KDD 2022 Research Track

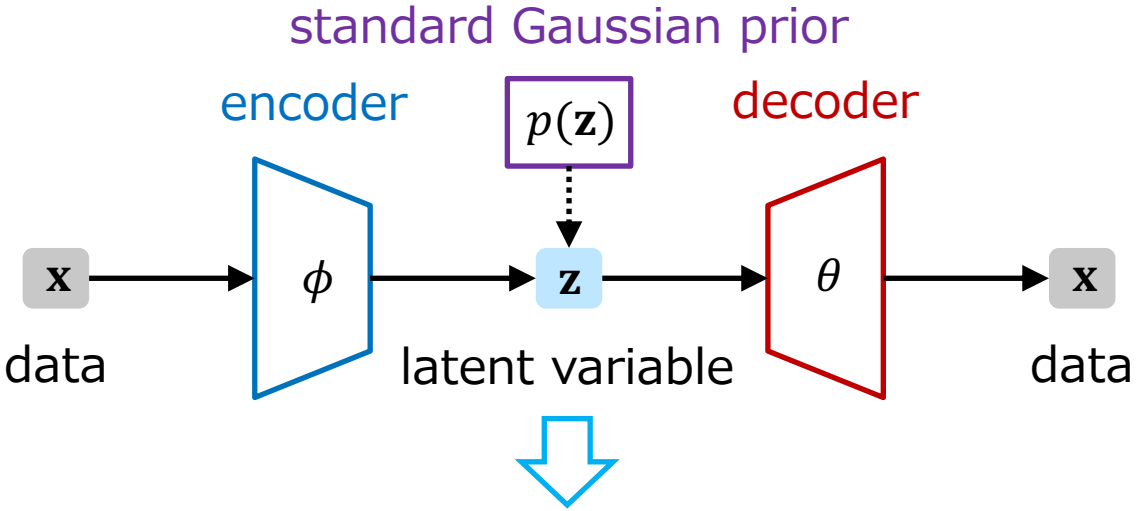
Learning Optimal Priors for Task-Invariant Representations in Variational Autoencoders

Hiroshi Takahashi[1], Tomoharu Iwata[1], Atsutoshi Kumagai[1], Sekitoshi Kanai[1], Masanori Yamada[1], Yuuki Yamanaka[1], Hisashi Kashima[2]

