Incomplete Multimodal Learning

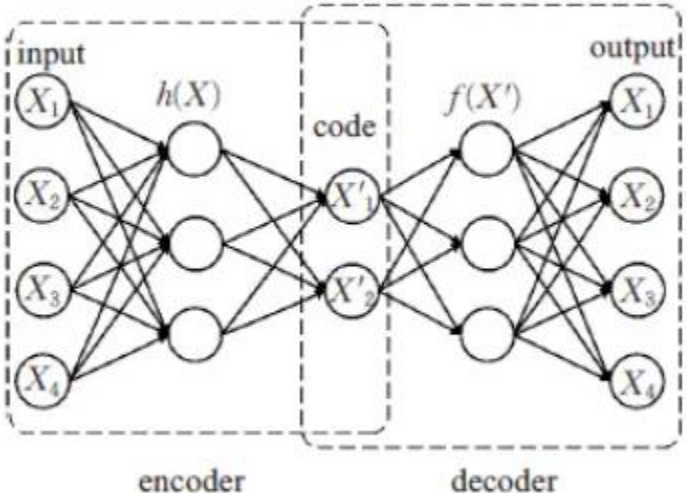
LI Xiaolou

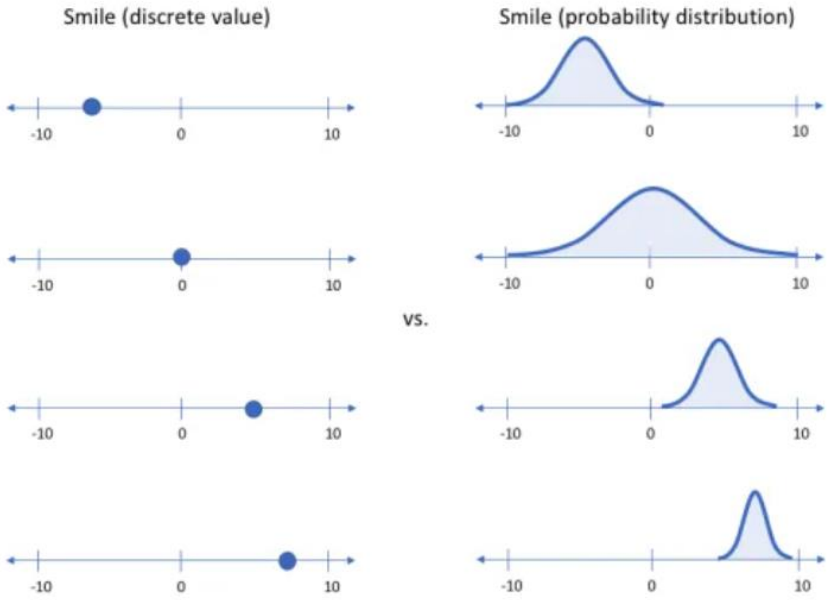
2023/4/28

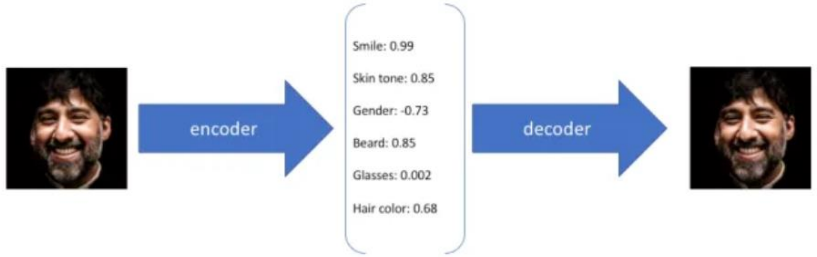
Paper

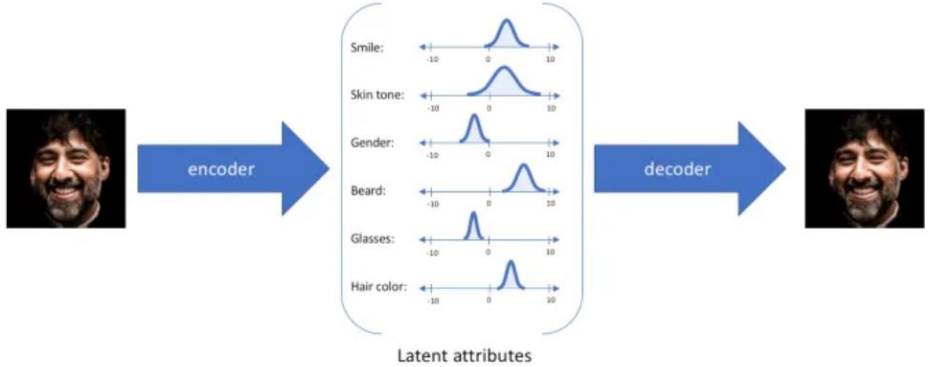
1. (VAE) Doersch, C., 2016. Tutorial on variational autoencoders. arXiv preprint arXiv:1606.05908.
2. (MVAE) Wu, M. and Goodman, N., 2018. Multimodal generative models for scalable weakly-supervised learning. Advances in neural information processing systems, 31.
