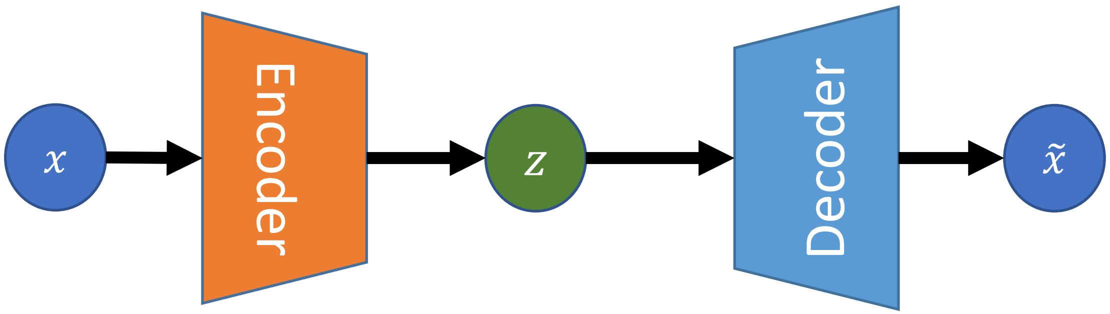
VAE

base and application

变分自编码器

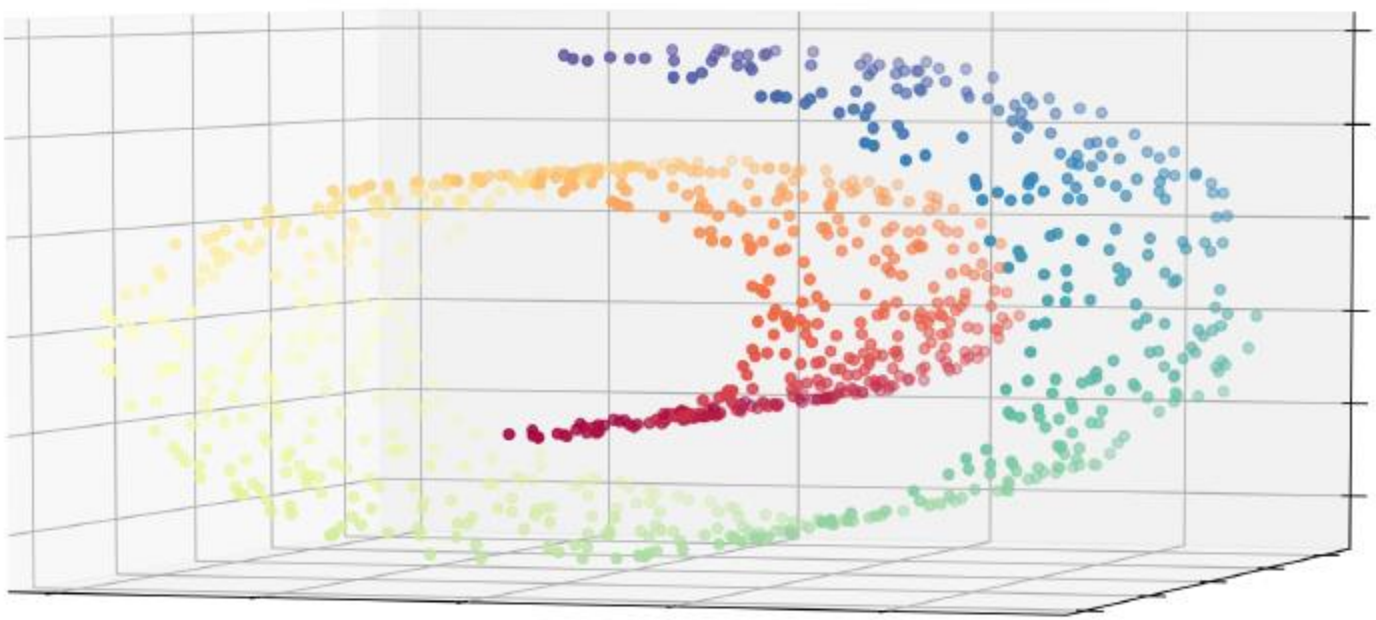
蔡泽锋 5.8

Auto-Encoder



Method	Parametric	Convex
PCA / classical MDS	N	Y (Dense)
Kernel PCA	N	Y (Dense)
Isomap	N	Y (Dense)
LLE	N	Y (Sparse)
Laplacian Eigenmaps	N	Y (Sparse)
tSNE	N	N
Autoencoder	Y	N

Auto-Encoder



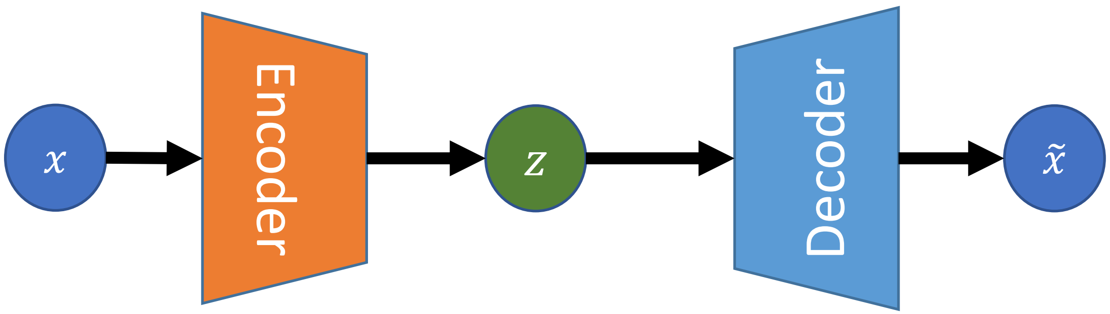
2d瑞士卷流形嵌入了3d

Manifold Hypothesis: the observed data lie on a low-dimensional manifold embedded in a higher-dimensional space.

Auto-Encoder

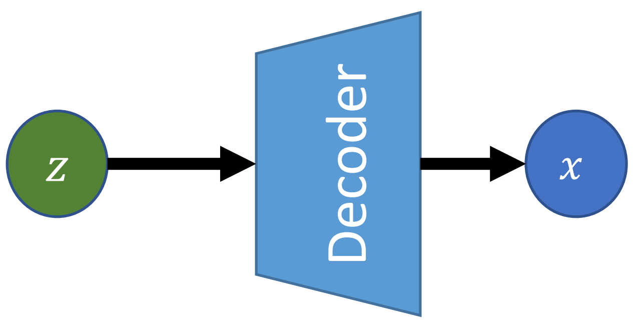
MLP-AE 隐空间 + tSNE plot

VAE



How to generate?

Intuition (remember from autoencoders!):
x is an image, **z** is latent factors used to generate **x:** attributes, orientation, etc.



VAE

Data likelihood: $p_\theta(x) = \int p_\theta(z) p_\theta(x|z) dz$

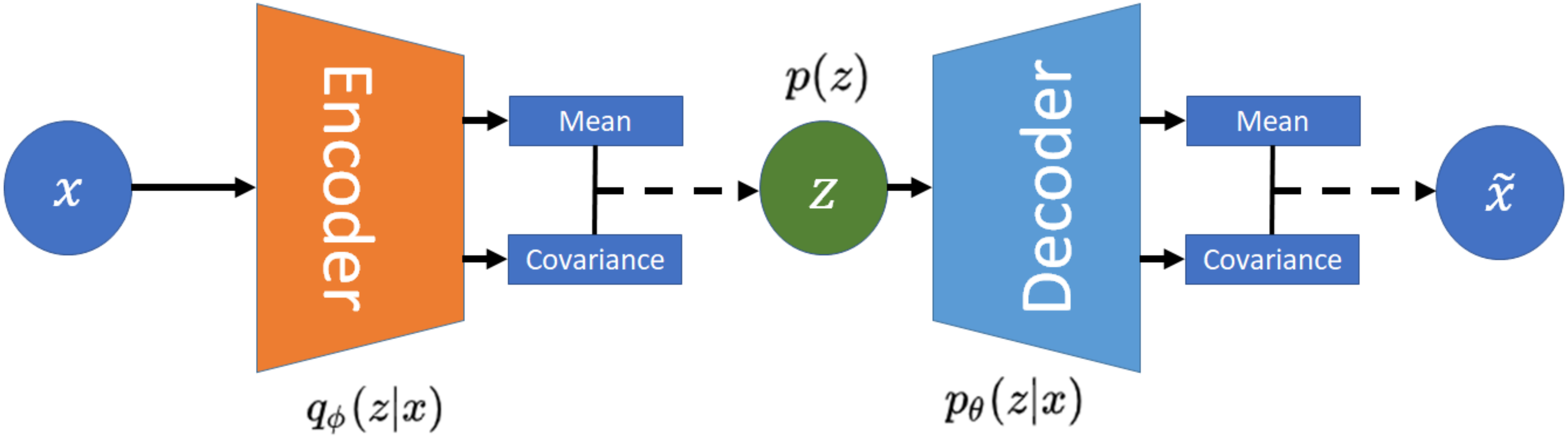
Intractible to compute
p(x|z) for every z!