[1]NTT, [2]Kyoto University

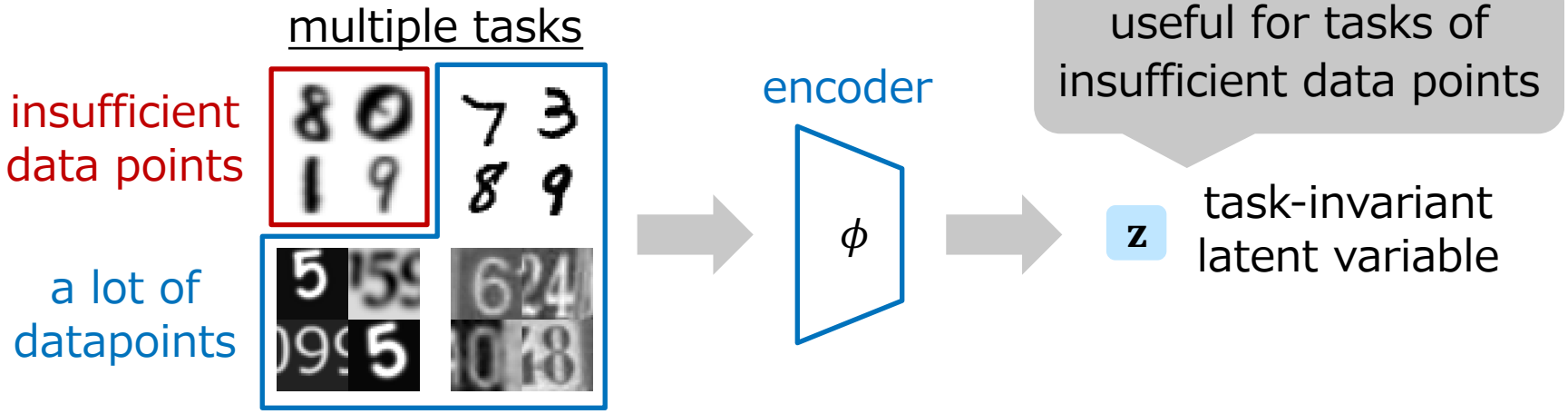
[Introduction] Variational Autoencoder

- The variational autoencoder (VAE) is a powerful latent variable model for unsupervised representation learning.



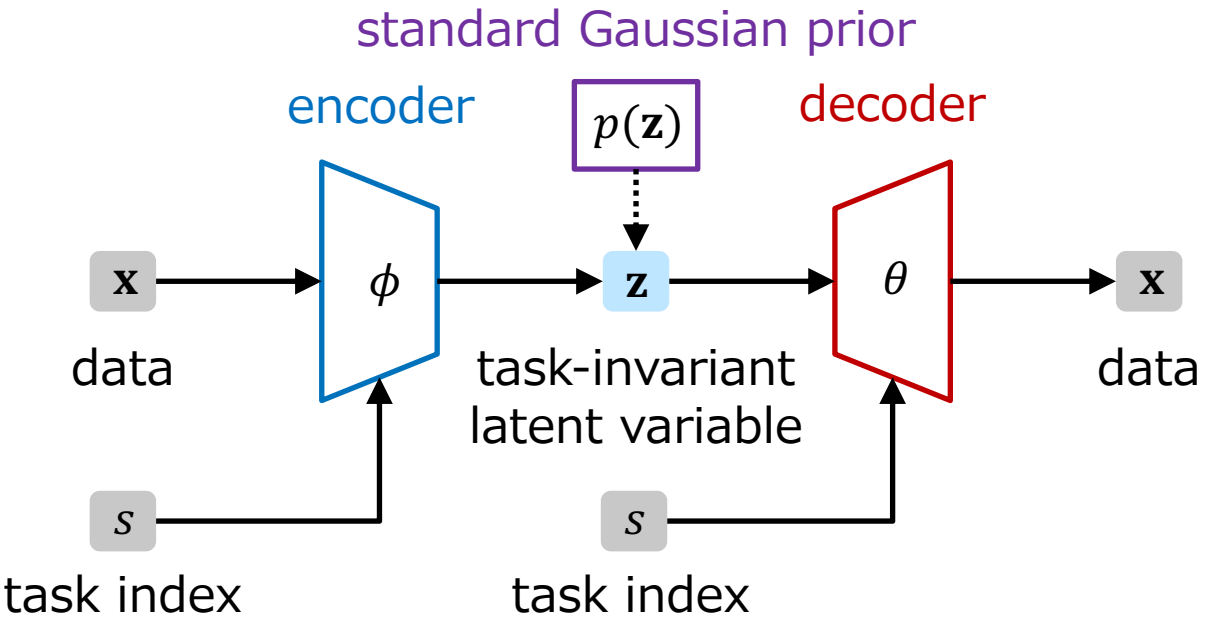
[Introduction] Multi-Task Learning

- However, the VAE cannot perform well with insufficient data points since it depends on neural networks.
- To solve this, we focus on obtaining task-invariant latent variable from multiple tasks.



[Introduction] Conditional VAE

- For multiple tasks, the conditional VAE (CVAE) is widely used, which tries to obtain task-invariant latent variable.



[Introduction] Problem and Contribution

NTT

- Although the CVAE can reduce the dependency of $\mathbf{z}$ on s to some extent, this dependency remains in many cases.
- The contribution of this study is as follows:

1. We investigate the cause of the task-dependency in the CVAE and reveal that the **simple prior** is one of the causes.
2. We introduce the **optimal prior** to reduce the task-dependency.
3. We theoretically and experimentally show that our learned representation works well on multiple tasks.