3. (IWAE) Burda, Y., Grosse, R. and Salakhutdinov, R., 2015. Importance weighted autoencoders. arXiv preprint arXiv:1509.00519.
4. (MMVAE) Shi, Y., Paige, B. and Torr, P., 2019. Variational mixture-of-experts autoencoders for multi-modal deep generative models. Advances in Neural Information Processing Systems, 32.

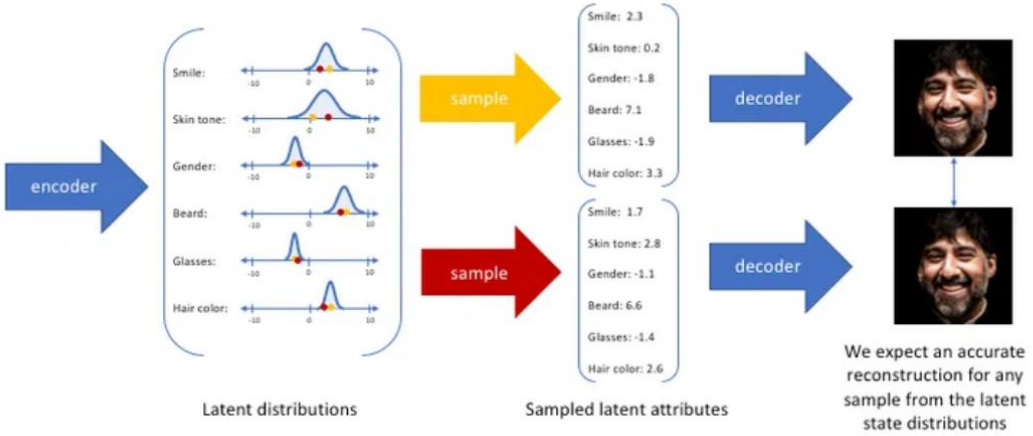
VAE


input
$h(X)$
code
$f(X')$
output
X_1
X_2
X_3
X_4
X'_1
X'_2
encoder
decoder

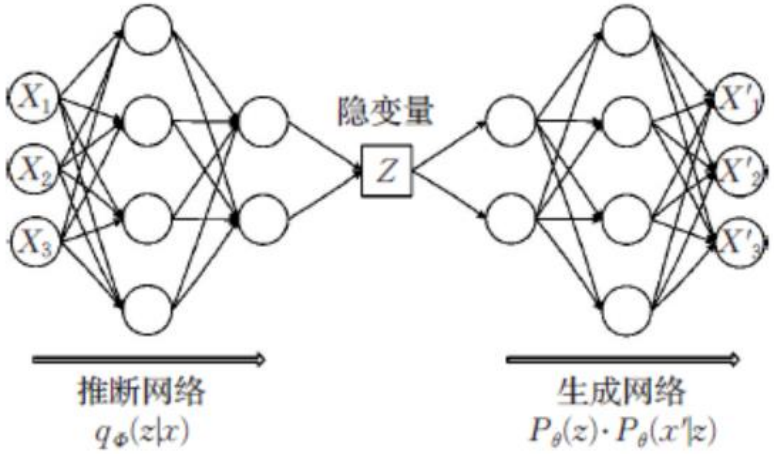

Smile (discrete value)
Smile (probability distribution)
-10
0
10
vs.