Posterior density also intractable: $p_\theta(z|x) = p_\theta(x|z) p_\theta(z) / p_\theta(x)$

Intractable data likelihood

Solution: In addition to decoder network modeling $p_\theta(x|z)$, define additional encoder network $q_\phi(z|x)$ that approximates $p_\theta(z|x)$

VAE



$$p(x) = \int p(z)p(x|z)dz$$

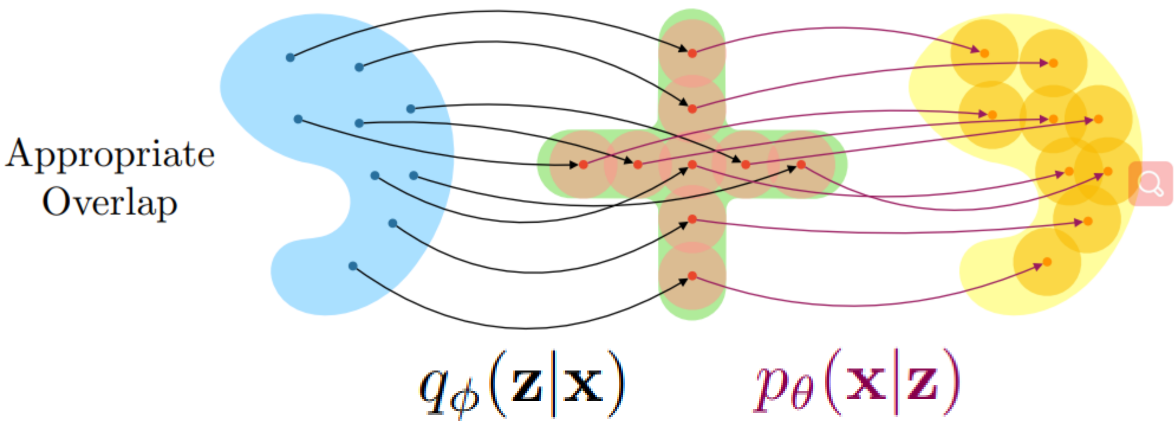
VAE

$$
\begin{aligned}
\log p_\theta(x^{(i)}) &= \mathbf{E}_{z\sim q_\phi(z|x^{(i)})}\left[\log p_\theta(x^{(i)})\right] \qquad (p_\theta(x^{(i)}) \text{ Does not depend on } z) \\
&= \mathbf{E}_z\left[\log \frac{p_\theta(x^{(i)} \mid z)p_\theta(z)}{p_\theta(z \mid x^{(i)})}\right] \qquad \text{(Bayes' Rule)} \\
&= \mathbf{E}_z\left[\log \frac{p_\theta(x^{(i)} \mid z)p_\theta(z)}{p_\theta(z \mid x^{(i)})}\frac{q_\phi(z \mid x^{(i)})}{q_\phi(z \mid x^{(i)})}\right] \qquad \text{(Multiply by constant)} \\
&= \mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - \mathbf{E}_z\left[\log \frac{q_\phi(z \mid x^{(i)})}{p_\theta(z)}\right] + \mathbf{E}_z\left[\log \frac{q_\phi(z \mid x^{(i)})}{p_\theta(z \mid x^{(i)})}\right] \qquad \text{(Logarithms)} \\
&= \mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z)) + D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z \mid x^{(i)}))
\end{aligned}
$$

$$
\underbrace{-KL(p_\theta(z|x)\|p(z))}_{KL-loss} + \underbrace{E_{p_\theta(z|x)}(q_\phi(x|z))}_{reconstruction-loss}
$$

VAE

Target
Structure
$p(\mathbf{z})$



Appropriate
Overlap
$q_\phi(\mathbf{z}|\mathbf{x})$
$p_\theta(\mathbf{x}|\mathbf{z})$

VAE

$p(z)$

图 3：VAE插值从左到右的渐变

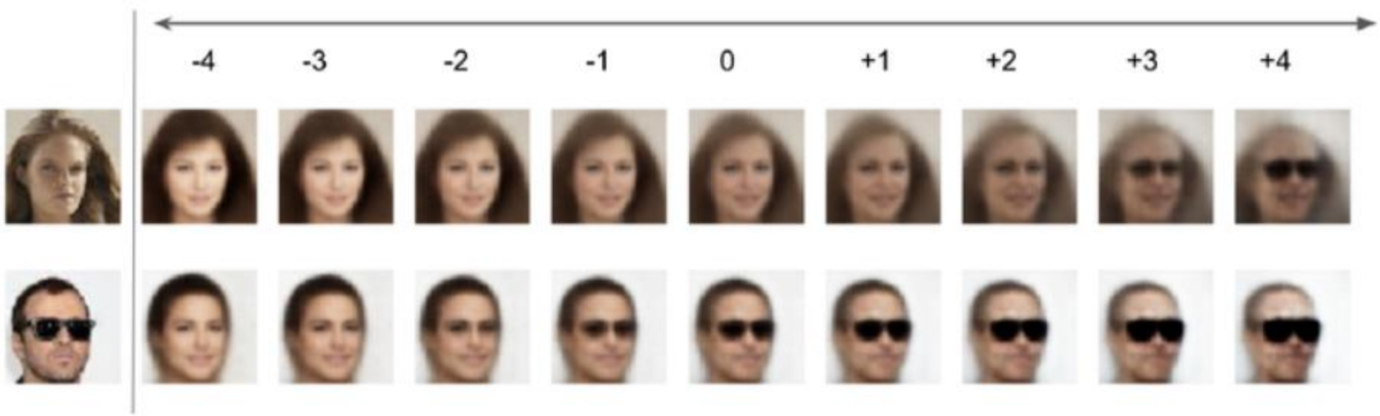


图 4：VAE外推，-去掉太阳镜，+加上太阳镜，0对应原始的左图

	s_1		s_2		s_3		s_1 - s_2 + s_3
RG	I am one.	-	I am two.	+	You are two.	=	You ready no.
RGP	I am one.	-	I am two.	+	You are two.	=	**You are one.**
RG	A word in a phrase.	-	A tree in a phrase.	+	A tree is green.	=	A word is purevy?
RGP	A word in a phrase.	-	A tree in a phrase.	+	A tree is green.	=	**A word is green.**
RP	A large number of people want to work.	-	A small number of people want to work.	+	A small sentence is enough.	=	A large senselfeir in or evacce.
RGP	A large number of people want to work.	-	A small number of people want to work.	+	A small sentence is enough.	=	A large sector for challenge.

Text extrapolation

t =0	in new york the company declined comment
t =0.1	in new york the company declined comment
t =0.2	in new york the transaction was suspended
t =0.3	in the securities company said yesterday
t =0.4	in other board the transaction had disclosed
t =0.5	other of those has been available
t =0.6	both of companies have been unchanged
t =0.7	both men have received a plan to restructure
t =0.8	and to reduce that it owns
t =0.9	and to continue to make prices
t =1	and they plan to buy more today

Text Interpolation

Disentangled Representation Learning for Non-Parallel Text Style Transfer

Vineet John
University of Waterloo
vineet.john@uwaterloo.ca

Lili Mou
AdeptMind Research
doublepower.mou@gmail.com
lili@adeptmind.ai

Hareesh Bahuleyan
University of Waterloo
hpallika@uwaterloo.ca

Olga Vechtomova
University of Waterloo
ovechtom@uwaterloo.ca

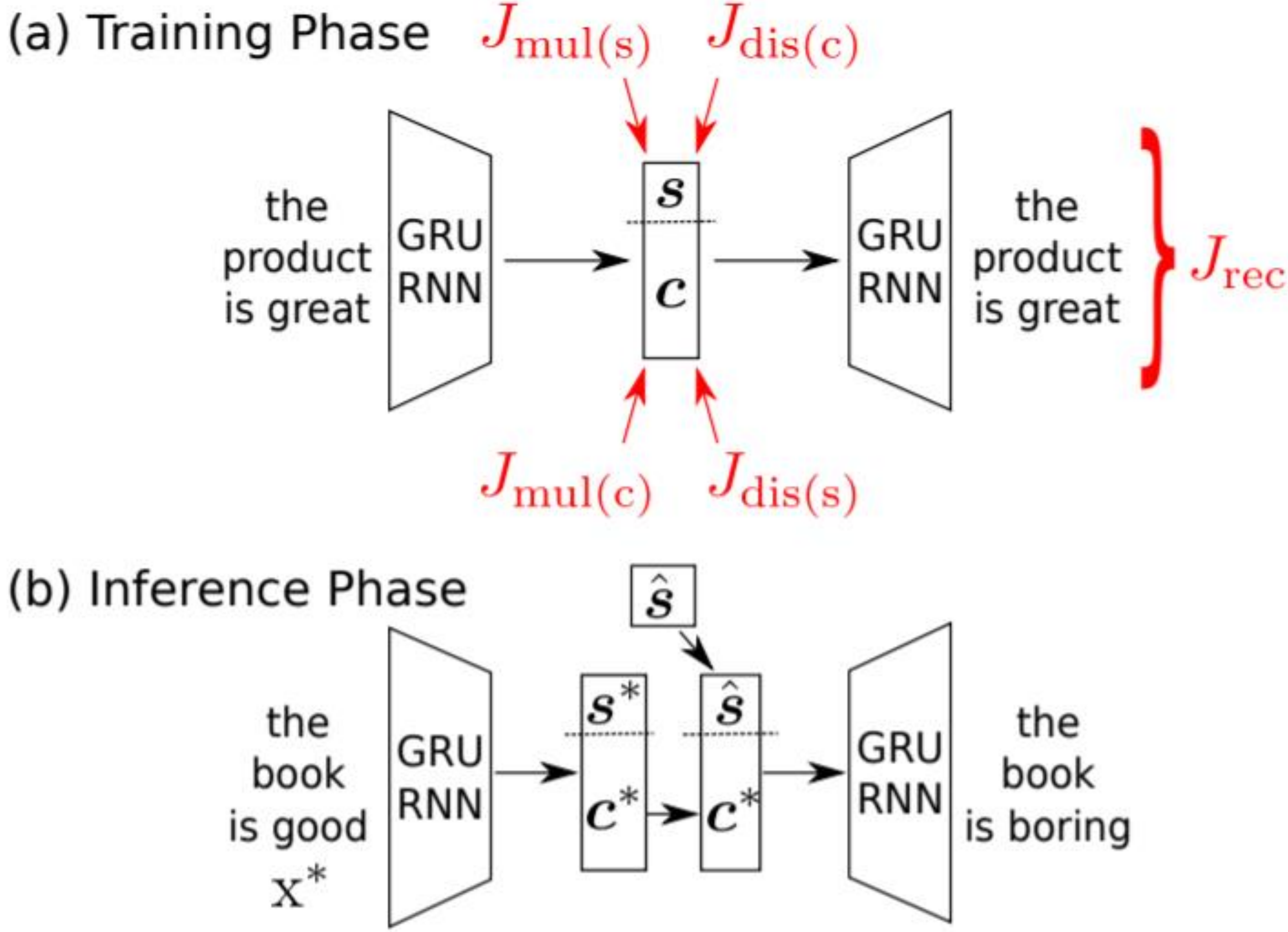


Figure 1: Overview of our approach.

