

# Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models

Kseniia Petrushina

MIPT, 2023

November 21, 2023

1 Motivation & Background

2 Theory

3 Empirical results

# Motivation

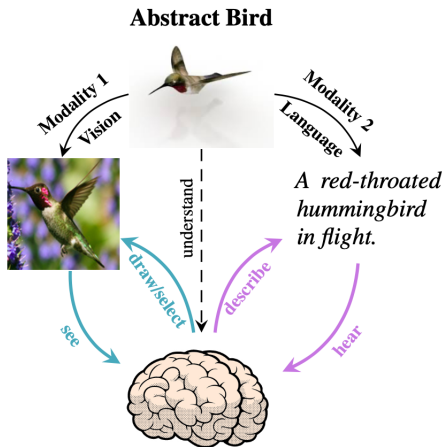


Figure: A schematic for multi-modal perception.

# Motivation

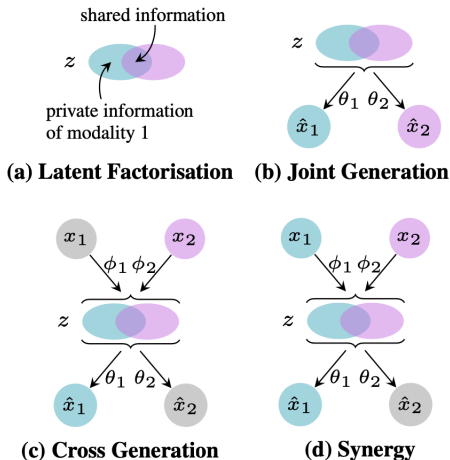


Figure: The four criteria for multi-modal generation models.

# Background

## VAE

**Goal:**

$$p_{\Theta}(\mathbf{z}, \mathbf{x}_{1,\dots,M}) = p(\mathbf{z}) \prod_{m=1}^M p_{\theta_m}(\mathbf{x}_m | \mathbf{z})$$

**Training objective:**

$$p_{\Theta}(\mathbf{x}_{1:M}) \rightarrow \max_{\Theta}$$

True posterior  $p_{\Theta}(\mathbf{z} | \mathbf{x}_{1:M}) \rightarrow$  variational posterior  $q_{\Phi}(\mathbf{z} | \mathbf{x}_{1:M})$

## ELBO

$$\mathcal{L}_{\text{ELBO}}(\mathbf{x}_{1:M}) = \mathbb{E}_{\mathbf{z} \sim q_{\Phi}(\mathbf{z} | \mathbf{x}_{1:M})} \left[ \log \frac{p_{\Theta}(\mathbf{z}, \mathbf{x}_{1:M})}{q_{\Phi}(\mathbf{z} | \mathbf{x}_{1:M})} \right]$$

$$\mathcal{L}_{\text{IWAE}}(\mathbf{x}_{1:M}) = \mathbb{E}_{\mathbf{z}^{1:K} \sim q_{\Phi}(\mathbf{z} | \mathbf{x}_{1:M})} \left[ \log \sum_{k=1}^K \frac{1}{K} \frac{p_{\Theta}(\mathbf{z}^k, \mathbf{x}_{1:M})}{q_{\Phi}(\mathbf{z}^k | \mathbf{x}_{1:M})} \right]$$

# Joint variational posterior

## Product of experts

$$q_{\Phi}(\mathbf{z}|\mathbf{x}_{1:M}) = \prod_{m=1}^M q_{\phi_m}(\mathbf{z}|\mathbf{x}_m)$$

- Low total density if one of the the factors has low density – each expert has veto power
- Overconfidence in predictions has detrimental consequences for the model

## Mixture of experts

$$q_{\Phi}(\mathbf{z}|\mathbf{x}_{1:M}) = \frac{1}{M} \sum_{m=1}^M q_{\phi_m}(\mathbf{z}|\mathbf{x}_m)$$

# MoE-multimodal VAE

## Objective

$$\mathcal{L}_{\text{IWAE}}^{\text{MoE}}(\mathbf{x}_{1:M}) = \frac{1}{M} \sum_{m=1}^M \mathbb{E}_{\mathbf{z}_m^{1:K} \sim q_{\phi_m}(\mathbf{z}|\mathbf{x}_m)} \left[ \log \frac{1}{K} \sum_{k=1}^K \frac{p_{\Theta}(\mathbf{z}_m^k, \mathbf{x}_{1:M})}{q_{\Phi}(\mathbf{z}_m^k | \mathbf{x}_{1:M})} \right]$$

## Theorem

$$\mathcal{L}_{\text{ELBO}}(\mathbf{x}_{1:M}) \leq \mathcal{L}_{\text{IWAE}}^{\text{MoE}}(\mathbf{x}_{1:M})$$