encoder
Smile: 0.99
Skin tone: 0.85
Gender: -0.73
Beard: 0.85
Glasses: 0.002
Hair color: 0.68
decoder
Latent attributes


encoder
Smile:
Skin tone:
Gender:
Beard:
Glasses:
Hair color:
decoder
Latent attributes


encoder
Smile:
Skin tone:
Gender:
Beard:
Glasses:
Hair color:
sample
sample
Smile: 2.3
Skin tone: 0.2
Gender: -1.8
Beard: 7.1
Glasses: -1.9
Hair color: 3.3
Smile: 1.7
Skin tone: 2.8
Gender: -1.1
Beard: 6.6
Glasses: -1.4
Hair color: 2.6
decoder
decoder
Latent distributions
Sampled latent attributes
We expect an accurate reconstruction for any sample from the latent state distributions

VAE



Original Data: $X = \{x_i\}_{i=1}^N$

Generated Data: $X' = \{x_i'\}_{i=1}^N$

Inference Net: $q_\phi(z|x)$

Generation Net: $P_\theta(z)P_\theta(x'|z)$

Why q(z|x)?

$$P(x1, x2, x3, \dots, xn) = P(x1)P(x2|x1)P(x3|x2, x1) \dots P(xn|xn-1, xn-2, xn-3, \dots, x1)$$

$\downarrow$ Mean Field Theory

$$Q(x1, x2, x3, \dots, xn) = Q(x1)Q(x2)Q(x3) \dots Q(xn)$$

VAE

$q(z|x) \rightarrow p(z|x)?$

$minKL(q(z|x)||p(z|x))$

L^v: evidence lower bound

$$
\begin{aligned}
L &= \log(p(x)) \\
&= \sum_z q(z|x) \log(p(x)) \\
&= \sum_z q(z|x) \log\left(\frac{p(z,x)}{p(z|x)}\right) \\
&= \sum_z q(z|x) \log\left(\frac{p(z,x)}{q(z|x)}\frac{q(z|x)}{p(z|x)}\right) \\
&= \sum_z q(z|x) \log\left(\frac{p(z,x)}{q(z|x)}\right) + \sum_z q(z|x) \log\left(\frac{q(z|x)}{p(z|x)}\right) \\
&= L^v + D_{KL}(q(z|x)||p(z|x))
\end{aligned}
$$

$$
\begin{aligned}
L^v &= \sum_z q(z|x) \log\left(\frac{p(z,x)}{q(z|x)}\right) \\
&= \sum_z q(z|x) \log\left(\frac{p(x|z)p(z)}{q(z|x)}\right) \\
&= \sum_z q(z|x) \log\left(\frac{p(z)}{q(z|x)}\right) + \sum_z q(z|x) \log(p(x|z))
\end{aligned}
$$

$$L^V = -D_{KL}(q(z|x)||p(z)) + \mathrm{E}_{q(z|x)}(log(p(x|z)))$$

$$L_1 = \frac{1}{2}\sum_{j=1}^{J}[1+\log((\sigma_j)^2)-(u_j)^2-(\sigma_j)^2]$$

$$L_2 = \mathrm{E}_{q(z|x)}(log(p(x|z))) \approx \frac{1}{L}\sum_{l=1}^{L} log\, p(x|z^{(l)})$$

$z \sim N(\mu, \sigma^2)$

MVAE

Motivation:

The incompleteness of the mode is very common, and the work done by the predecessors either does not learn the joint distribution or requires additional calculations. And this paper uses a multimodal VAE

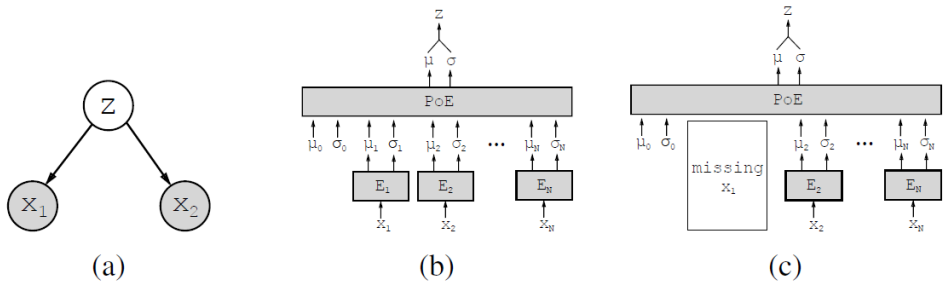


Figure 1: (a) Graphical model of the MVAE. Gray circles represent observed variables. (b) MVAE architecture with N modalities. E_i represents the i-th inference network; μ_i and σ_i represent the i-th variational parameters; μ_0 and σ_0 represent the prior parameters. The product-of-experts (PoE) combines all variational parameters in a principled and efficient manner. (c) If a modality is missing during training, we drop the respective inference network. Thus, the parameters of $E_1, ..., E_N$ are shared across different combinations of missing inputs.

