KDD 2022 Research Track

Learning Optimal Priors for Task-Invariant Representations in Variational Autoencoders

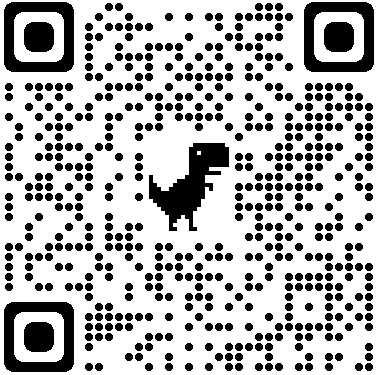
Hiroshi Takahashi[1], Tomoharu Iwata[1], Atsutoshi Kumagai[1], Sekitoshi Kanai[1], Masanori Yamada[1], Yuuki Yamanaka[1], Hisashi Kashima[2]

[1]NTT, [2]Kyoto University

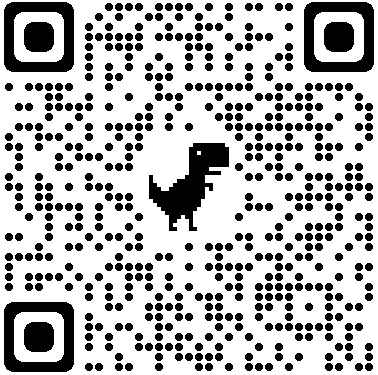
自己紹介

- 名前: 高橋 大志
- 所属: ドコモ <- NTT研究所 (SIC・CD研) / 京大 鹿島研 D3
- 研究: Variational Autoencoder (VAE) の性能改善