Multi-Task Loss for Style.

$$\boldsymbol{y}_s = \mathrm{softmax}(W_{\text{mul(s)}}\boldsymbol{s} + \boldsymbol{b}_{\text{mul(s)}})$$

$$J_{\text{mul(s)}}(\boldsymbol{\theta}_{\text{E}}; \boldsymbol{\theta}_{\text{mul(s)}}) = -\sum\nolimits_{l\in\text{labels}} t_s(l) \log y_s(l)$$

Adversarial Loss for Style.

$$\boldsymbol{y}_s = \mathrm{softmax}(W_{\text{dis(s)}}\boldsymbol{c} + \boldsymbol{b}_{\text{dis(s)}})$$

$$J_{\text{dis(s)}}(\boldsymbol{\theta}_{\text{dis(s)}}) = -\sum\nolimits_{l\in\text{labels}} t_c(l) \log y_s(l)$$

$$J_{\text{adv(s)}}(\boldsymbol{\theta}_{\text{E}}) = \mathcal{H}(\boldsymbol{y}_s|\boldsymbol{c}; \boldsymbol{\theta}_{\text{dis(s)}}) \quad \mathcal{H}(\boldsymbol{p}) = -\sum\nolimits_{i\in\text{labels}} p_i \log p_i$$

Multi-Task Loss for Content.

$$\boldsymbol{y}_c = \mathrm{softmax}(W_{\text{mul(c)}}\boldsymbol{c} + \boldsymbol{b}_{\text{mul(c)}})$$

$$J_{\text{mul(c)}}(\boldsymbol{\theta}_{\text{E}}; \boldsymbol{\theta}_{\text{mul(c)}}) = -\sum\nolimits_{w\in\text{vocab}} t_c(w) \log y_c(w)$$

Adversarial Loss for Content.

$$\boldsymbol{y}_c = \mathrm{softmax}({W_{\text{dis(c)}}}^\top \boldsymbol{s} + \boldsymbol{b}_{\text{dis(c)}})$$

$$J_{\text{dis(c)}}(\boldsymbol{\theta}_{\text{dis(c)}}) = -\sum\nolimits_{w\in\text{vocab}} t_c(w) \log y_c(w)$$

$$J_{\text{adv(c)}}(\boldsymbol{\theta}_{\text{E}}) = \mathcal{H}(\boldsymbol{y}_c|\boldsymbol{s}; \boldsymbol{\theta}_{\text{dis(c)}})$$

$$\boxed{\begin{aligned} J_{\text{ovr}} =& J_{\text{AE}}(\boldsymbol{\theta}_{\text{E}}, \boldsymbol{\theta}_{\text{D}}) \\ &+ \lambda_{\text{mul(s)}} J_{\text{mul(s)}}(\boldsymbol{\theta}_{\text{E}}, \boldsymbol{\theta}_{\text{mul(s)}}) - \lambda_{\text{adv(s)}} J_{\text{adv(s)}}(\boldsymbol{\theta}_{\text{E}}) \\ &+ \lambda_{\text{mul(c)}} J_{\text{mul(c)}}(\boldsymbol{\theta}_{\text{E}}, \boldsymbol{\theta}_{\text{mul(c)}}) - \lambda_{\text{adv(c)}} J_{\text{adv(c)}}(\boldsymbol{\theta}_{\text{E}}) \end{aligned}}$$

$$\hat{\boldsymbol{s}} = \frac{\sum_{i \in \text{target style}} \boldsymbol{s}_i}{\#\ \text{target style samples}}$$

Experiment I: Disentangling Latent Space

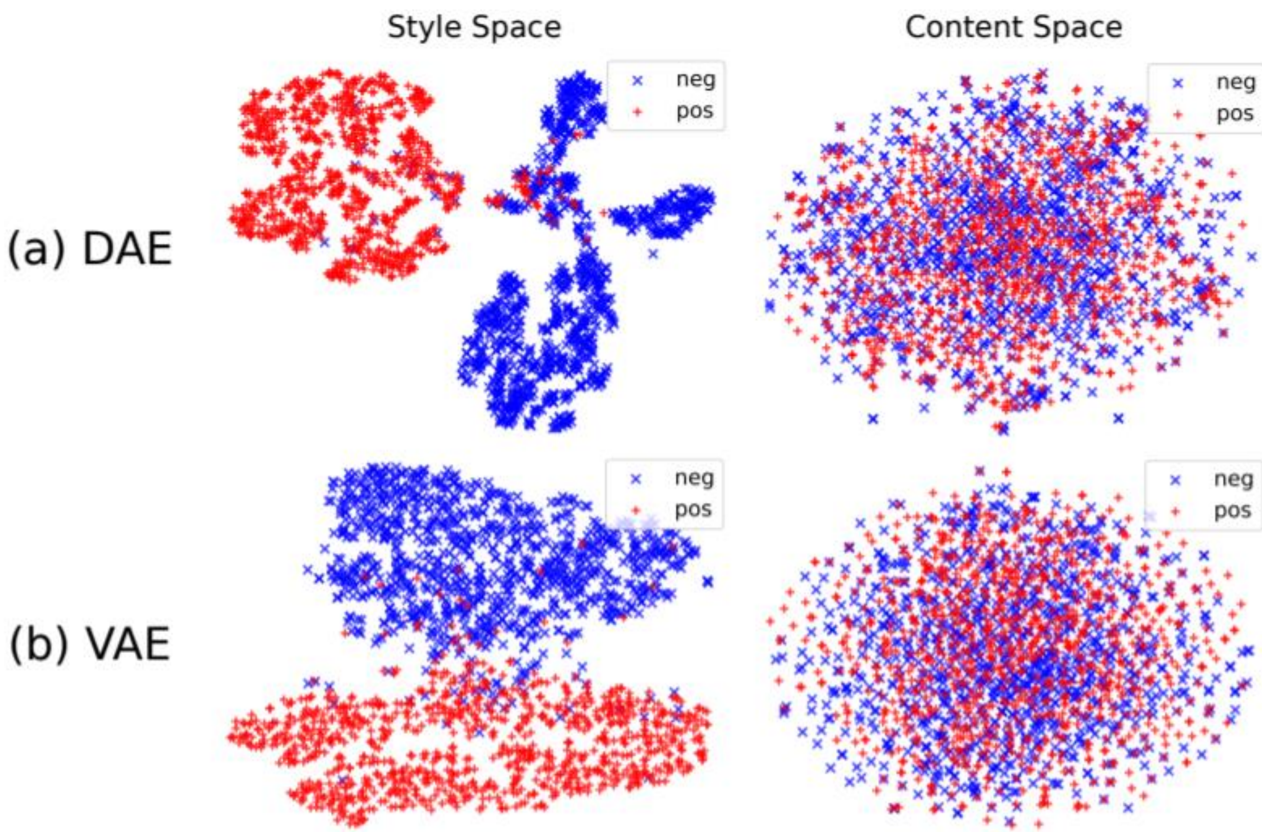


Figure 2: t-SNE plots of the disentangled style and content spaces (with all auxiliary losses on the Yelp dataset).