Method:

1. MVAE
2. PoE (Product-of-expert inference network)
3. Sub-sampled training

$$\mathrm{ELBO}(x) \triangleq \mathbb{E}_{q_\phi(z|x)}[\lambda \log p_\theta(x|z)] - \beta\, \mathrm{KL}[q_\phi(z|x), p(z)]$$

conditional independence

$$p(x1, x2, \ldots, xn, z) = p(z)p(x1|z)p(x2|z) \ldots p(xn|z)$$

$$\mathrm{ELBO}(X) \triangleq \mathbb{E}_{q_\phi(z|X)}[\sum_{x_i \in X} \lambda_i \log p_\theta(x_i|z)] - \beta\, \mathrm{KL}[q_\phi(z|X), p(z)].$$

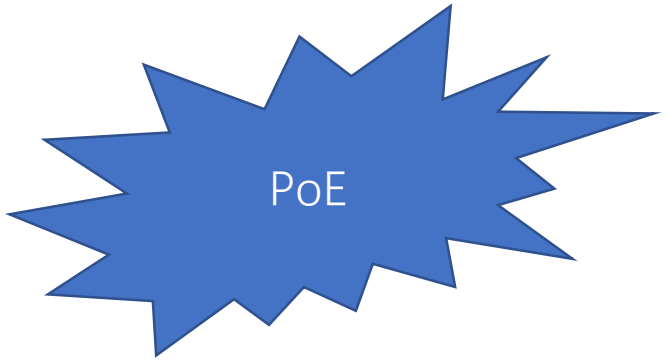
MVAE

Approximating the joint posterior

$$
\begin{aligned}
p(z|x_1,...,x_N) &= \frac{p(x_1,...,x_N|z)p(z)}{p(x_1,...,x_N)} = \frac{p(z)}{p(x_1,...,x_N)}\prod_{i=1}^{N} p(x_i|z) \\
&= \frac{p(z)}{p(x_1,...,x_N)}\prod_{i=1}^{N}\frac{p(z|x_i)p(x_i)}{p(z)} = \frac{\prod_{i=1}^{N} p(z|x_i)}{\prod_{i=1}^{N-1} p(z)}\cdot\frac{\prod_{i=1}^{N} p(x_i)}{p(x_1,...,x_N)} \qquad (3) \\
&\propto \frac{\prod_{i=1}^{N} p(z|x_i)}{\prod_{i=1}^{N-1} p(z)}
\end{aligned}
$$

Alternatively, if we approximate $p(z|x_i)$ with $q(z|x_i) \equiv \tilde{q}(z|x_i)p(z)$, where $\tilde{q}(z|x_i)$ is the underlying inference network, we can avoid the quotient term:

$$
p(z|x_1,...,x_N) \propto \frac{\prod_{i=1}^{N} p(z|x_i)}{\prod_{i=1}^{N-1} p(z)} \approx \frac{\prod_{i=1}^{N}[\tilde{q}(z|x_i)p(z)]}{\prod_{i=1}^{N-1} p(z)} = p(z)\prod_{i=1}^{N}\tilde{q}(z|x_i). \qquad (4)
$$



Why Sub-sampled Training?

1. Can't handle modality lost
2. Unable to get inter-modal info

- the ELBO using the product of all N Gaussians
- all ELBO terms using a single modality
- k ELBO terms using k randomly chosen subsets

$$
\mathrm{ELBO}(x_1,...,x_N) + \sum_{i=1}^{N}\mathrm{ELBO}(x_i) + \sum_{j=1}^{k}\mathrm{ELBO}(X_j)
$$