twitter

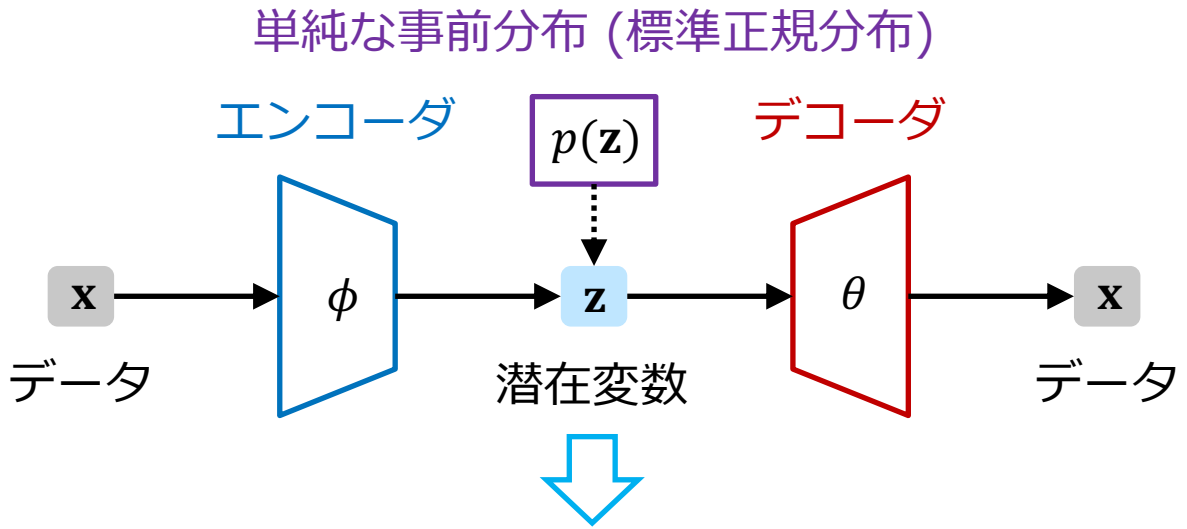


個人ページ



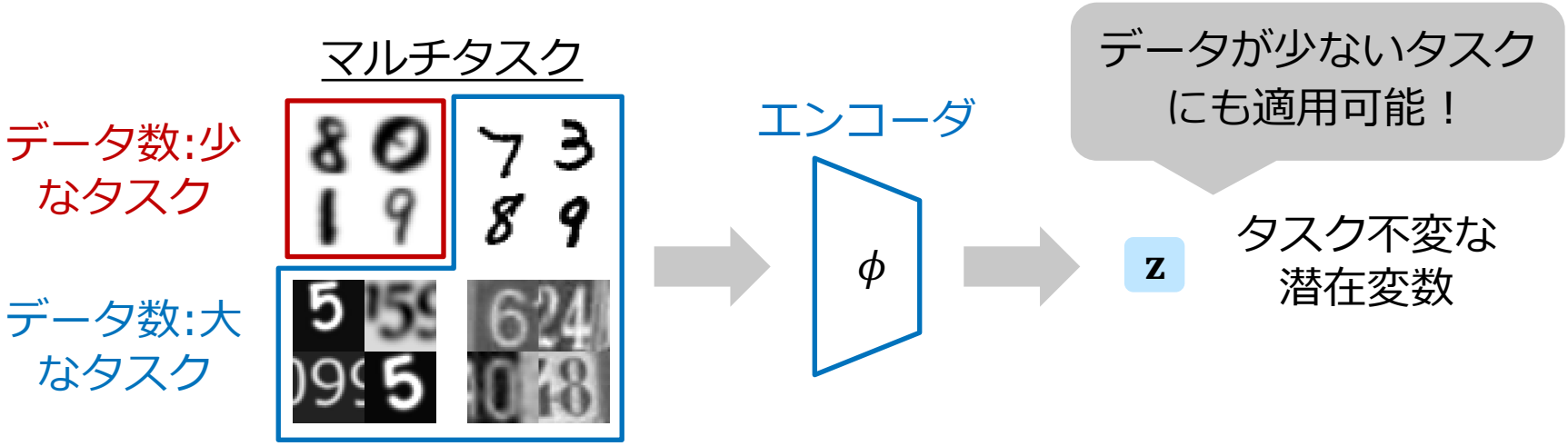
[Introduction] Variational Autoencoder

- Variational autoencoder (VAE) は、教師なし表現学習のための強力な潜在変数モデル



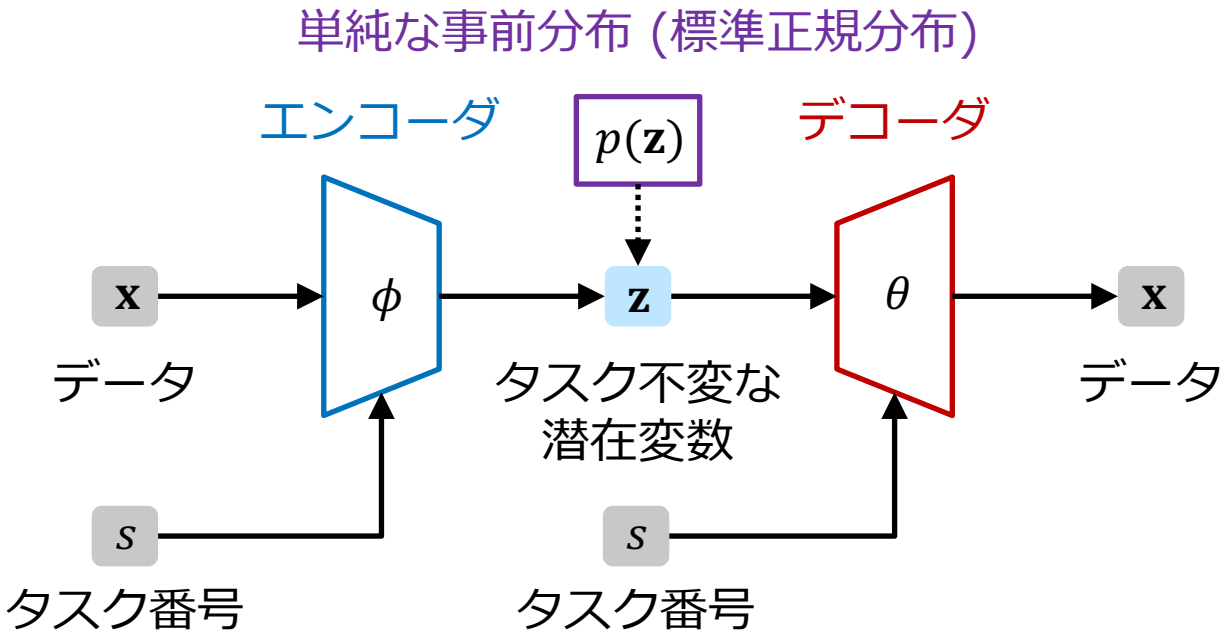
[Introduction] マルチタスク学習

- VAEは強力だが、ニューラルネットワークを用いているため、データ数が不十分な時は性能が極端に低下してしまう
- この問題を解決するため、本研究では複数のタスクからタスク不変な潜在変数を学習することに着目



