Make Interpretable Discrete Representation via Vector Quantized-Variational AutoEncoder

Anonymous ACL submission

Abstract

VAE, unsupervised. discrete representations, loss to distance vector quantization, reversible generator.

1 Introduction

[1]

2 Model

2.1 VAE

$$\log q_\Phi(z|x^i) = \log \mathcal{N}(z; \mu^{(i)}, \sigma^{2(i)} I) \tag{1}$$

$$L(\theta, \phi; x^{(i)}) \simeq \frac{1}{2}\sum_{j=1}^{J}\left(1 + \log((\sigma_j^{(i)})^2) - (\mu_j^{(i)})^2 - (\sigma_j^{(i)})^2\right) + \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(x^{(i)}|z^{(i,l)})$$
$$\text{where} \quad z^{(i,l)} = \mu^{(i)} + \sigma^{(i)} \odot \epsilon^{(l)} \quad \text{and} \quad \epsilon^{(l)} \sim \mathcal{N}(0, I) \tag{2}$$

Semi-supervised

$$\log\ p_\theta(\mathbf{x}) = -\log \int p_\theta(\mathbf{z}) p_\theta(\mathbf{x}|\mathbf{z}) d\mathbf{z}$$
$$\geq\ \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathbf{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})).$$

2.2 Toward Controlled Generation of Text

2.2.1 Generator

$$\begin{aligned}\hat{\boldsymbol{x}} \sim G(\boldsymbol{z}, \boldsymbol{c}) &= p_G(\hat{\boldsymbol{x}}|\boldsymbol{z}, \boldsymbol{c}) \\ &= \prod\nolimits_t p(\hat{x}_t|\hat{\boldsymbol{x}}^{<t}, \boldsymbol{z}, \boldsymbol{c}),\end{aligned} \tag{3}$$

$$\hat{x}_t \sim \mathbf{softmax}(\boldsymbol{o}_t/\tau), \tag{4}$$

$$\boldsymbol{z} \sim E(\boldsymbol{x}) = q_E(\boldsymbol{z}|\boldsymbol{x}). \tag{5}$$

$$\begin{aligned}\mathcal{L}_{\textbf{VAE}}(\boldsymbol{\theta}_G, \boldsymbol{\theta}_E; \boldsymbol{x}) = &-\mathbf{KL}(q_E(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z})) \\ &+ \mathbb{E}_{q_E(\boldsymbol{z}|\boldsymbol{x})q_D(\boldsymbol{c}|\boldsymbol{x})}\left[\log p_G(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})\right],\end{aligned} \tag{6}$$

$$D(\boldsymbol{x}) = q_D(\boldsymbol{c}|\boldsymbol{x}). \tag{7}$$

$$\mathcal{L}_{\textbf{Attr},c}(\boldsymbol{\theta}_G) = \mathbb{E}_{p(\boldsymbol{z})p(\boldsymbol{c})}\left[\log q_D(\boldsymbol{c}|\widetilde{G}_\tau(\boldsymbol{z}, \boldsymbol{c}))\right]. \tag{8}$$

$$\mathcal{L}_{\textbf{Attr},z}(\boldsymbol{\theta}_G) = \mathbb{E}_{p(\boldsymbol{z})p(\boldsymbol{c})}\left[\log q_E(\boldsymbol{z}|\widetilde{G}_\tau(\boldsymbol{z}, \boldsymbol{c}))\right]. \tag{9}$$

$$\min\nolimits_{\boldsymbol{\theta}_G} \mathcal{L}_G = \mathcal{L}_{\textbf{VAE}} + \lambda_c \mathcal{L}_{\textbf{Attr},c} + \lambda_z \mathcal{L}_{\textbf{Attr},z}, \tag{10}$$

2.2.2 Discriminator

$$\mathcal{L}_s(\boldsymbol{\theta}_D) = \mathbb{E}_{\mathcal{X}_L}\left[\log q_D(\boldsymbol{c}_L|\boldsymbol{x}_L)\right]. \tag{11}$$

$$\mathcal{L}_u(\boldsymbol{\theta}_D) = \mathbb{E}_{p_G(\hat{\boldsymbol{x}}|\boldsymbol{z},\boldsymbol{c})p(\boldsymbol{z})p(\boldsymbol{c})}\big[\log q_D(\boldsymbol{c}|\hat{\boldsymbol{x}}) + \beta\mathcal{H}(q_D(\boldsymbol{c}'|\hat{\boldsymbol{x}}))\big], \tag{12}$$

$$\min\nolimits_{\boldsymbol{\theta}_D} \mathcal{L}_D = \mathcal{L}_s + \lambda_u \mathcal{L}_u, \tag{13}$$