MVAE

Model	BinaryMNIST	MNIST	FashionMNIST	MultiMNIST	CelebA
Estimated log $p(x_1)$					
VAE	-86.313	-91.126	-232.758	-152.835	-6237.120
BiVCCA	-87.354	-92.089	-233.634	-202.490	-7263.536
JMVAE	-86.305	-90.697	-232.630	-152.787	-6237.967
MVAE-Q	-91.665	-96.028	-236.081	-166.580	-6290.085
MVAE	**-86.026**	**-90.619**	**-232.535**	**-152.761**	-6236.923
MVAE19	–	–	–	–	**-6236.109**
Estimated log $p(x_1, x_2)$					
JMVAE	-86.371	**-90.769**	**-232.948**	**-153.101**	-6242.187
MVAE-Q	-92.259	-96.641	-236.827	-173.615	-6294.861
MVAE	**-86.255**	-90.859	-233.007	-153.469	-6242.034
MVAE19	–	–	–	–	**-6239.944**
Estimated log $p(x_1 \| x_2)$					
CVAE	**-83.448**	**-87.773**	**-229.667**	–	-6228.771
JMVAE	-83.985	-88.696	-230.396	**-145.977**	**-6231.468**
MVAE-Q	-90.024	-94.347	-234.514	-163.302	-6311.487
MVAE	-83.970	-88.569	-230.695	-147.027	-6234.955
MVAE19	–	–	–	–	-6233.340

Table 2: Estimates (using $q(z|x_1)$) for marginal probabilities on the average test example. MVAE and JMVAE are roughly equivalent in data log-likelihood but as Table 1 shows, MVAE uses far fewer parameters. The CVAE is often better at capturing $p(x_1|x_2)$ but does not learn a joint distribution.

IWAW

Importance Weighted Autoencoders

LOSS of VAE: $E_{z\sim q(z|x)}[\log \frac{p(x,z)}{q(z|x)}]$

LOSS of IWAE: $E_{z_1,z_2,\ldots,z_k\sim q(z|x)}[\log \frac{1}{k}\sum_{i=1}^{k} \frac{p(x,z_i)}{q(z_i|x)}]$

Why weighted?

Gradient:

VAE: $\frac{1}{k}\sum_{i=1}^{k} \nabla_\theta \log w_i$

w: $\frac{p(x,z|\theta)}{q(z|x,\theta)}$

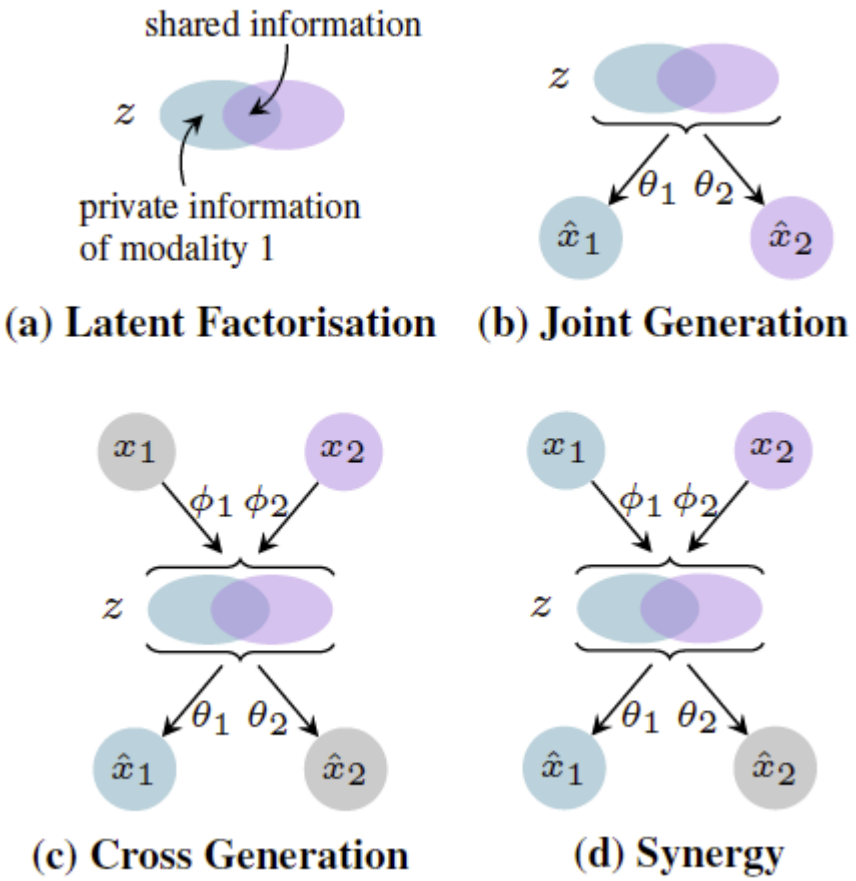
IWAE: $\sum_{i=1}^{k} \tilde{w}_i \nabla_\theta \log w_i$

$\frac{w_i}{\sum_{j=1}^{k} w_j} = \tilde{w}_i$

MMVAE

4 Standard:

- Latent Factorisation
- Coherent Joint Generation
- Coherent Cross Generation
- Synergy



(a) Latent Factorisation (b) Joint Generation

(c) Cross Generation (d) Synergy

MMVAE

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

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

Why MoE?