Experiment II: Non-Parallel Text Style Transfer

Model	Yelp Dataset				Amazon Dataset			
	Transfer Accuracy	Cosine Similarity	Word Overlap	Language Fluency	Transfer Accuracy	Cosine Similarity	Word Overlap	Language Fluency
Style-Embedding (Fu et al. 2018)	0.182	**0.959**	**0.666**	-16.17	0.400[†]	**0.930**[†]	**0.359**	-28.13
Cross-Alignment (Shen et al. 2017)	0.784[†]	0.892	0.209	-23.39	0.606	0.893	0.024	-26.31
Multi-Decoder (Zhao et al. 2018)	0.818[†]	0.883	0.272	-20.95	0.552	*0.926*	0.169	-34.70
Ours (DAE)	0.883	*0.915*	*0.549*	-10.17	0.720	0.921	*0.354*	-24.74
Ours (VAE)	**0.934**	0.904	0.473	**-9.84**	**0.822**	0.900	0.196	**-21.70**

Objectives	Transfer Accuracy	Cosine Similarity	Word Overlap	Language Fluency
J_{AE}	0.106	**0.939**	0.472	-12.58
J_{AE}, $J_{\text{mul(s)}}$	0.767	0.911	0.331	-12.17
J_{VAE}, $J_{\text{adv(s)}}$	0.782	0.886	0.230	-12.03
J_{VAE}, $J_{\text{mul(s)}}$, $J_{\text{adv(s)}}$	0.912	0.866	0.171	-9.59
J_{VAE}, $J_{\text{mul(s)}}$, $J_{\text{adv(s)}}$, $J_{\text{mul(c)}}$, $J_{\text{adv(c)}}$	**0.934**	0.904	**0.473**	**-9.84**

A Batch Normalized Inference Network Keeps the KL Vanishing Away

Qile Zhu[1*], Jianlin Su[*], Wei Bi[2], Xiaojiang Liu[2], Xiyao Ma[1], Xiaolin Li[3] and Dapeng Wu[1]
[1]University of Florida, [2]Tencent AI Lab, [3]AI Institute, Tongdun Technology
{valder,maxiy,dpwu}@ufl.edu
{victoriabi,kieranliu}@tencent.com
xiaolin.li@tongdun.net
bojone@spaces.ac.cn

An extension for the original ACL 2020 paper

Posterior collapse

$$\underbrace{-KL(p_\theta(z|x)\|p(z))}_{KL-loss} + \underbrace{E_{p_\theta(z|x)}(q_\phi(x|z))}_{reconstruction-loss}$$

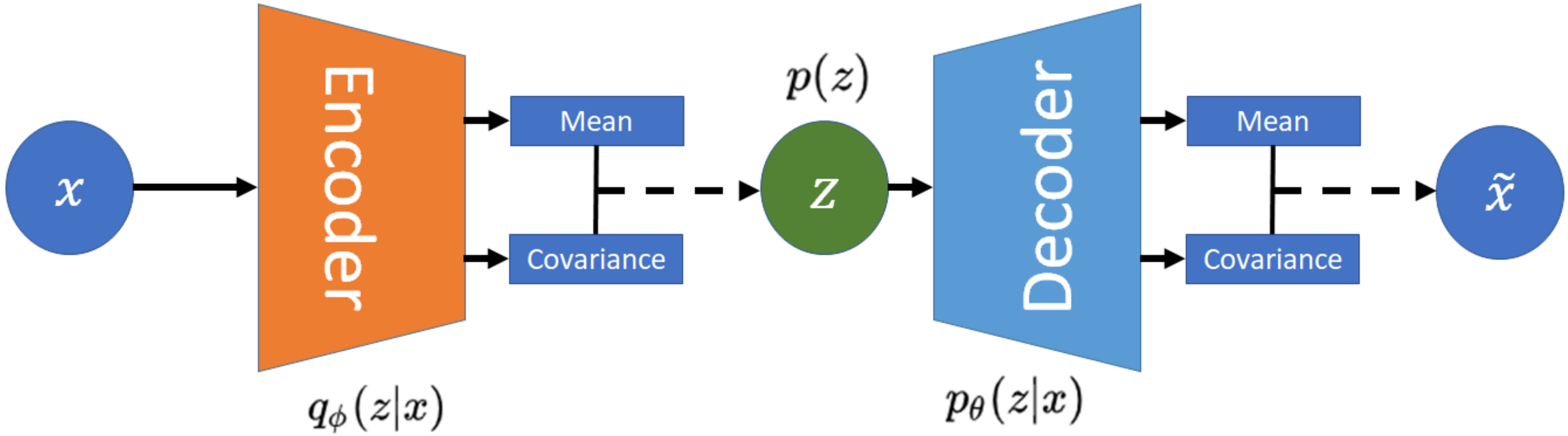
KL-Loss vanish!

As Nature Language is a discrete sequence, if we use a **autoregressive decoder**, the decoder may be too strong, and as encoder contain **noise**, finally it ignore the information passed from encoder !

Example:

Original X:I have a meeting today.
Normal Y: I have a meeting today.
KL vanish Y: I have a breakfast in the morning.

Original X: I am not want to sleep.
Normal X: I am not want to sleep.
KL vanish: I want to sleep.



$$\mu = f_\mu(\boldsymbol{x})$$

$$\boldsymbol{\Sigma} = diag(f_\Sigma(\boldsymbol{x}))$$

$$KL = \frac{1}{2b}\sum_{j=1}^{b}\sum_{i=1}^{n}(\mu_{ij}^2 + \sigma_{ij}^2 - \log\sigma_{ij}^2 - 1)$$

$$= \frac{1}{2}\sum_{i=1}^{n}(\frac{\sum_{j=1}^{b}\mu_{ij}^2}{b} + \frac{\sum_{j=1}^{b}\sigma_{ij}^2}{b}$$

$$- \frac{\sum_{j=1}^{b}\log\sigma_{ij}^2}{b} - 1).$$

$\sum_{j=1}^{b} \mu_{ij}^2 / b$ $\quad$ $\mathrm{E}[\mu_i^2] = \mathrm{Var}[\mu_i] + \mathrm{E}^2[\mu_i].$

$\sum_{j=1}^{b} \sigma_{ij}^2 / b$ $\quad$ $\mathrm{E}[\sigma_i^2]$

$\sum_{j=1}^{b} \log \sigma_{ij}^2 / b$ $\quad$ $\mathrm{E}[\log \sigma_i^2]$

$$
\begin{aligned}
\mathrm{E}[KL] = \frac{1}{2} \sum_{i=1}^{n} (\mathrm{Var}[\mu_i] + \mathrm{E}^2[\mu_i] \\
+ \mathrm{E}[\sigma_i^2] - \mathrm{E}[\log \sigma_i^2] - 1) \\
\geq \frac{1}{2} \sum_{i=1}^{n} (\mathrm{Var}[\mu_i] + \mathrm{E}^2[\mu_i]),
\end{aligned}
$$

$\mathrm{E}[\sigma_i^2 - \log \sigma_i^2] \geq 1$ $\quad$ $e^x - x \geq 1$

$$\hat{\mu}_i = \gamma \frac{\mu_i - \mu_{\mathcal{B}i}}{\sigma_{\mathcal{B}i}} + \beta,$$

$$\mathrm{E}[KL] \geq \frac{1}{2} \sum_i^n (\mathrm{Var}[\mu_i] + \mathrm{E}^2[\mu_i])$$
$$= \frac{n \cdot (\gamma^2 + \beta^2)}{2}.$$

Algorithm 1 BN-VAE training.

1: Initialize $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$.
2: **for** $i = 1, 2, \cdots$ Until Convergence **do**
3: Sample a mini-batch $\mathbf{x}$.
4: $\mu, \log \sigma^2 = f_\phi(\mathbf{x})$.
5: $\mu' = BN_{\gamma,\beta}(\mu)$.
6: Sample $\mathbf{z} \sim \mathcal{N}(\mu', \sigma^2)$ and reconstruct $\mathbf{x}$ from $f_\theta(\mathbf{z})$.
7: Compute gradients $\mathbf{g}_{\phi,\theta} \leftarrow \nabla_{\phi,\theta} \mathcal{L}(\mathbf{x}; \phi, \theta)$.
8: Update ϕ, θ using $\mathbf{g}_{\phi,\theta}$.
9: **end for**

Further Extension

$$q(z) = \int \hat{p}(x) p(z|x) dx$$
$$= \int \hat{p}(x) \mathcal{N}(z; \mu(x), \sigma(x)) dx.$$

$$0 = \int \hat{p}(x) \mu(x) dx = \mathrm{E}_{x \sim \hat{p}(x)}[\mu(x)].$$

$$\beta_\mu = 0$$

$$1 = \int \hat{p}(x) [\mu(x)^2 + \sigma(x)^2] dx$$
$$= \mathrm{E}_{x \sim \hat{p}(x)}[\mu(x)^2] + \mathrm{E}_{x \sim \hat{p}(x)}[\sigma(x)^2]$$

$$\beta_\mu^2 + \gamma_\mu^2 + \beta_\sigma^2 + \gamma_\sigma^2 = 1$$

$$\beta_\mu = \beta_\sigma = 0$$
$$\gamma_\mu = \sqrt{\tau + (1 - \tau) \cdot sigmoid(\theta)}$$
$$\gamma_\sigma = \sqrt{(1 - \tau) \cdot sigmoid(-\theta)},$$

where $\tau \in (0, 1)$ and θ is a trainable parameter.

