathbf{m}) \) through its evidence lower bound (ELBO):
\[
\mathcal{L}_{\text{ELBO}}(\mathbf{x}_m; \mathbf{z}_m) = \mathbb{E}_{q_{\phi}(\mathbf{z}_m | \mathbf{x}_m)} \left[ \log p(\mathbf{x}_m | \mathbf{z}_m) \right] - \mathbb{E}_{q_{\phi}(\mathbf{z}_m | \mathbf{x}_m)} \left[ \log q_{\phi}(\mathbf{z}_m | \mathbf{x}_m) \right]
\]
(1)
MoE-IWAE The importance weighted autoencoder (IWAE) objective, proposed by [2], computes a tighter lower bound to \( \mathcal{L}_{\text{ELBO}}(\mathbf{x}_m; \mathbf{z}_m) \) than \( \mathcal{L}_{\text{ELBO}}(\mathbf{x}_m; \mathbf{z}_m) \), defined as
\[
\mathcal{L}_{\text{IWAE}}(\mathbf{x}_m; \mathbf{z}_m) = \mathbb{E}_{q_{\phi}(\mathbf{z}_m | \mathbf{x}_m)} \left[ \log \sum_k \frac{1}{M} p_{\Theta}(\mathbf{x}_m | \mathbf{z}_m) \right]
\]
(2)
With the MoE joint posterior, \( \mathcal{L}_{\text{IWAE}}(\mathbf{x}_m; \mathbf{z}_m) \) can be written as
\[
\mathcal{L}_{\text{IWAE}}(\mathbf{x}_m; \mathbf{z}_m) = \sum_k \mathbb{E}_{q_{\phi}(\mathbf{z}_m | \mathbf{x}_m)} \left[ \log \sum_k \frac{1}{M} p_{\Theta}(\mathbf{x}_m | \mathbf{z}_m) \right]
\]
(3)
Why MoE? Some alternatives —
► Explicitly parametrisation \( q_{\phi}(\mathbf{z} | \mathbf{x}, \mathbf{m}) \) as an encoder: requires all modalities to be presented at all time, violating requirement (c);
► Product of experts (PoE), i.e. \( q_{\phi}(\mathbf{z} | \mathbf{x}, \mathbf{m}) = \prod_m q_{\phi_m}(\mathbf{z} | \mathbf{x}, \mathbf{m}) \) seen in [3]: prone to bias towards experts from modality with higher precision, undesirable for multi-modal reconstruction.

Latent factorisation: visualisation
The UMAP visualisation of the posterior and prior distribution demonstrates a few possibilities of latent factorisation, using different objectives including:
► Our objective (see Eq. (5))
► A modality-biased version of Eq. (5)
► Eq. (5) with MMD regularisation

Experiments: MNIST-SVHN

Table 3: Correlation of image (I)-Sentence (S) pair for generation.

Figure 2: Qualitative evaluation of MMVAE (ours) and MVAE [3].

Figure 3: Latent dimensions that affect SVHN, MNIST, and both.

Figure 4: Qualitative evaluation of MMVAE (ours) and MVAE [3].

Figure 5: GradCAM (4) visualisation of vision → language: The heatmaps visualise attention in the vision domain activated by the selected keywords in generated sentences (highlighted in yellow).

References

Experiments: CUB Image-Caption

Table 1: Latent digit classification accuracy (%).

Table 2: Probability of digit matching (%) for joint and cross generation.