[Introduction] Conditional VAE

- マルチタスクに対して、タスク不変な潜在変数を学ぼうとする Conditional VAE (CVAE) が広く使われている



[Introduction] CVAEの問題と本研究の貢献

- CVAEは潜在変数へのタスクへの依存性をある程度減らせるが、多くの場合は依存性が残ってしまうことが知られている
- 本研究の貢献は下記の3点:

1. CVAEのタスク依存性の原因を調査し、単純な事前分布を用いていることが一因であることを明らかにした
2. タスク依存性を減らすための最適な事前分布を提案
3. 提案手法を用いて学習した表現が、マルチタスク上で良い性能を発揮することを、理論的・実験的に明らかにした

[Preliminaries] CVAEの定式化

- タスク s が与えられたもとでの $\mathbf{x}$ の確率を以下で定義する:

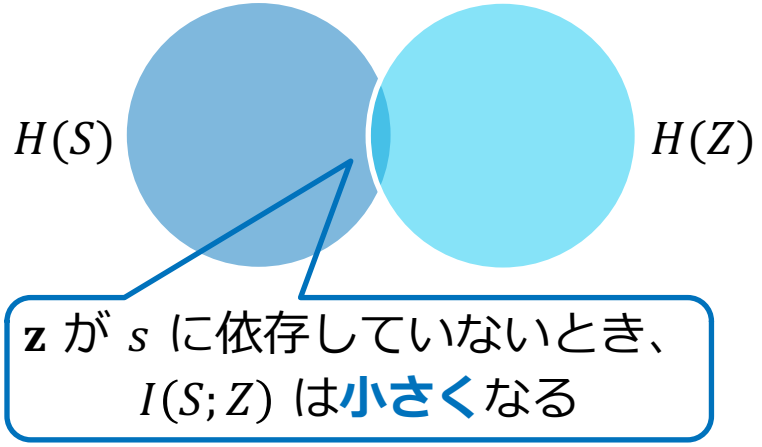
$$p_\theta(\mathbf{x}|s) = \int \underbrace{p_\theta(\mathbf{x}|\mathbf{z},s)}_{\text{デコーダ}}\underbrace{p(\mathbf{z})}_{\text{事前分布}}\mathrm{d}\mathbf{z} = \mathbb{E}_{\underbrace{q_\phi(\mathbf{z}|\mathbf{x},s)}_{\text{エンコーダ}}}\left[\frac{p_\theta(\mathbf{x}|\mathbf{z},s)p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x},s)}\right]$$

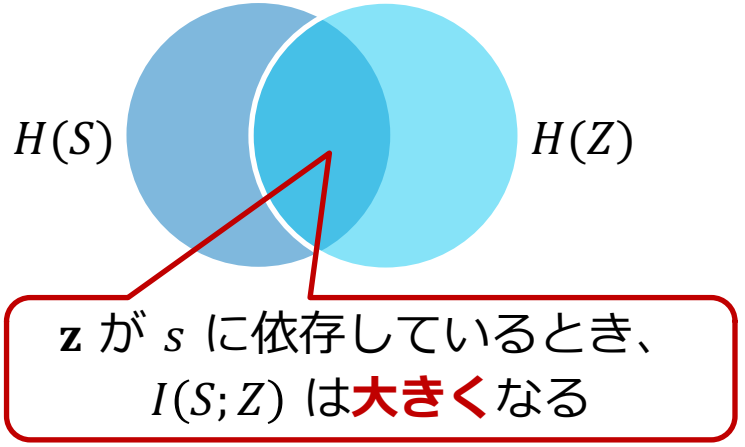
- CVAEは、対数尤度の下界である変分下界 (ELBO)を最大化するように学習される