Experiments

	Yahoo				Yelp			
Model	**NLL**	**KL**	**MI**	**AU**	**NLL**	**KL**	**MI**	**AU**
				Without a pretrained AE encoder				
CNN-VAE	$\leq$332.1	10.0	-	-	$\leq$359.1	7.6	-	-
LSTM-LM	328	-	-	-	351.1	-	-	-
VAE	328.6	0.0	0.0	0.0	357.9	0.0	0.0	0.0
β-VAE (0.4)	328.7	6.3	2.8	8.0	358.2	4.2	2.0	4.2
cyclic *	330.6	2.1	2.0	2.3	359.5	2.0	1.9	4.1
Skip-VAE *	328.5	2.3	1.3	8.1	357.6	1.9	1.0	7.4
SA-VAE	327.2	5.2	2.7	9.8	**355.9**	2.8	1.7	8.4
Agg-VAE	**326.7**	5.7	2.9	15.0	**355.9**	3.8	2.4	11.3
FB (4)	331.0	4.1	3.8	3.0	359.2	4.0	1.9	32.0
FB (5)	330.6	5.7	2.0	3.0	359.8	4.9	1.3	32.0
δ-VAE (0.1) *	330.7	3.2	0.0	0.0	359.8	3.2	0.0	0.0
vMF-VAE (13) *	327.4	2.0	-	32.0	357.5	2.0	-	32.0
BN-VAE (0.6) *	**326.7**	6.2	5.6	32.0	356.5	6.5	5.4	32.0
BN-VAE (0.7) *	327.4	8.8	7.4	32.0	**355.9**	9.1	7.4	32.0
				With a pretrained AE encoder				
cyclic *	333.1	25.8	9.1	32.0	361.5	20.5	9.3	32.0
FB (4) *	**326.2**	8.1	6.8	32.0	356.0	7.6	6.6	32.0
δ-VAE (0.15) *	331.0	5.6	1.1	11.2	359.4	5.2	0.5	5.9
vMF-VAE (13) *	328.4	2.0	-	32.0	357.0	2.0	-	32.0
BN-VAE (0.6) *	326.7	6.4	5.8	32.0	**355.5**	6.6	5.9	32.0
BN-VAE (0.7) *	326.5	9.1	7.6	32.0	355.7	9.1	7.5	32.0

#label	100	500	1k	2k	10k
AE	81.1	86.2	90.3	89.4	94.1
VAE	66.1	82.6	88.4	89.6	94.5
δ-VAE	61.8	61.9	62.6	62.9	93.8
Agg-VAE	80.9	85.9	88.8	90.6	93.7
cyclic	62.4	75.5	80.3	88.7	94.2
FB (9)	79.8	84.4	88.8	91.12	94.7
AE+FB (6)	87.6	90.2	92.0	93.4	94.9
BN-VAE (0.7)	**88.8**	**91.6**	**92.5**	**94.1**	**95.4**

Table 3: Accuracy on Yelp.

Model	CVAE	CVAE (BOW)	BN-VAE
PPL	36.40	24.49	30.67
KL	0.15	9.30	5.18
BLEU-4	10.23	8.56	8.64
A-bow Prec	95.87	96.89	96.64
A-bow Recall	90.93	93.95	94.43
E-bow Prec	86.26	83.55	84.69
E-bow Recall	77.91	81.13	81.75

Table 4: Comparison on dialogue generation.

Model	Fluency			Relevance			Informativeness		
	Avg	#Accept	#High	Avg	#Accept	#High	Avg	#Accept	#High
CVAE	2.11 (0.58)	87%	23%	1.90 (0.49)	82%	8%	1.39 (0.59)	34%	5%
CVAE (BOW)	2.08 (0.73)	84%	23%	1.86 (0.58)	75%	11%	**1.54** (0.65)	46%	8%
BN-CVAE	**2.16** (0.71)	88%	27%	**1.92** (0.67)	80%	12%	**1.54** (0.67)	43%	10%

Table 5: Human evaluation results. Numbers in parentheses is the corresponding variance on 200 test samples.

Topic: ETHICS IN GOVERNMENT		
Context: have trouble drawing lines as to what's illegal and what's not		
Target (statement): well i mean the other problem is that they're always up for		
CVAE	CVAE (BOW)	BN-CVAE
1. yeah	1. yeah	1. it's not a country
2. yeah	2. oh yeah they're not	2. it is the same thing that's what i think is about the state is a state
3. yeah	3. no it's not too bad	3. yeah it's

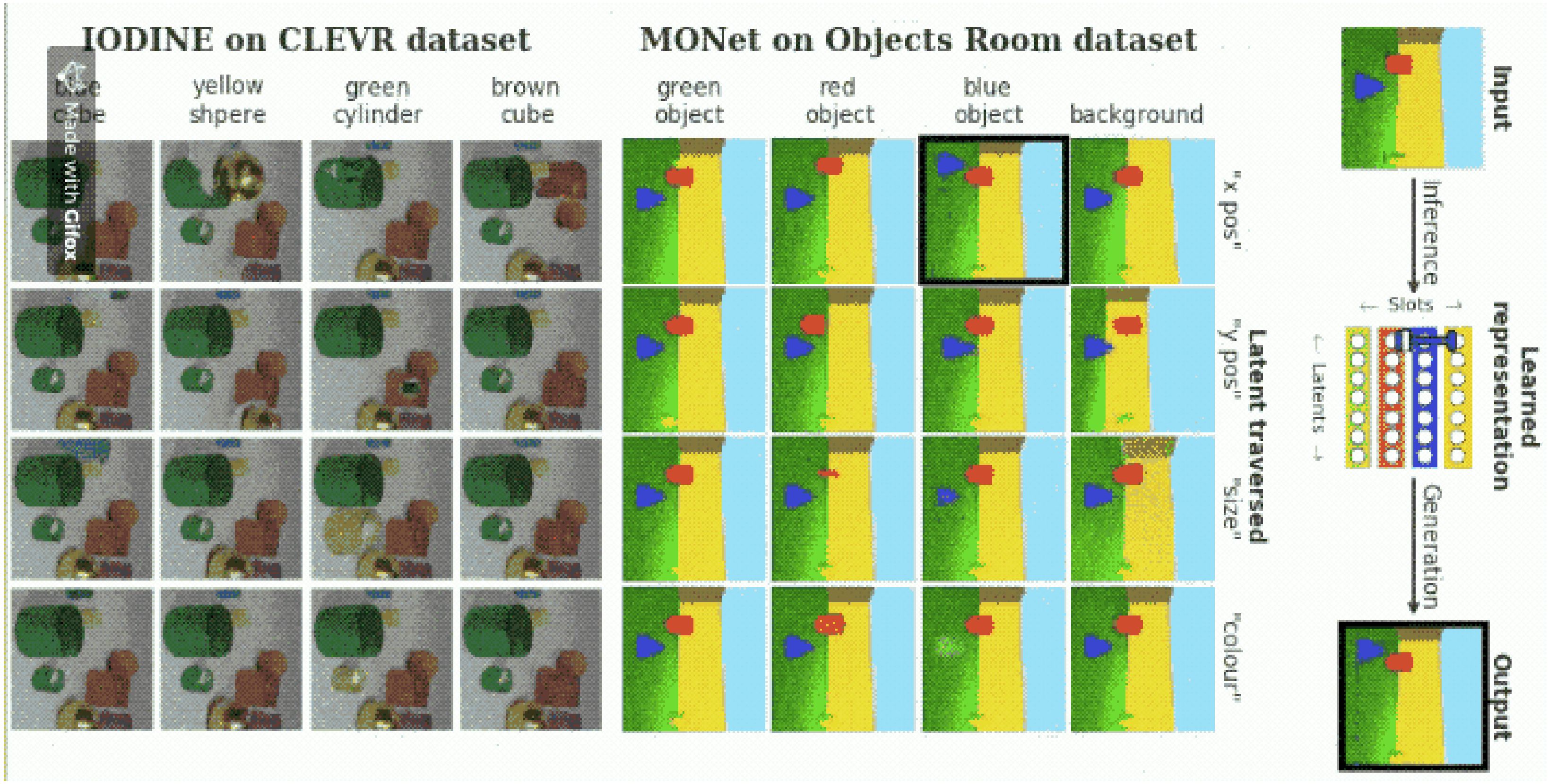
Table 6: Sampled generated responses. Only the last sentence in the context is shown here.

Disentangling Disentanglement in Variational Autoencoders

Emile Mathieu [*1] **Tom Rainforth** [*1] **N. Siddharth** [*2] **Yee Whye Teh** [1]

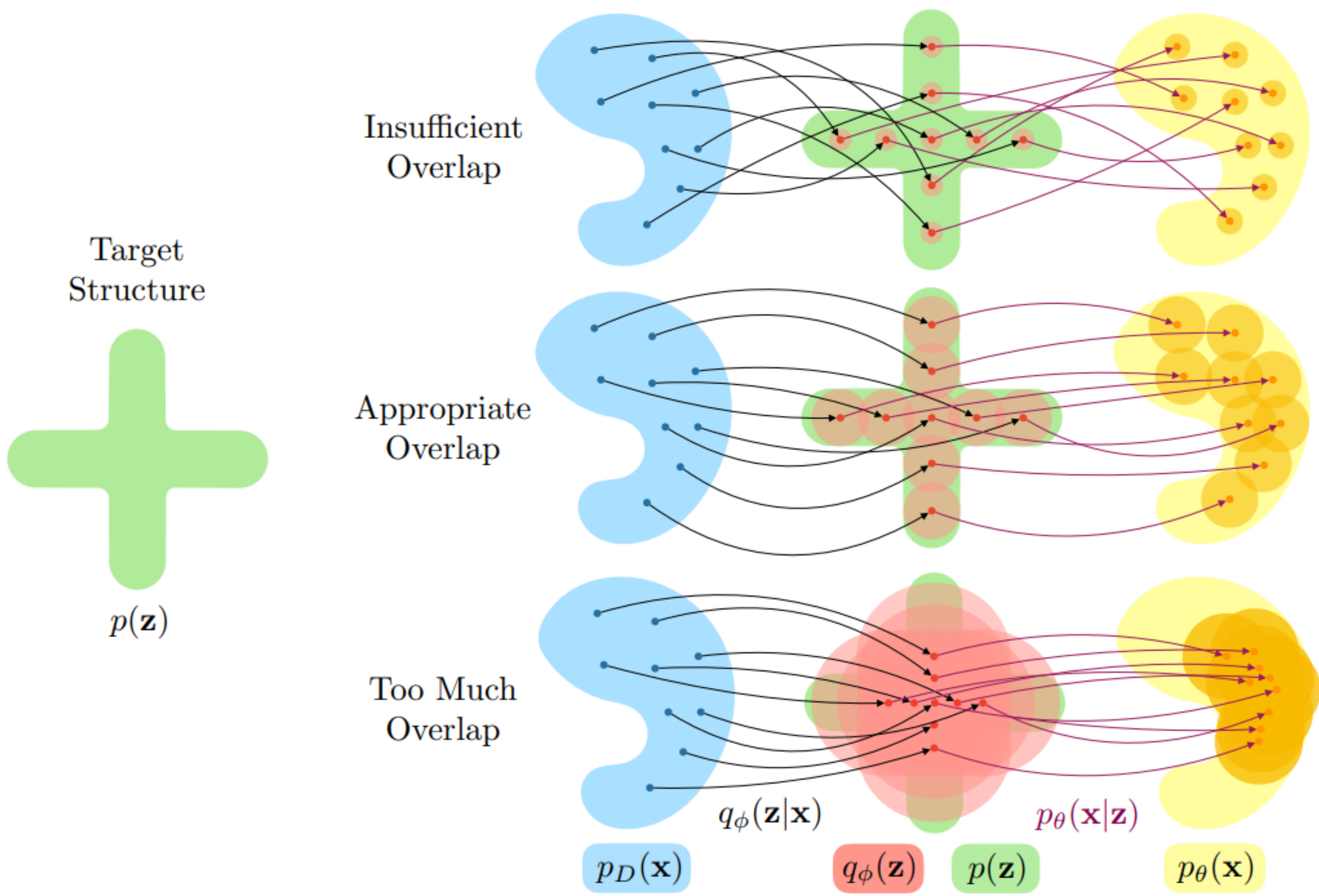
Proceedings of the 36^{th} *International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019.