## Proof

$$\begin{aligned} \mathcal{L}_{\text{ELBO}}(\mathbf{x}_{1:M}) &= \mathbb{E}_{\mathbf{z} \sim q_{\Phi}(\mathbf{z}|\mathbf{x}_{1:M})} \left[ \log \frac{p_{\Theta}(\mathbf{z}, \mathbf{x}_{1:M})}{q_{\Phi}(\mathbf{z}|\mathbf{x}_{1:M})} \right] = \\ &= \frac{1}{M} \sum_{m=1}^M \mathbb{E}_{\mathbf{z}_m \sim q_{\phi_m}(\mathbf{z}|\mathbf{x}_m)} \left[ \log \frac{p_{\Theta}(\mathbf{z}_m, \mathbf{x}_{1:M})}{q_{\Phi}(\mathbf{z}_m | \mathbf{x}_{1:M})} \right] \leq \mathcal{L}_{\text{IWAE}}^{\text{MoE}}(\mathbf{x}_{1:M}) \end{aligned}$$

# MNIST-SVHN

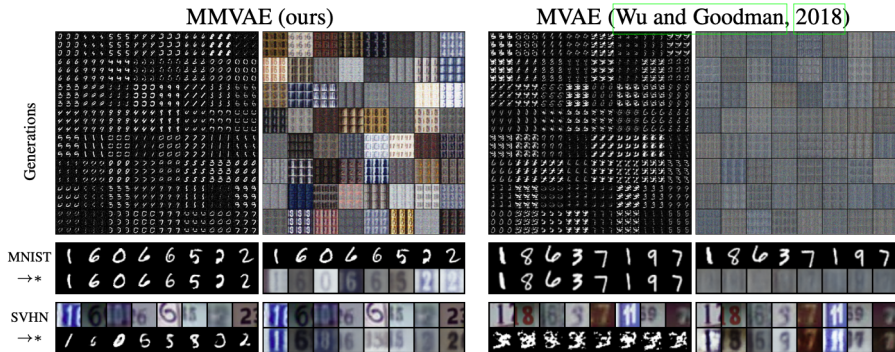


Figure: Qualitative results on MNIST-SVHN dataset pair.



# MNIST-SVHN

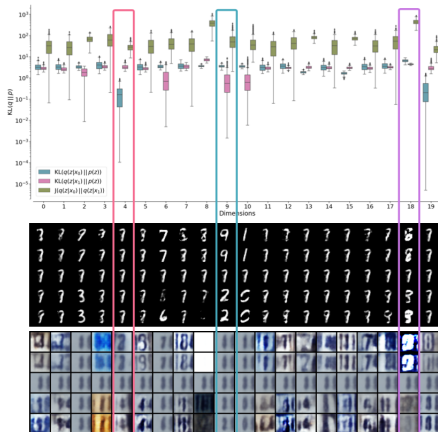


Figure: Per-dimension latent traversals for a pair of datapoints indicating dimensions that affect only **SVHN**, only **MNIST**, and both **MNIST & SVHN**.

# MNIST-SVHN

	$\log p(\mathbf{x}_m   \mathbf{x}_m, \mathbf{x}_n)$	$\log p(\mathbf{x}_m   \mathbf{x}_m)$	$\log p(\mathbf{x}_m   \mathbf{x}_n)$
m=MNIST, n=SVHN	868.76	868.37	628.31
m=SVHN, n=MNIST	3441.01	3441.01	2337.56

**Table:** Log-likelihoods for different arrangements of MNIST and SVHN.

joint marginal likelihood  $\geq$  single marginal likelihood

# Caltech-UCSD Birds

## vision → vision



## Generation from prior samples:



this bird is black with a a  
and and a black beak.

this bird is white with black and has  
a long, pointy beak.

this bird has wings that are brown  
and a thick bill.

## language → language

this small, white bellied bird  
has a brown head and a  
read tipped beak.

this small bird bird with a  
brown head and a red bill.

this bird has a grey  
back, a white and black  
spotted belly and breast  
and a yellow eyebrow.

this bird has a white  
belly, a brown back, and  
a brown brown eyebrow.

a black bird with a curved  
neck, and a long silver  
beak with blue eyes.

a large bird with black  
neck, and a long long  
beak and and and bill.

a small sized bird that has  
a yellow facial marking with  
a pointed bill.

a small bird with a  
yellow belly with a  
pointed bill.

## language → vision



this small, white bellied bird  
has a brown head and a  
read tipped beak.

this bird has a grey  
back, a white and black  
spotted belly and breast  
and a yellow eyebrow.

a black bird with a curved  
neck, and a long silver  
beak with blue eyes.

a small sized bird that has  
a yellow facial marking  
with a pointed bill.

## vision → language



this bird has wings that are  
black and has a white bill.

this bird is brown with white  
and a very short beak.

this bird is black black  
with with a and and a  
red beak.

this bird is yellow and  
yellow, white and and  
and a and and beak.

Figure: Qualitative results on CUB dataset.

- 1 **Main article** Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models.