[Preliminaries] Reviewing CVAE

- The CVAE models a conditional probability of $\mathbf{x}$ given s as:

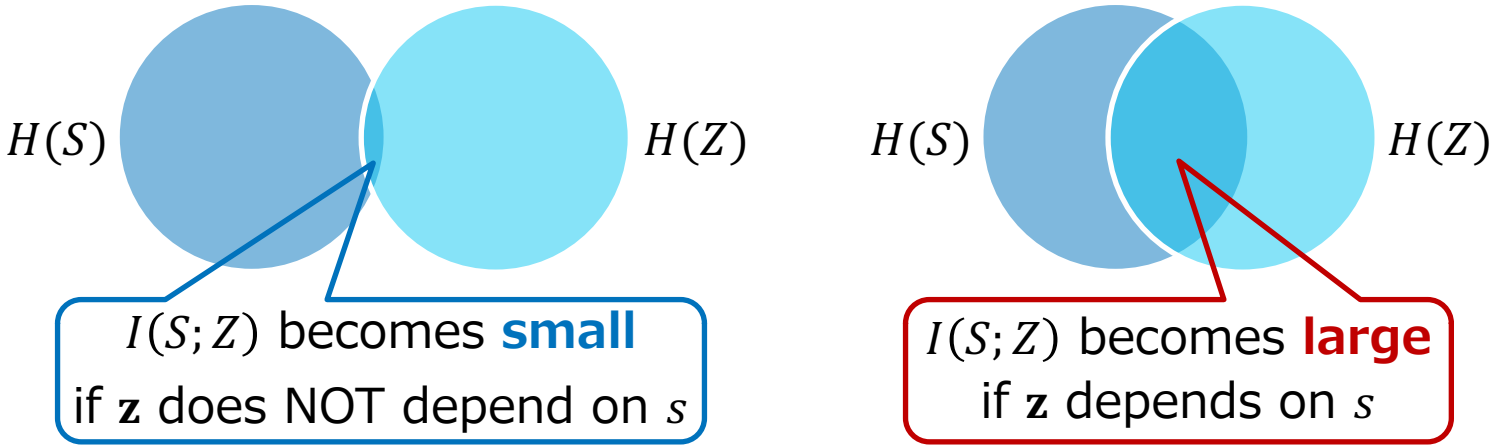
$$p_\theta(\mathbf{x}|s) = \int \underbrace{p_\theta(\mathbf{x}|\mathbf{z},s)}_{\text{decoder}} \underbrace{p(\mathbf{z})}_{\text{prior}} \mathrm{d}\mathbf{z} = \mathbb{E}_{\underbrace{q_\phi(\mathbf{z}|\mathbf{x},s)}_{\text{encoder}}} \left[\frac{p_\theta(\mathbf{x}|\mathbf{z},s)p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x},s)} \right]$$

- The CVAE is trained by maximizing the ELBO that is the lower bound of the log-likelihoods as follows:

$$\mathcal{F}_{\mathrm{CVAE}}(\theta,\phi) = \mathbb{E}_{\underbrace{p_D(\mathbf{x},s)}_{\text{data distribution}} q_\phi(\mathbf{z}|\mathbf{x},s)} \left[\ln p_\theta(\mathbf{x}|\mathbf{z},s) \right] - \underbrace{\mathbb{E}_{p_D(\mathbf{x},s)} \left[D_{KL}(q_\phi(\mathbf{z}|\mathbf{x},s) \| p(\mathbf{z})) \right]}_{= \mathcal{R}(\phi)}$$

[Preliminaries] Mutual Information

- To investigate the cause of dependency of $\mathbf{z}$ on s, we introduce the mutual information $I(S;Z)$, which measures the mutual dependence between two random variables.