Disentanglement ?



How to Do it ?

a) The latent encodings of the data having an appropriate level of overlap.



ELBO

$$\mathcal{L}_\beta(\boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \beta\,\mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})\,\|\,p(\boldsymbol{z})).$$

Theorem 1. *The β-VAE target $\mathcal{L}_\beta(\boldsymbol{x})$ can be interpreted in terms of the standard ELBO, $\mathcal{L}(\boldsymbol{x}; \pi_{\theta,\beta}, q_\phi)$, for an adjusted target $\pi_{\theta,\beta}(\boldsymbol{x},\boldsymbol{z}) \triangleq p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) f_\beta(\boldsymbol{z})$ with annealed prior $f_\beta(\boldsymbol{z}) \triangleq p(\boldsymbol{z})^\beta / F_\beta$ as*

$$\mathcal{L}_\beta(\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x}; \pi_{\theta,\beta}, q_\phi) + (\beta - 1) H_{q_\phi} + \log F_\beta \qquad (3)$$

where $F_\beta \triangleq \int_{\boldsymbol{z}} p(\boldsymbol{z})^\beta d\boldsymbol{z}$ is constant given β, and H_{q_ϕ} is the entropy of $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$.

b) the aggregate encoding of the data conforming to a desired structure, represented through the prior.

Theorem 2. *If* $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; 0, \sigma I)$ *and* $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \mu_\phi(\boldsymbol{x}), S_\phi(\boldsymbol{x}))$, *then for all rotation matrices* R,

$$\mathcal{L}_\beta(\boldsymbol{x}; \theta, \phi) = \mathcal{L}_\beta(\boldsymbol{x}; \theta^\dagger(R), \phi^\dagger(R)) \tag{6}$$

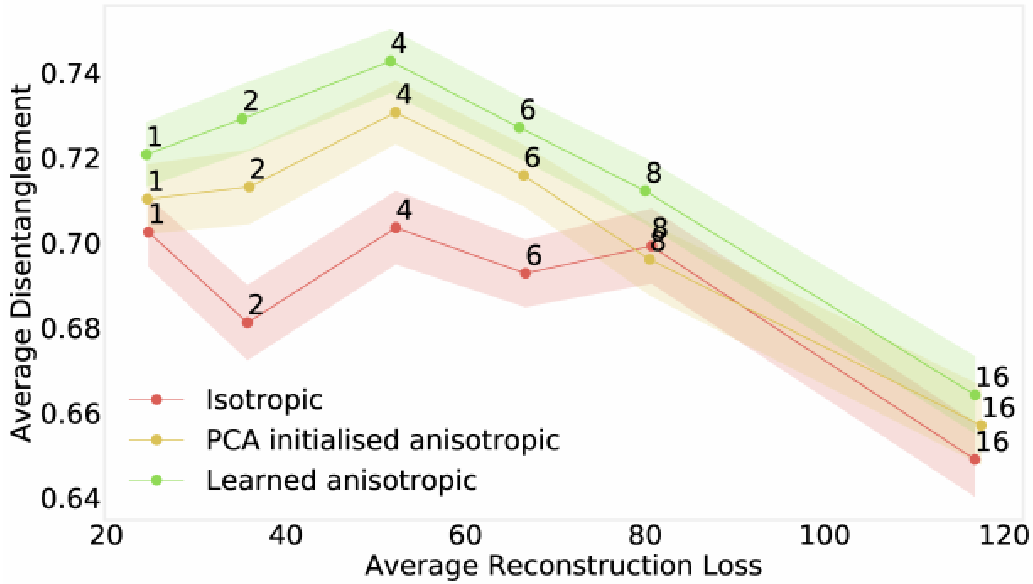
where $\theta^\dagger(R)$ *and* $\phi^\dagger(R)$ *are transformed networks such that*

$$p_{\theta^\dagger}(\boldsymbol{x} \mid \boldsymbol{z}) = p_\theta(\boldsymbol{x} \mid R^T \boldsymbol{z}),$$
$$q_{\phi^\dagger}(\boldsymbol{z} | \boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z}; R\mu_\phi(\boldsymbol{x}), RS_\phi(\boldsymbol{x})R^T\right).$$

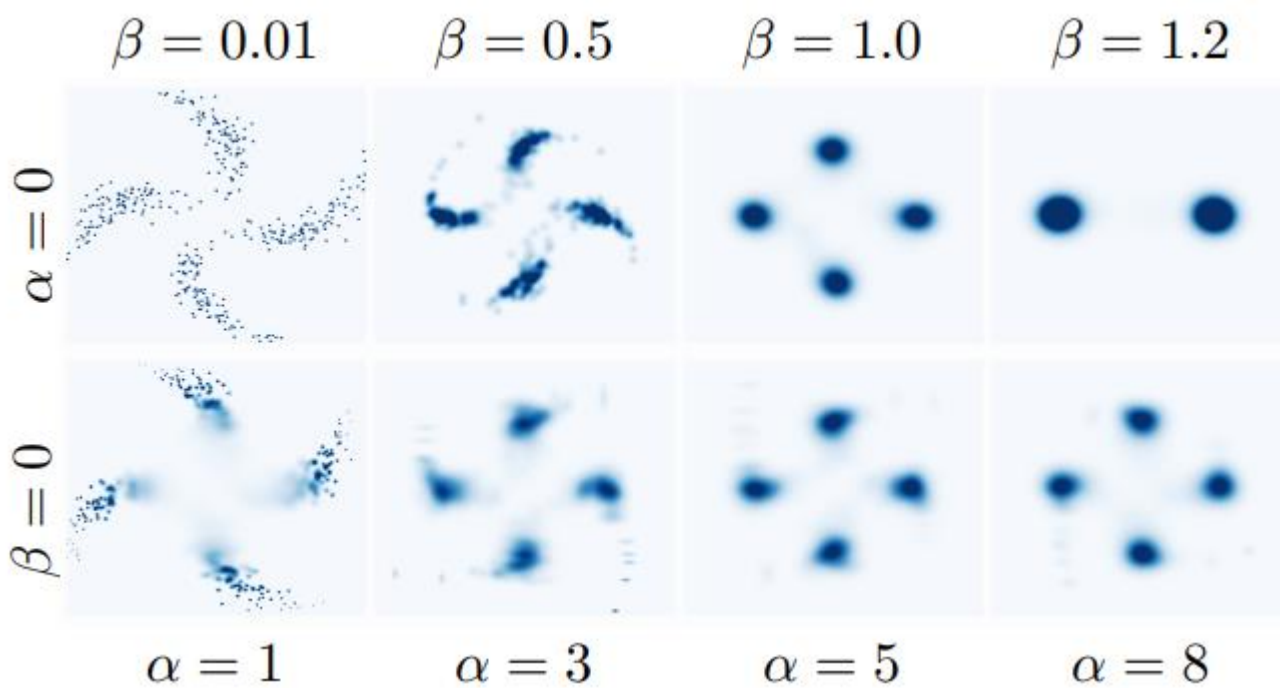
Add a regularization term

$$\mathcal{L}_{\alpha,\beta}(\boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] - \beta \ \mathrm{KL}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \,\|\, p(\boldsymbol{z})) - \boxed{\alpha\, \mathbb{D}(q_\phi(\boldsymbol{z}), p(\boldsymbol{z}))}$$

Prior for Axis-Aligned Disentanglement



Clustered Prior



Prior for Sparsity

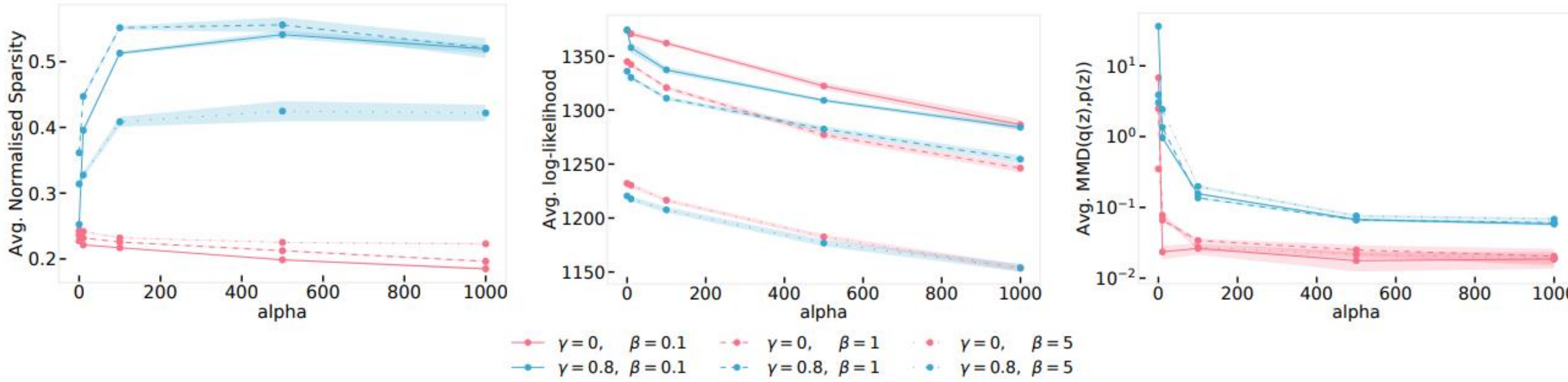
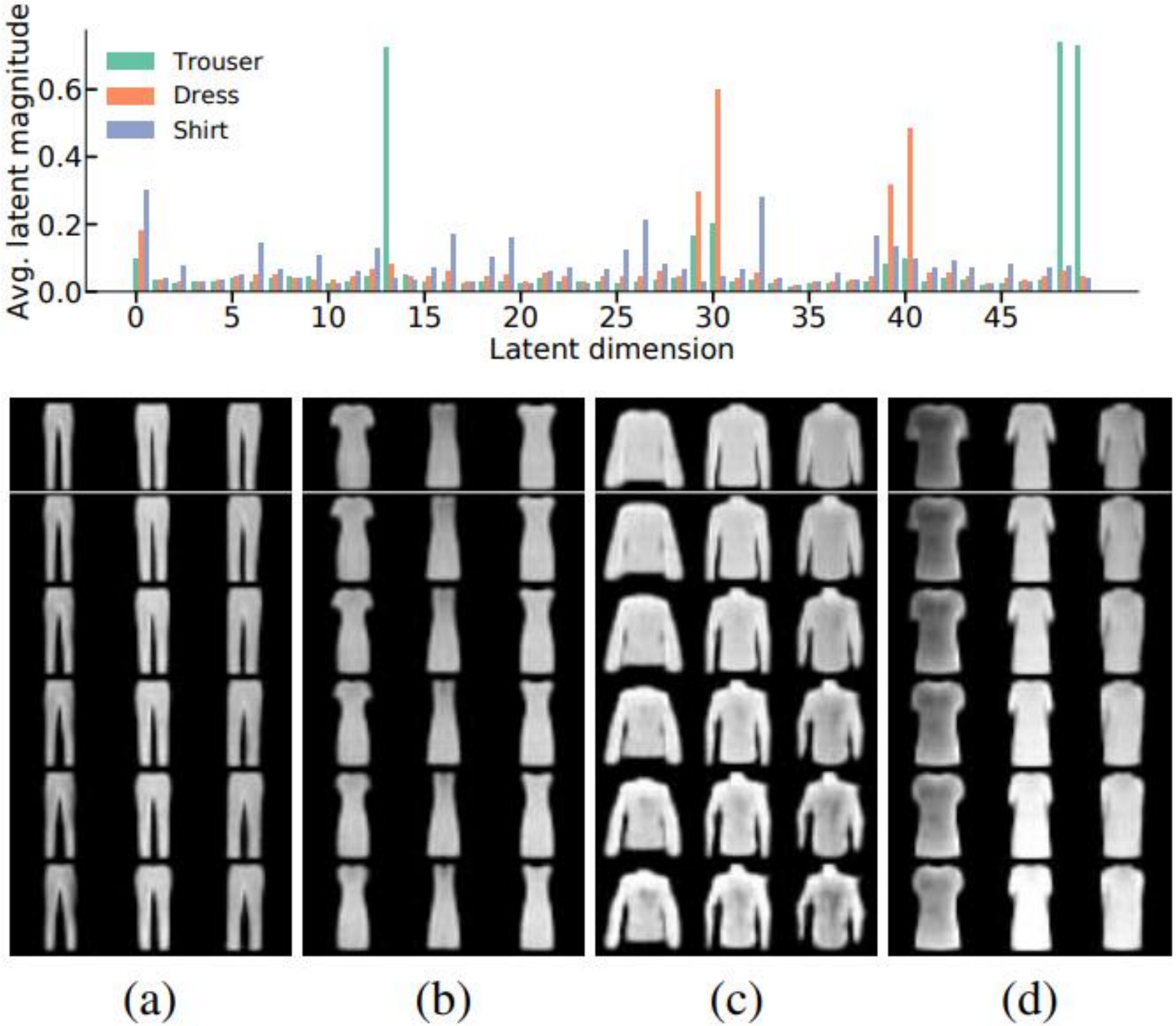


Figure 4. [Left] Sparsity vs regularisation strength α (c.f. (7), high better). [Center] Average reconstruction log-likelihood $\mathbb{E}_{p_D(\boldsymbol{x})}[\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]]$ vs α (higher better). [Right] Divergence (MMD) vs α (lower better). Note here that the different values of γ represent regularizations to different distributions, with regularization to a Gaussian (i.e. $\gamma = 0$) much easier to achieve than the sparse prior, hence the lower divergence. Shaded areas represent ± 2 standard errors in the mean estimate calculated using 8 separately trained networks. See Appendix B for full experimental details.

$$p(\boldsymbol{z}) = \prod_d \, (1-\gamma)\, \mathcal{N}(z_d; 0, 1) + \gamma\, \mathcal{N}(z_d; 0, \sigma_0^2)$$



Trouser
Dress
Shirt
Avg. latent magnitude
0.0
0.2
0.4
0.6
0 5 10 15 20 25 30 35 40 45
Latent dimension
(a)
(b)
(c)
(d)

Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations

Francesco Locatello [1,2] Stefan Bauer [2] Mario Lucic [3] Gunnar Rätsch [1] Sylvain Gelly [3] Bernhard Schölkopf [2]
Olivier Bachem [3]

(1) Be explicit about the role of inductive biases and (implicit) supervision

(2) Investigate concrete benefits of enforcing disentanglement of the learned representations