2.3 Neural Discrete Representation Learning

$$q(z = k|x) = \begin{cases} 1 & \textbf{for k = argmin}_j \|z_e(x) - e_j\|_2\textbf{,} \\ 0 & \textbf{otherwise} \end{cases}, \tag{14}$$

$$z_q(x) = e_k, \quad \textbf{where} \quad k = \textbf{argmin}_j \|z_e(x) - e_j\|_2 \tag{15}$$

$$L = \log p(x|z_q(x)) + \|\textbf{sg}[z_e(x)] - e\|_2^2 + \beta\|z_e(x) - \textbf{sg}[e]\|_2^2, \tag{16}$$

[1] https://github.com/. –Code demo here?

Algorithm 1 Controlled Generation of Text

Require: A large corpus of unlabeled sentences $\mathcal{X} = \{\boldsymbol{x}\}$
A few sentence attribute labels $\mathcal{X}_L = \{(\boldsymbol{x}_L, \boldsymbol{c}_L)\}$
Parameters: $\lambda_c, \lambda_z, \lambda_u, \beta$ – balancing parameters

1: Initialize the base VAE by minimizing Eq.(6) on $\mathcal{X}$ with $\boldsymbol{c}$ sampled from prior $p(\boldsymbol{c})$
2: **repeat**
3: Train the discriminator D by Eq.(13)
4: Train the generator G and the encoder E by Eq.(10) and minimizing Eq.(6), respectively.
5: **until** convergence

Ensure: Sentence generator G conditioned on disentangled representation $(\boldsymbol{z}, \boldsymbol{c})$

2.4 Soft-to-Hard Vector Quantization

$$\phi(z) := \textbf{softmax}(-\sigma[\|z-c_1\|^2, \ldots, \|z-c_L\|^2]) \in \mathbb{R}^L, \quad (17)$$

$$\lim_{\sigma \to \infty} \phi_j(z) = \begin{cases} 1 & \textbf{if } j = \textbf{arg min}_{j' \in [L]} \|z - c_{j'}\| \\ 0 & \textbf{otherwise} \end{cases} \quad (18)$$

$$\tilde{Q}(z) := \sum_{j=1}^{L} c_j \phi_i(z) = C\phi(\mathbf{z}), \quad (19)$$

$$\begin{aligned} \hat{Z} = D(E(Z)) &= [\hat{Q}(z^{(1)}), \cdots, \hat{Q}(z^{(m)})] \\ &= [\hat{\phi}(\mathbf{z^{(1)}}), \cdots, \hat{\phi}(\mathbf{z^{(m)}})]. \end{aligned}$$

......

2.5 Wasserstein AutoEncoder

Theorem 1 *For any function $G \colon \mathcal{Z} \to \mathcal{X}$ we have*

$$\inf_{\Gamma \in \mathcal{P}(X \sim P_X, Y \sim P_G)} \mathbb{E}_{(X,Y)\sim\Gamma} \left[c(X, Y)\right] =$$

where Q_Z is the marginal distribution of Z when $X \sim P_X$ and $Z \sim Q(Z|X)$.

2.6 Visualization for Discrete

2.7 Gumbel-softmax?

2.8 RevNet

use fewer samples (update 2 times once)

2.9 Contributions

- propose RVQVAE which
- do experiments on
- compare with soft-to-hard, extension to Wasswerstein like WAE
- Visualization for discrete
- Reversible??

3 Experiments

generation(VQVAE, small), attribute accurancy(), Samples

3.1 sentiment

IMDB, SST

3.2 question type

TREC

3.3 data augmentation, semi-supervised

3.4 compare to S-VAE

3.5 language model?

3.6 synthetic: LSTM?

3.7 prior

3.8 language model(WK2) & inputless + (latent=3*3) + Sample + condition generate(semi-supervise) + visualization + prior(something like style transfer)

3.9 soft-to-hard? how to map? (use model?) distance measurement?

3.10 probability? Wasserstein?

3.11 compare with soft-to-hard, Wasserstein !!

3.12 unsupervised conditional together

3.13 revertible generator(sequence)

3.14 discussion

compare with soft-to-hard

extend to Wasserstein(WAE)

combine reversible generator(fewer samples/epochs to convergence?)

References

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *CoRR* abs/1511.06349. http://arxiv.org/abs/1511.06349.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Controllable text generation. *CoRR* abs/1703.00955. http://arxiv.org/abs/1703.00955.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. *CoRR* abs/1711.00937. http://arxiv.org/abs/1711.00937.

Model	**Standard**		**Inputless Decoder**	
	validation ppl	test ppl	validation ppl	test ppl
RNNLM(with dropout=0.2)	122.26	118.49	682.76	642.71
RVAE(w/o dropout)	152.13	144.47	171.53	157.83
VQVAE(w/o dropout)	148.94	141.50	128.72	118.12

Table 1: Language Model Results on PTB.

Model	**Standard**		**Inputless Decoder**	
	validation ppl	test ppl	validation ppl	test ppl
RNNLM(with dropout=0.2)	128.43	122.80	993.59	934.34
RVAE(w/o dropout)	229.47	207.51	88.35??	79.48??
VQVAE(w/o dropout)	132.17	118.90	208.78	189.48

Table 2: Language Model Results on WK2. Need check?or omit?

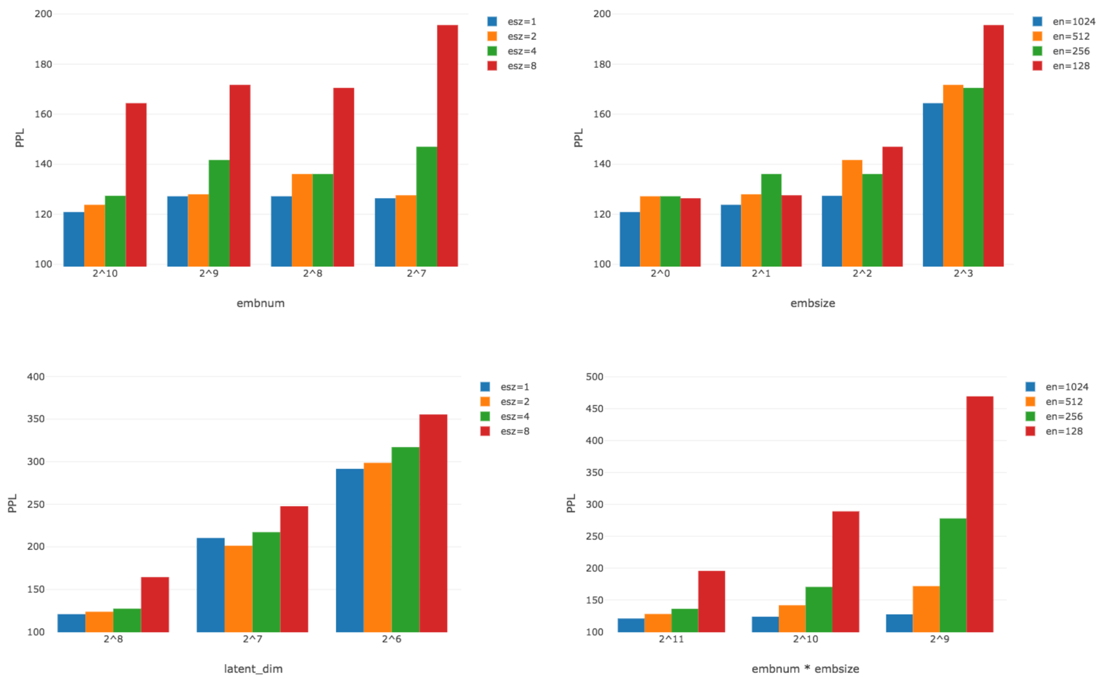


Figure 1: Latent comparisions.

latentdim = latent1 * latent2 * embsize, embnum

Table 3: Different latent size. Maybe figure is better?

Samples(Various) (Inputless)

Table 4: Samples.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. 2017. Improved variational autoencoders for text modeling using dilated convolutions. *CoRR* abs/1702.08139. http://arxiv.org/abs/1702.08139.

A appendix

(van den Oord et al., 2017) (Bowman et al., 2015) (Yang et al., 2017) (Hu et al., 2017) (?)

single	the food was good but the service was horrible.	1
all in order	the food was good , but the service was terrible .	1
permutations-rand	came here for the first time last night .	1
random	food was good , service was a little slow ,	1
reduce any	food was very good , service was fast and friendly.	1
absolutely single	1	11

Table 5: Samples via different ways.

VISUAL RESULTS (unsupervised)

Figure 2: Visualizations of learned latent representations.

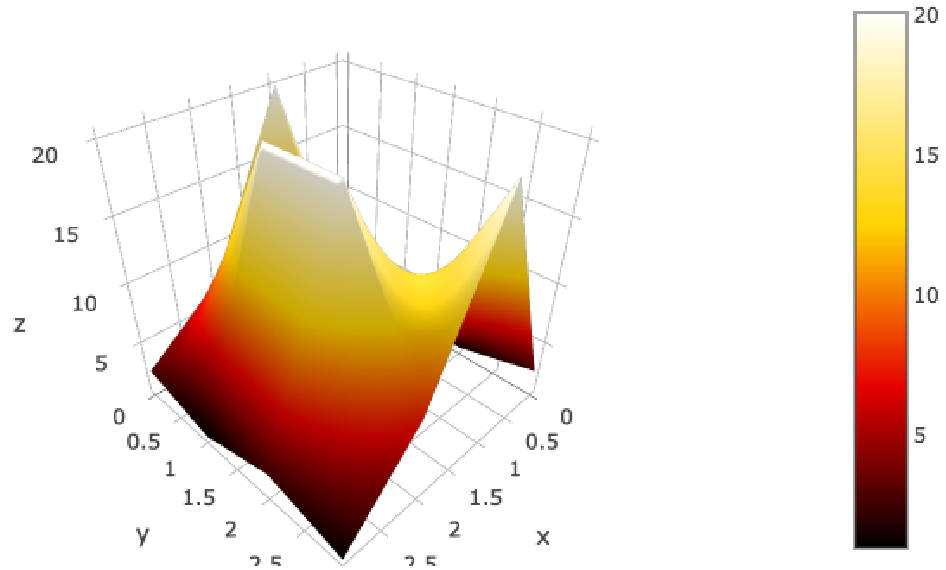


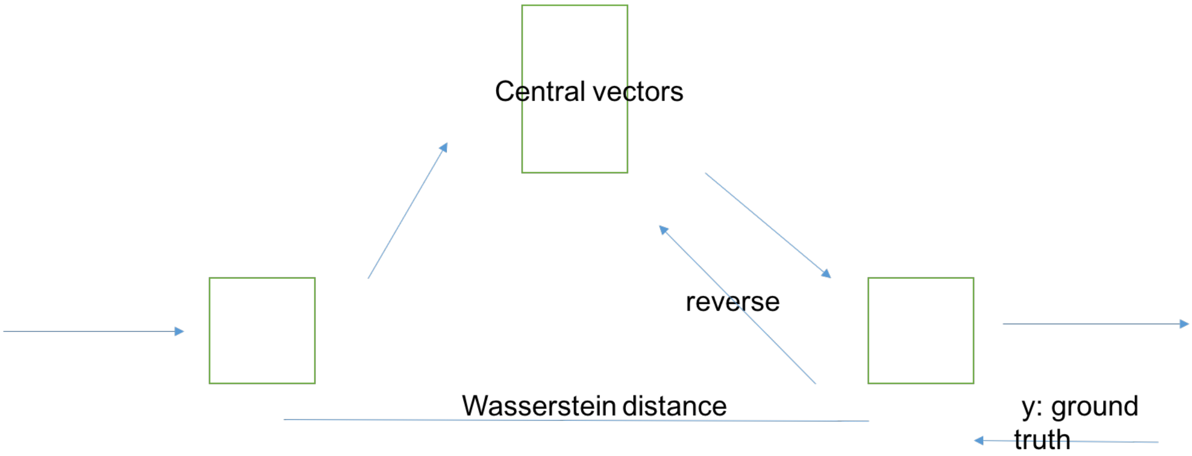
Figure 3: Latent visualizationss.

Model	**Length**		**Diversity**	
	bos	rand	bos	rand
single	3.82	6.86	1	11
all in order	2.0	4.43	1	11
permutations-rand	3.82?	6.84?	1	11
random	4.57	6.73	1	11
single absolutely	1	11	1	11

Table 6: sample length comparisions.(capability of store information) NOT NEED BEAM-SEARCH?

Semi-supervised Results

Table 7: Semi-supervised Results.



R-VQVAE ARCHITECTURE: 1 VAE + 2 DISCRETE REPRESENTATIONS

Figure 4: R-VQVAE

PRIOR COMPARISION

Figure 5: Prior comparision

Prior	
i thought the movie was too bland and too much	this was one of the outstanding thrillers of the last decade
i guess the movie is too bland and too much	this is one of the outstanding thrillers of the all time
i guess the film will have been too bland	this will be one of the great thrillers of the all time

Table 8: Each triple of sentences is [origin, VQVAE, VQVAE on prior].