MoE(mixture of experts)

$$q_\Phi(\boldsymbol{z} \mid \boldsymbol{x}_{1:M}) = \sum_m \alpha_m \cdot q_{\phi_m}(\boldsymbol{z} \mid \boldsymbol{x}_m)$$

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

MMVAE

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

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

Multimodal IWAE!!!!

MMVAE

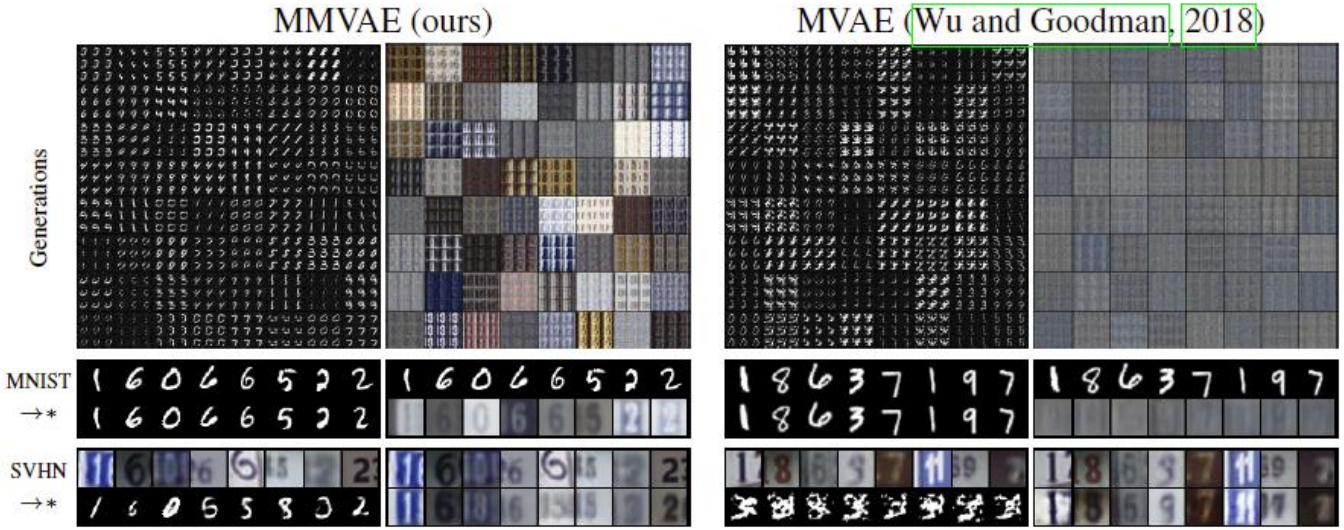


Table 2: Probability of digit matching (%) for joint and cross generatio

	Joint	Cross (M→S)	Cross (S→M)
MMVAE	42.1	86.4	69.1
MVAE	12.7	9.5	9.3

Table 3: Evaluating the different log likelihoods for different arrangements of MNIST and SVHN.

		$\log p(\boldsymbol{x}_m, \boldsymbol{x}_n)$	$\log p(\boldsymbol{x}_m \mid \boldsymbol{x}_m, \boldsymbol{x}_n)$	$\log p(\boldsymbol{x}_m \mid \boldsymbol{x}_m)$	$\log p(\boldsymbol{x}_m \mid \boldsymbol{x}_n)$
m = MNIST, n = SVHN	**MMVAE**	6261.40	868.76	868.37	628.31
	MVAE	2961.80	−176.68	−107.46	−778.20
m = SVHN, n = MNIST	**MMVAE**	6261.40	3441.01	3441.01	2337.56
	MVAE	2961.80	3395.12	3536.86	−12747.50

Table 1: Digit classification accuracy (%) of latent variables in different models.

	MMVAE	MVAE (single)	MVAE (both)	single-VAE
MNIST	91.3	95.7	94.9	85.3
SVHN	68.0	9.1	90.1	20.7