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

[Preliminaries] 相互情報量

- 潜在変数 $\mathbf{z}$ の タスク s への依存性を調べるために、2つの確率変数間の依存性を測定する**相互情報量** $I(S;Z)$ を導入する





[Proposed] 定理1

- CVAEは、相互情報量 $I(S;Z)$ をその上界である $\mathcal{R}(\phi)$を最小化することで最小化している:

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

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

$s = k$ の時の $\mathbf{x}, \mathbf{z}$ 間の相互情報量

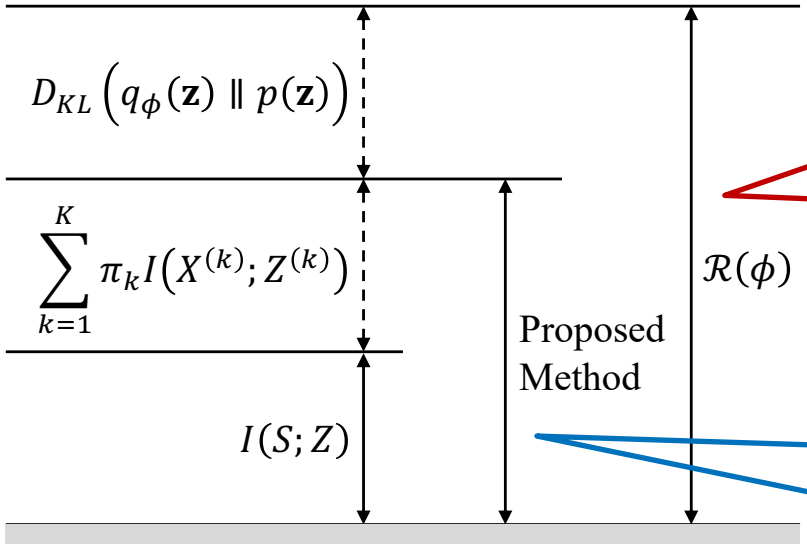
$q_\phi(\mathbf{z}) = \int q_\phi(\mathbf{z}|\mathbf{x},s)p_D(\mathbf{x},s)\mathrm{d}\mathbf{x}$

$\pi_k = p(s=k)$

- しかし, $D_{KL}(q_\phi(\mathbf{z})||p(\mathbf{z}))$ が通常大きい値を取るため、$\mathcal{R}(\phi)$ は $I(S;Z)$ のタイトな上界ではない

[Proposed] 事前分布の効果

- つまり、単純な事前分布 $p(\mathbf{z})$ が**タスク依存性の一因**であり、$q_\phi(\mathbf{z})$ がタスク依存性を減らすための**最適な事前分布**である



$D_{KL}(q_\phi(\mathbf{z})||p(\mathbf{z}))$ が通常大きい値を取るため $\mathcal{R}(\phi)$ は $I(S;Z)$ に対するタイトな上界ではなくなっている

$p(\mathbf{z}) = q_\phi(\mathbf{z})$の時、 $\mathcal{R}(\phi)$ は $I(S;Z)$ に対して最もタイトな上界となる

(他項は事前分布に依存しないため)

[Proposed] 定理2

- 最適な事前分布を用いた変分下界 $\mathcal{F}_{\mathrm{Proposd}}(\theta, \phi)$ は、常に元々の変分下界 $\mathcal{F}_{\mathrm{CVAE}}(\theta, \phi)$ よりも大きい値を取る:

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

- つまり、$\mathcal{F}_{\mathrm{Proposd}}(\theta, \phi)$ は $\mathcal{F}_{\mathrm{CVAE}}(\theta, \phi)$ と比べて、**より良い対数尤度の下界**になっている
- 対数尤度を大きくするほうがより良い生成モデルとなるため、提案手法のほうがより良い表現を学習できる

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

- $\mathcal{F}_{\mathrm{Proposd}}(\theta, \phi) = \mathcal{F}_{\mathrm{CVAE}}(\theta, \phi) + D_{KL}(q_\phi(\mathbf{z})||p(\mathbf{z}))$ は、 KL情報量 $D_{KL}(q_\phi(\mathbf{z})||p(\mathbf{z}))$ を計算することで最適化できる:

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

- $q_\phi(\mathbf{z})/p(\mathbf{z})$ は、2つの確率分布の比を、両分布からのサンプルを用いて近似できる**密度比推定**を用いて近似することができる (Section 3.3 参照)

[Proposed] 理論的な貢献

- 本研究の理論的な貢献は下記:

定理1

- **単純な事前分布**がタスク依存性の一因であることを明らかにした
- タスク依存性を減らす**最適な事前分布**として$q_\phi(\mathbf{z})$ を導入

定理2

- $\mathcal{F}_{\text{Proposd}}(\theta, \phi)$ は**良い対数尤度の下界**であり、CVAEよりも良い表現の学習を可能にする

- 続いて、実験的に提案手法の評価を行う

[Experiments] データセット

- 手書き数字 (USPS and MNIST)、住居番号 (SynthDigits and SVHN)、顔画像 (Frey, Olivetti, and UMist) のデータセットを用いて評価する

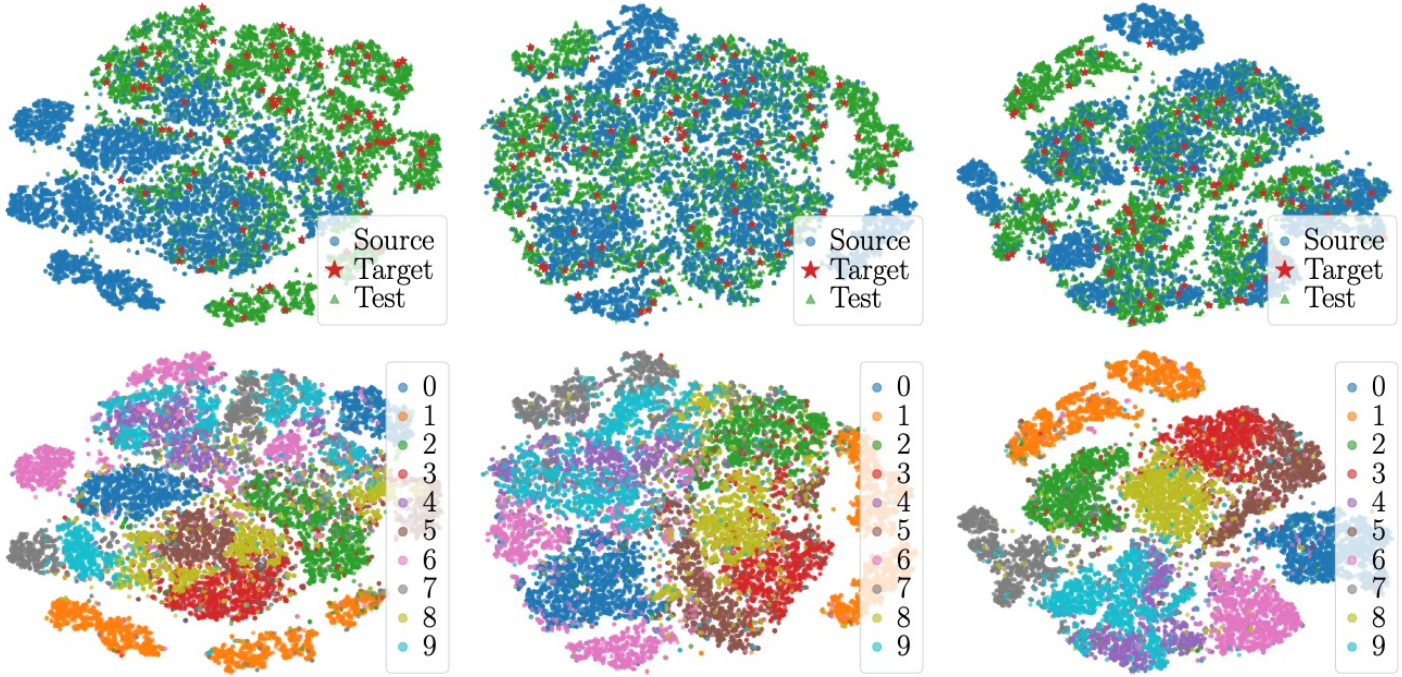
	Dimension	Train size	Valid size	Test size
USPS	784	6,438	1,000	1,860
MNIST	784	10,000	10,000	10,000
SynthDigits	1,024	10,000	10,000	9,553
SVHN	1,024	10,000	10,000	26,032
Frey	560	1,565	200	200
Olivetti	560	150	100	150
UMist	560	300	75	200