[Proposed] Theorem 1

- The CVAE tries to minimize the mutual information $I(S;Z)$ by minimizing its upper bound $\mathcal{R}(\phi)$:

$$\mathcal{R}(\phi) \equiv \mathbb{E}_{p_D(\mathbf{x},s)}\left[D_{KL}(q_\phi(\mathbf{z}|\mathbf{x},s)\|p(\mathbf{z}))\right]$$

$$= I(S;Z) + D_{KL}(q_\phi(\mathbf{z})\|p(\mathbf{z})) + \sum_{k=1}^{K} \pi_k I(X^{(k)};Z^{(k)})$$

mutual information between $\mathbf{x}$ and $\mathbf{z}$ when $s=k$

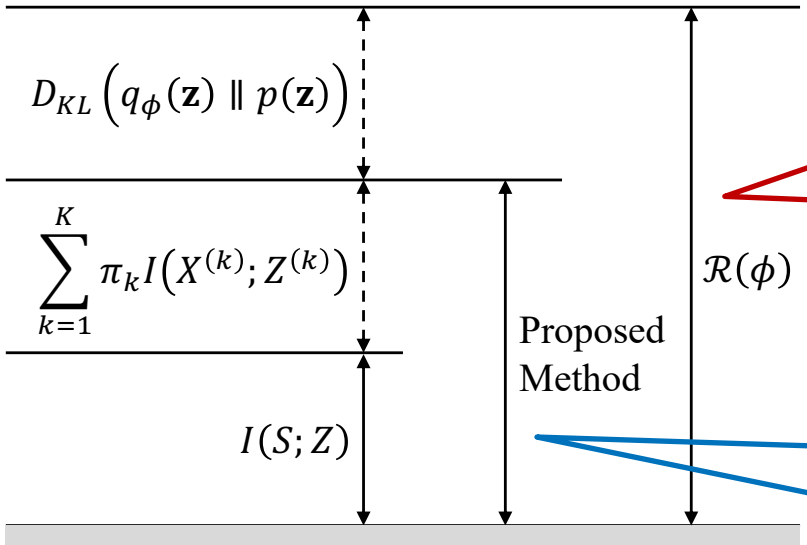
$q_\phi(\mathbf{z}) = \int q_\phi(\mathbf{z}|\mathbf{x},s)p_D(\mathbf{x},s)\mathrm{d}\mathbf{x}$

$\pi_k = p(s=k)$

- However, $\mathcal{R}(\phi)$ is NOT a tight upper bound of $I(S;Z)$ since $D_{KL}(q_\phi(\mathbf{z})||p(\mathbf{z}))$ usually gives a large value.

[Proposed] Effects of Priors

- That is, the simple prior $p(\mathbf{z})$ is **one causes of the task-dependency**, and $q_\phi(\mathbf{z})$ is the **optimal prior** to reduce it.



$\mathcal{R}(\phi)$ is NOT a tight upper bound of $I(S;Z)$ since $D_{KL}(q_\phi(\mathbf{z})||p(\mathbf{z}))$ usually gives a large value.

When $p(\mathbf{z}) = q_\phi(\mathbf{z})$, $\mathcal{R}(\phi)$ becomes the tightest upper bound of $I(S;Z)$.

[Proposed] Theorem 2

- The ELBO with this optimal prior $\mathcal{F}_{\mathrm{Proposd}}(\theta, \phi)$ is always larger than or equal to original ELBO $\mathcal{F}_{\mathrm{CVAE}}(\theta, \phi)$:

$$\mathcal{F}_{\mathrm{Proposed}}(\theta, \phi) = \mathcal{F}_{\mathrm{CVAE}}(\theta, \phi) + D_{KL}(q_\phi(\mathbf{z}) \| p(\mathbf{z})) \geq \mathcal{F}_{\mathrm{CVAE}}(\theta, \phi)$$

- That is, $\mathcal{F}_{\mathrm{Proposd}}(\theta, \phi)$ is also a **better lower bound of the log-likelihood** than $\mathcal{F}_{\mathrm{CVAE}}(\theta, \phi)$.
- This contributes to obtaining better representation for the improved performance on the target tasks.