(3) Consider a reproducible experimental setup covering several data sets

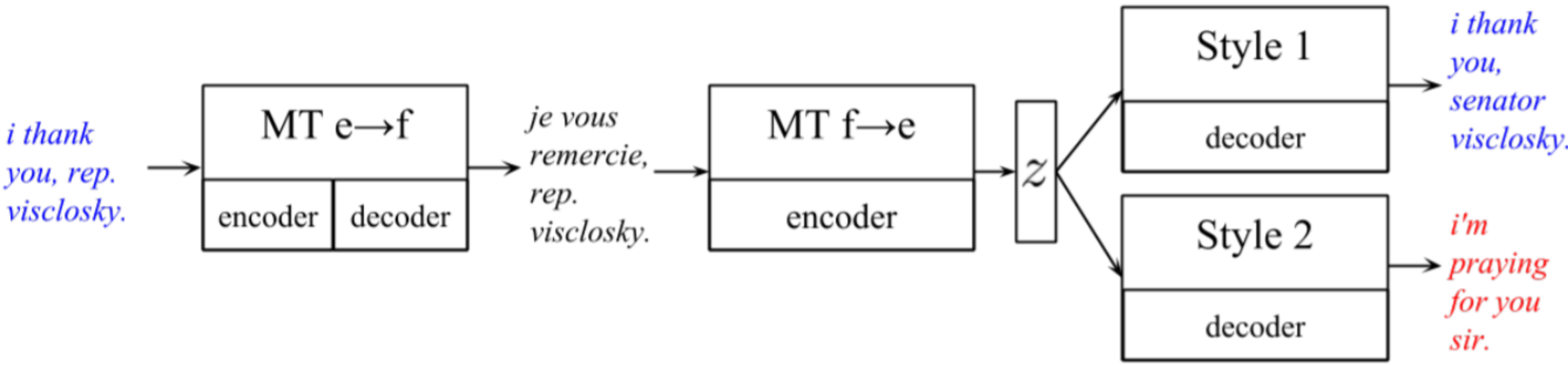
Fin

Fin

Style Transfer Through Back-Translation

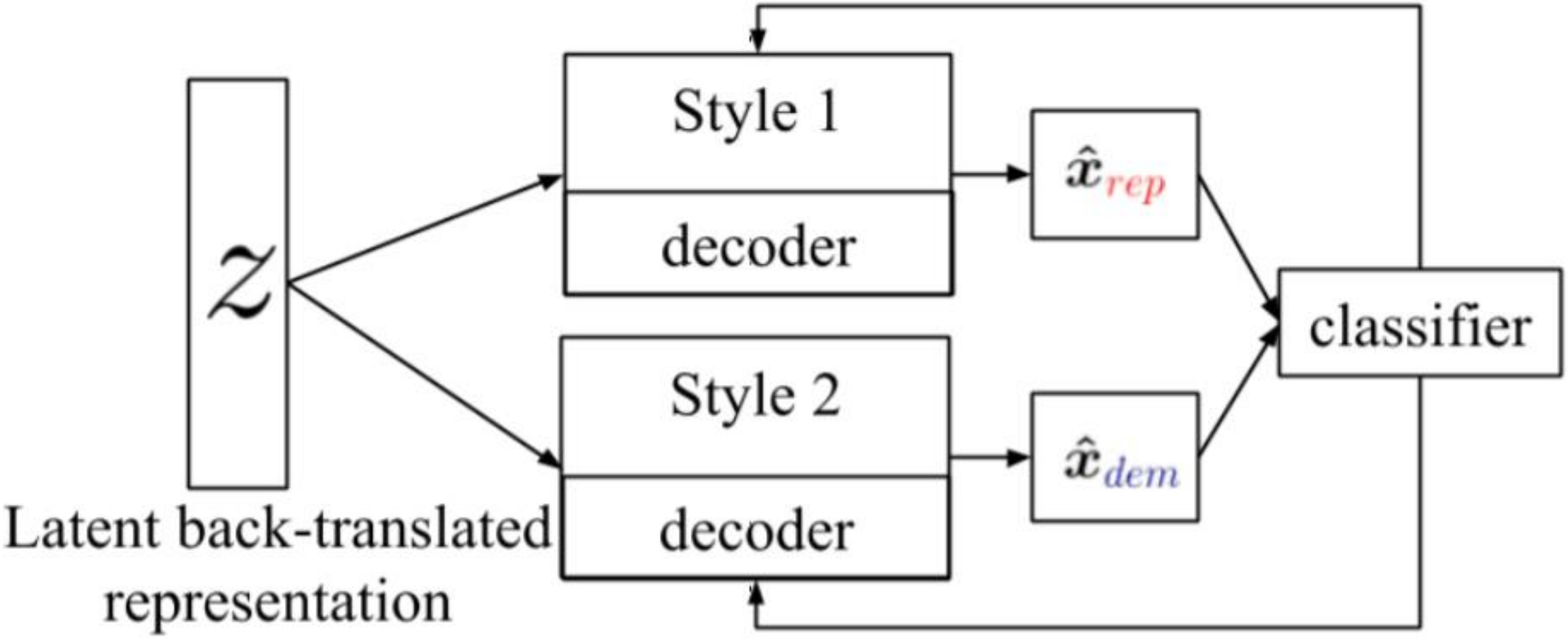
Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, Alan W Black
Carnegie Mellon University, Pittsburgh, PA, USA
{sprabhum,ytsvetko,rsalakhu,awb}@cs.cmu.edu

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 866–876
Melbourne, Australia, July 15 - 20, 2018.



Goal for latent variable z:
(1) represents the meaning of the input sentence grounded in back-translation
(2) weakens the style attributes of author's traits.

Prior work has shown that the process of translating a sentence from a source language to a target language retains the meaning of the sentence but does not preserve the stylistic features related to the author's traits (Rabinovich et al., 2016).



$$\mathcal{L}_{class}(\boldsymbol{\theta}_C) = \mathbb{E}_{\boldsymbol{X}}[\log q_C(\boldsymbol{s}|\boldsymbol{x})].$$

$$\mathcal{L}_{recon}(\boldsymbol{\theta}_G; \boldsymbol{x}) = \mathbb{E}_{q_E(\boldsymbol{z}|\boldsymbol{x})}[\log p_{gen}(\boldsymbol{x}|\boldsymbol{z})]$$

$$\min_{\theta_{gen}} \mathcal{L}_{gen} \quad = \quad \mathcal{L}_{recon} + \lambda_c \mathcal{L}_{class}$$

Experiment result

Experiment	CAE	BST
Gender	**60.40**	57.04
Political slant	75.82	**88.01**
Sentiment	80.43	**87.22**

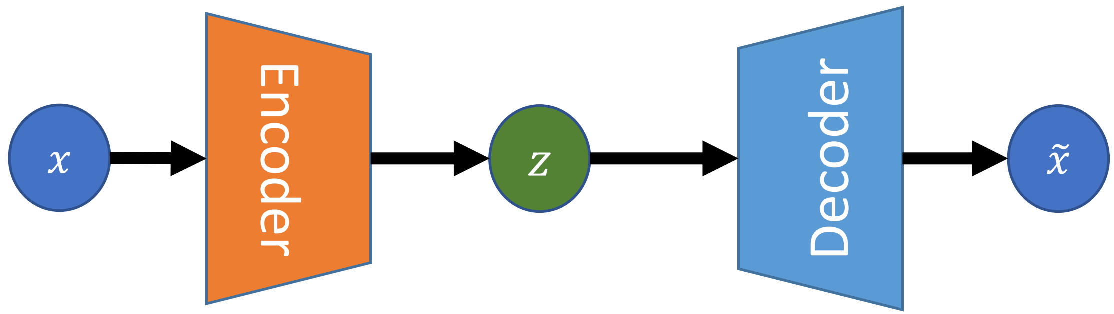
Table 4: Accuracy of the style transfer in generated sentences.

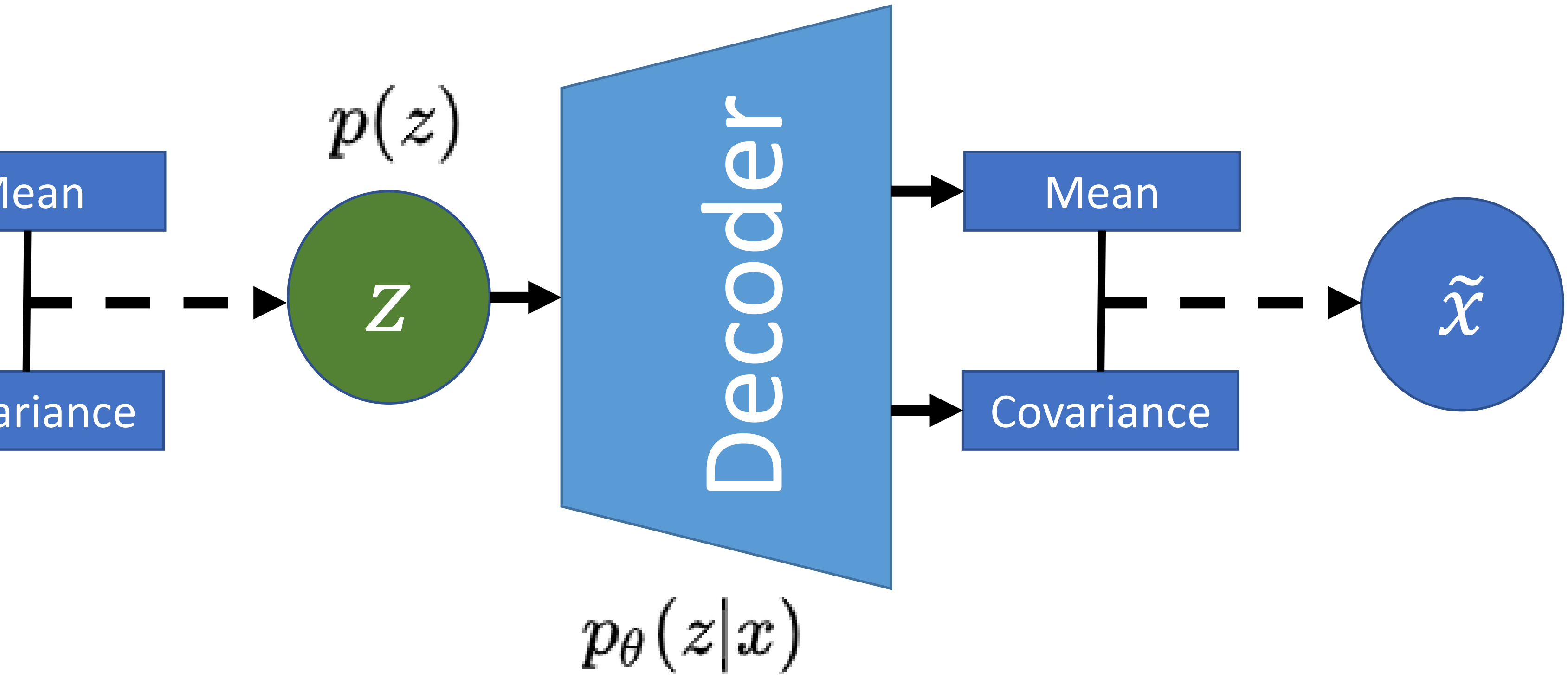
Experiment	CAE	No Pref.	BST
Gender	15.23	41.36	**43.41**
Political slant	14.55	**45.90**	39.55
Sentiment	35.91	**40.91**	23.18

Table 5: Human preference for meaning preservation in percentages.

Experiment	CAE	BST
Gender	2.42	**2.81**
Political slant	2.79	**2.87**
Sentiment	3.09	**3.18**
Overall	2.70	**2.91**
Overall Short	3.05	**3.11**
Overall Long	2.18	**2.62**

Fluency of the generated sentences.


x
Encoder
z
Decoder
$\tilde{x}$


$p(z)$
Mean
Variance
z
Decoder
Mean
Covariance
$\tilde{x}$
$p_\theta(z|x)$