[Experiments] 設定

- 手書き数字に対して、2タスク上で学習し、ターゲットタスク上で性能を評価する実験を行う:
  - ソースタスクは大量のデータがある
  - ターゲットタスクは100個だけデータがある
  - ペアは (USPS→MNIST)、(MNIST→USPS)、(SynthDigits→SVHN)、(SVHN→SynthDigits) の4通り
- 顔画像に対して、1個の学習器を3タスク上で学習し、各タスク上での性能を評価する実験を行う
  - 顔画像はデータ数が少ないため、全てのタスクでデータが少ない場合での性能を評価できる

[Results] 定性評価: 表現の可視化

Visualization of latent variables on USPS→MNIST



[Results] 定量評価: 密度推定

	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	$\mathbf{913.08 \pm 5.05}$

他の手法と比べて同等もしくはそれ以上の性能を達成

[Results] 定量評価: 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	$\mathbf{0.26 \pm 0.00}$

他の手法と比べて同等もしくはそれ以上の性能を達成

まとめ

- 本研究の貢献は下記の通り：

定理1

- **単純な事前分布**がタスク依存性の一因であることを明らかにした
- タスク依存性を減らす**最適な事前分布**として$q_\phi(\mathbf{z})$ を導入

定理2

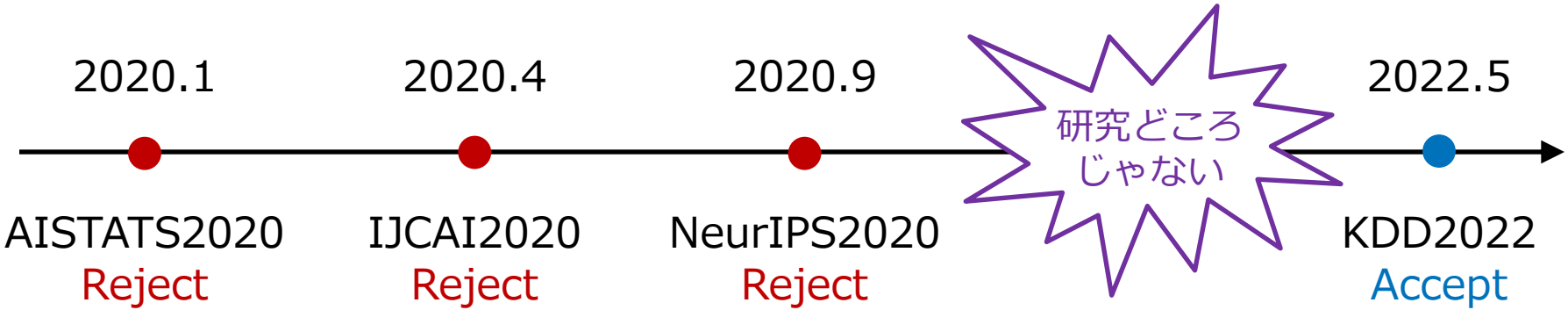
- $\mathcal{F}_{\text{Proposd}}(\theta, \phi)$ は**良い対数尤度の下界**であり、CVAEよりも良い表現の学習を可能にする

実験

- 提案手法が複数のデータセット上でより良い性能を達成

Road to KDD2022 Acceptance

- 他の会議で3回リジェクトされました (しかも2年越しです)



- 一番大きかったのは論文のストーリーの変更
  - 元々は「VAEを用いたマルチタスク密度推定」というストーリーでしたが、「マルチタスク表現学習」に変更したところ、無事採択されました
  - 心が折れた時には一から見直すのが良いという学びでした

ご清聴ありがとうございます

論文、スライド、ポスターは下記にあります