[Proposed] Optimizing $\mathcal{F}_{\mathrm{Proposd}}(\theta, \phi)$

- We optimize $\mathcal{F}_{\mathrm{Proposd}}(\theta, \phi) = \mathcal{F}_{\mathrm{CVAE}}(\theta, \phi) + D_{KL}(q_{\phi}(\mathbf{z})||p(\mathbf{z}))$ by approximating the KL divergence $D_{KL}(q_{\phi}(\mathbf{z})||p(\mathbf{z}))$:

$$D_{KL}(q_{\phi}(\mathbf{z})\|p(\mathbf{z})) = \int q_{\phi}(\mathbf{z}) \ln \frac{q_{\phi}(\mathbf{z})}{p(\mathbf{z})} \mathrm{d}\mathbf{z}$$

- We approximate $q_{\phi}(\mathbf{z})/p(\mathbf{z})$ by **density ratio trick**, which can estimate the density ratio between two distributions using samples from both distribution (See Section 3.3).

[Proposed] Theoretical Contributions

- Our theoretical contributions are summarized as follows:

Theorem 1 shows:

- The **simple prior** is one of the causes of the task-dependency.
- $q_\phi(\mathbf{z})$ is the **optimal prior** to reduce the task-dependency.

Theorem 2 shows:

- $\mathcal{F}_{\mathrm{Proposd}}(\theta, \phi)$ gives a **better lower bound of the log-likelihood**, which enables us to obtain better representation than the CVAE.

- We next evaluate our representation on various datasets.

[Experiments] Datasets

- We used two handwritten digits (USPS and MNIST), two house number digits (SynthDigits and SVHN), and three face datasets (Frey, Olivetti, and UMist).

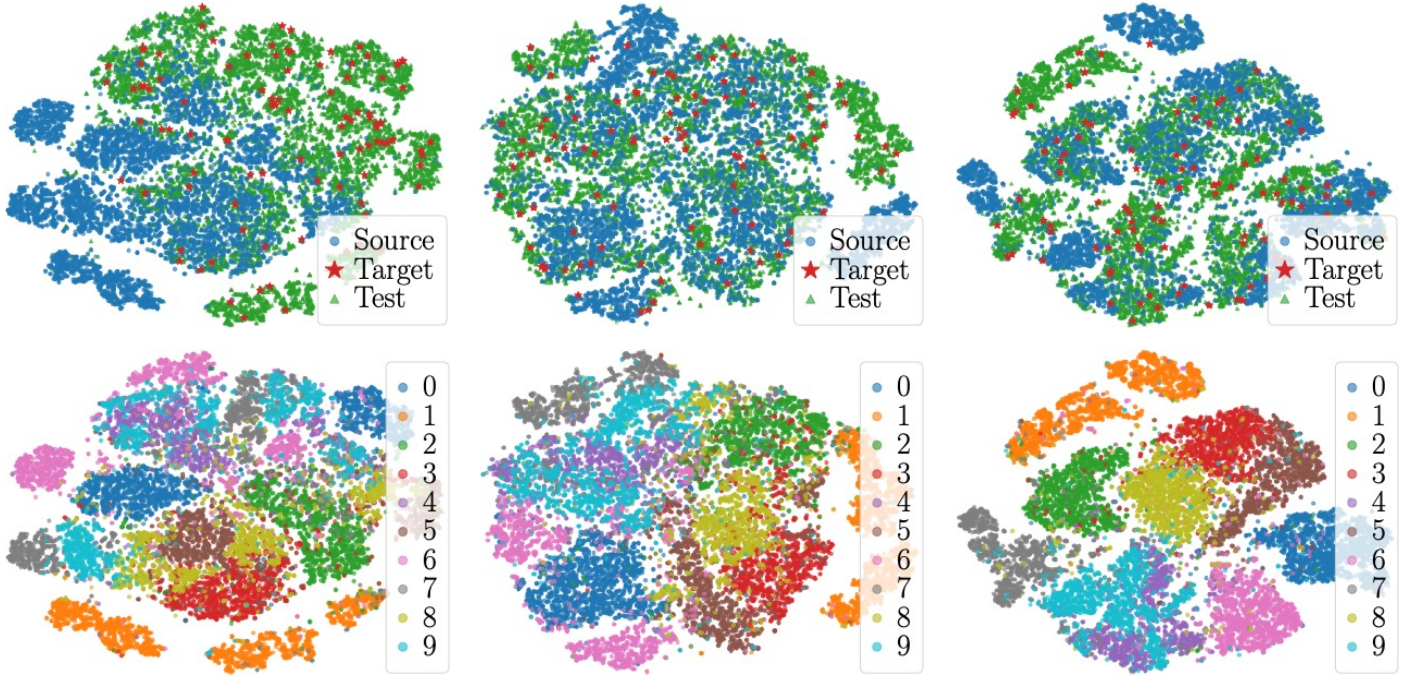
	Dimension	Train size	Valid size	Test size
USPS	784	6,438	1,000	1,860
MNIST	784	10,000	10,000	10,000
SynthDigits	1,024	10,000	10,000	9,553
SVHN	1,024	10,000	10,000	26,032
Frey	560	1,565	200	200
Olivetti	560	150	100	150
UMist	560	300	75	200

[Experiments] Settings

- On digits datasets, we conducted two-task experiments, which estimate the performance on the target tasks:
  - The source task has a lot of training data points.
  - The target task has only 100 training data points.
  - Pairs are (USPS→MNIST), (MNIST→USPS), (SynthDigits→SVHN), and (SVHN→SynthDigits).
- On face datasets, we conducted three-task experiment, which simultaneously evaluates the performance on each task using a single estimator.

[Results] Visualizing Representations

Visualization of latent variables on USPS→MNIST



[Results] Density Estimation Performance

NTT

	VAE	CVAE	Proposed
USPS→MNIST	-163.25 ± 2.15	-152.32 ± 1.64	$\mathbf{-149.08 \pm 0.86}$
MNIST→USPS	-235.23 ± 1.54	$\mathbf{-211.18 \pm 0.55}$	$\mathbf{-212.11 \pm 1.48}$
Synth→SVHN	1146.04 ± 35.65	1397.36 ± 10.89	$\mathbf{1430.27 \pm 11.44}$
SVHN→Synth	760.66 ± 8.85	814.63 ± 10.09	$\mathbf{855.51 \pm 11.41}$
Face Datasets	895.41 ± 2.98	902.99 ± 3.69	$\mathbf{913.08 \pm 5.05}$

Almost equal to or better performance than other approaches

[Results] Downstream Classification

	VAE	CVAE	Proposed
USPS→MNIST	0.52 ± 2.15	0.53 ± 0.02	$\mathbf{0.68 \pm 0.01}$
MNIST→USPS	0.64 ± 0.01	0.67 ± 0.01	$\mathbf{0.74 \pm 0.02}$
Synth→SVHN	0.20 ± 0.00	$\mathbf{0.21 \pm 0.00}$	0.19 ± 0.00
SVHN→Synth	0.25 ± 0.01	0.25 ± 0.00	$\mathbf{0.26 \pm 0.00}$

Almost equal to or better performance than other approaches

Conclusion

- Our contribution for the CVAE are summarized as follows:

Theorem 1 shows:

- The **simple prior** is one of the causes of the task-dependency.
- We propose the **optimal prior** to reduce the task-dependency.

Theorem 2 shows:

- Our approach gives a **better lower bound of the log-likelihood**, which enable us to obtain better representation than the CVAE.

Experiments shows:

- Our approach achieves better performance on various datasets.

Thank you for listening!

My paper, slide, and poster are here: