# Controlled Text Generation
# &
# Persona-based Dialogue

NLP&AI LAB

장윤나

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Contents

1. Controlled Generation
   1. Introduction
   2. On Variational Learning of Controllable Representations for Test without Supervision
   3. CoCon – Content Conditioner
2. Persona-based Dialogue
   1. Method & Experiments

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# 1. Controlled Text Generation

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# 1. Controlled Text Generation

## Image Manipulation

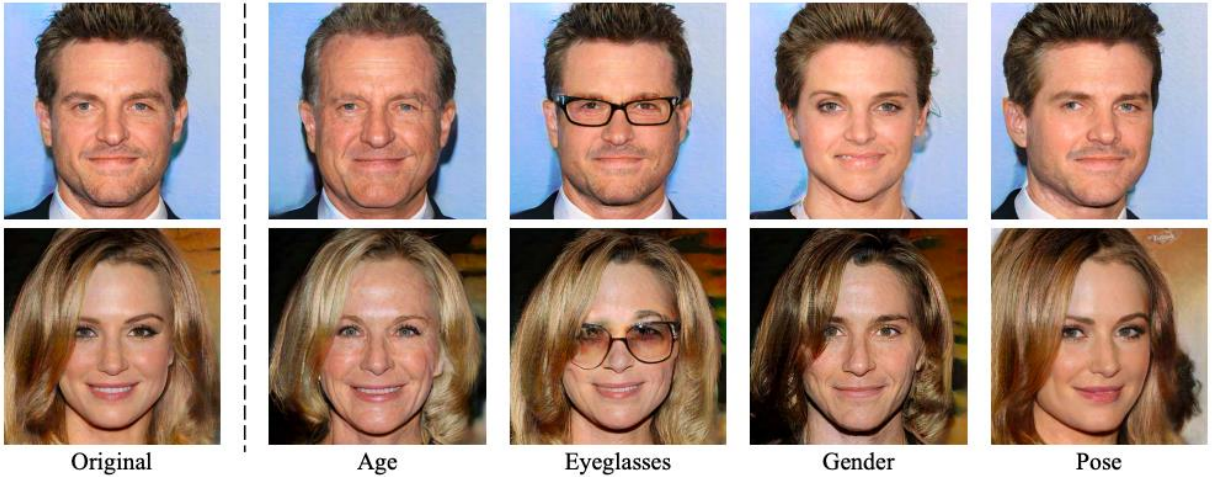



Figure 1: Manipulating various facial attributes through varying the latent codes of a well-trained GAN model. The first column shows the original synthesis from PGGAN [21], while each of the other columns shows the results of manipulating a specific attribute.

- InterFaceGAN: age, eyeglasses, gender, pose라는 attribute에 대하여 조작

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# 1. Controlled Text Generation

## Text Manipulation

Text attribute transfer - 다른 내용은 그대로, 특정한 attribute만 바꿈

Conditioned generation - 특정한 조건에 맞게 텍스트 생성

- Sentiment(Yelp dataset, Amazon dataset)
- Spoken/written, tense
- Persona(Persona-chat, Image Chat)

| Image | Personality | Generated comment |
|---|---|---|
| | Anxious | I love cats but i always get so scared that they will scratch me. |
| | Happy | That cat looks SO happy to be outside. |
| | Vague | That's a nice cat. Or is it a lion? |
| | Dramatic | That cat looks so angry; it might claw your eyes out! |
| | Charming | Awww, sweet kitty. You are so handsome! |
| | Sentimental | The arena reminded me of my childhood. |
| | Argumentative | I dislike the way the arena has been arranged |
| | Cultured | The length of this stadium coincides rather lovely with the width. |
| | Sweet | It was such a nice day at the game. These fans are the best. |
| | Romantic | Basking at the game with my love |

Image Chat

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# 1.2 On Variational Learning of Controllable Representations for Text without Supervision

ICML 2020

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Introduction

## Controllable generation & VAE

- Controllable generation 분야의 기존 방법은 supervised learning from annotated attributes에 의존 - high cost of labelling.
- 미리 정의된 attributes에 대해서만 조작이 가능함.

- VAE 는 High-dimensional data를 학습 - interaction of many complex factors(image: lighting, pose, text: style, content)
- VAE discovers **controllable latent factors** that allow manipulation through **conditional generation** from <u>interpolated or extrapolated latent codes.</u>
- Text VAE는 이미지보다 더 좋지 않은 성능을 보였음.

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

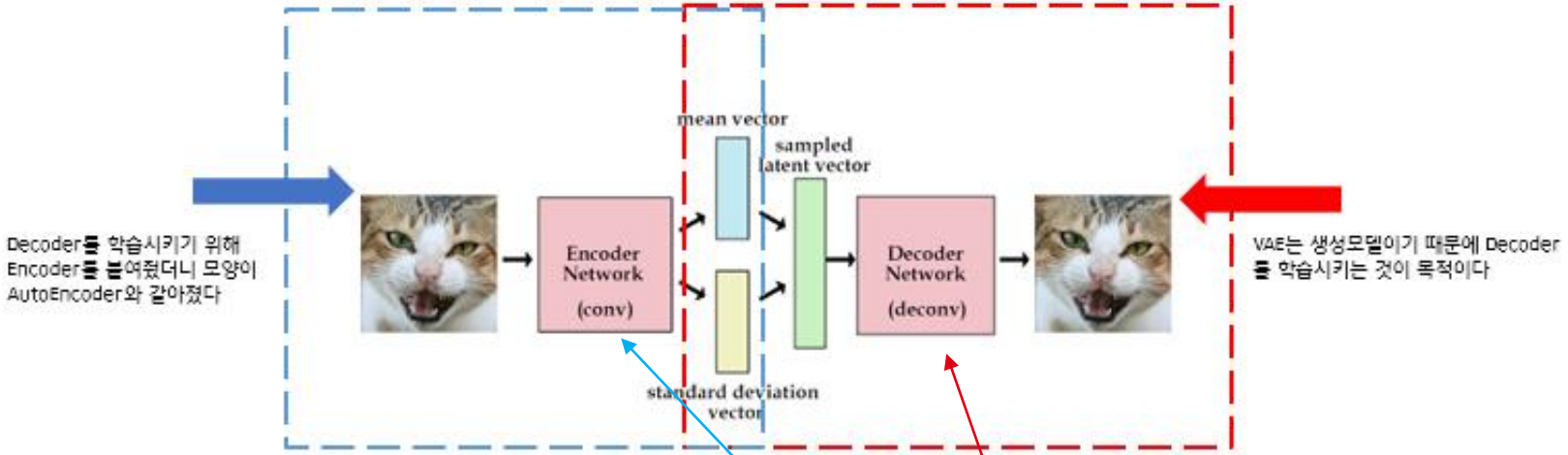
# Introduction

- 특정한 semantic variations을 위하여 latent factor를 조작하는 것은 posterior dist.의 low-density region으로 latent code를 이동시킨다는 것을 발견.
- Latent space안의 **vacant regions**은 decoding network가 해당 코드를 처리할 수 없게 만들며, generation 시 low quality result로 이어짐.
- CP-VAE(Constrained Posterior VAE) - to constrain the posterior mean to a learned probability simplex & only perform manipulation within the probability simplex.
- 2 regularizers added:
  - 1. enforces an orthogonal structure of the learned probability simplex.
  - 2. encourages this simplex to be filled without holes.

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Background: Variational Autoencoders

## VAE - generative model




$$\mathcal{L}_{\text{VAE}}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi}) = \mathbf{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})] - D_{\text{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})),$$

Decoder net은
aggregated posterior dist. q로부터
샘플된 z를 처리하는 것을 학습.

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Latent Vacancy Hypothesis

- Hypothesis: when trained on text data, the aggregated **posterior** $(q_\phi(z) = \mathrm{E}_{x\sim p_d(x)} q_\phi(z|x))$ of sequence-VAE **tend to have vacant regions of low density**, where the decoder may fail to generalize to.

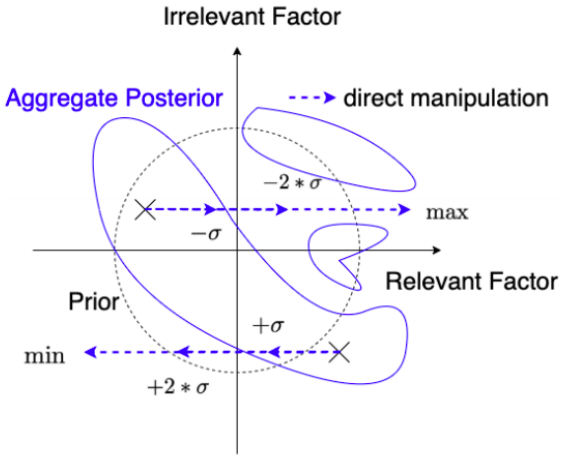



*Figure 1.* Illustration of why latent vacancy prevents effective manipulation in VAEs. The aggregated posterior shown has multiple disconnected areas and direct manipulations of the relevant factor may fall into vacant regions of low density.

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Latent Vacancy Hypothesis

## 1. Unsupervised Sentiment Manipulation

- Discovering a sentiment latent dimension & manipulating the sentiment attribute
- Yelp restaurant reviews dataset with $\beta$-VAE, $\beta$ -VAE with aggressive training and CP-VAE
- Latent space of 80-dim with a LSTM enc&dec (beta: 0.35)
- **One dim of latent code** achieves around 75% sentiment classification acc (others 50%) - **effective sentiment indicator dimension**

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Latent Vacancy Hypothesis

## 1. Unsupervised Sentiment Manipulation

- Sentiment manipulation by modifying latent dimension
- Transfer strength & content preservation

| | Example | Transfer Strength | Content Preservation | NLL Discrepancy |
|---|---|---|---|---|
| Source sentence | the pizza is offered without toppings and it 's lacking in flavor . | - | - | - |
| $\beta$-VAE w. aggr training ($\pm\sigma$) | the pizza is offered in toppings and it 's lacking in pittsburgh sauce . | Weak | Good | Small |
| $\beta$-VAE w. aggr training ($\pm 2 * \sigma$) | the pizza is more than fresh and your food is lacking in flavor | Medium | Medium | Medium |
| $\beta$-VAE w. aggr training (extremum) | the service is a great cut and the food is top notch in charlotte . | Strong | Bad | Large |
| CP-VAE (this work) | the pizza is full of spicy and it 's delicious . | Strong | Good | Small |

*Table 1.* Summary of the behaviours of $\beta$-VAE with aggressive training and our proposed CP-VAE. Detailed quantitative evaluations for transfer strength and content preservation are presented in Tab. 2.

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Latent Vacancy Hypothesis

## 2. NLL of the Codes under the Aggregated Posterior

- Increase of Negative log-likelihood (NLL) : the new codes land in regions of lower density.
- 1000 test samples' code before and after manipulation
- **Discrepancy** between the original and the manipulated codes becomes larger as we increase the magnitude of the manipulation

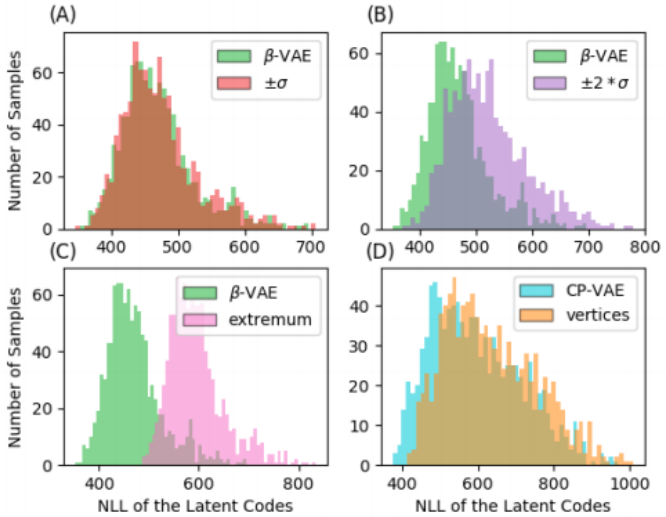



*Figure 2.* Histograms of all the test samples' negative log-likelihood (NLL) under the aggregated posterior, considering their original latent codes and manipulated ones. (A) (B) (C): three manipulation strategies for $\beta$-VAE with aggressive training; (D) CP-VAE.

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Method

## 1. Overview

1. To better match the aggregated posterior with the prior.
   - unsupervised learning에서 이러한 방법론이 성공한 적은 없음.

2. To directly **enforce** that the **aggregated posterior** itself has **no vacant region** anywhere where we would like to perform latent code manipulation.
   - **map** the **posterior Gaussian mean** to a **constrained space** (a learned probability simplex) and perform **manipulation** to be **within** this simplex

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Method

## 1. Overview

Splitting z into two parts z(1) and z(2) following prior work(Bao et al.,2019)

- z(1): **relevant factors that are dominant in the data** without an inductive bias from external signals - averaging Glove/BERT embeddings
- z(2): learns to encode the **remaining local information** that is useful for reconstructing the source sentences - LSTM encoding network

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \rightarrow q_{\boldsymbol{\phi}_1}(\boldsymbol{z}^{(1)}|\boldsymbol{x}) q_{\boldsymbol{\phi}_2}(\boldsymbol{z}^{(2)}|\boldsymbol{x}) \quad \text{where } \boldsymbol{\phi} = \boldsymbol{\phi}_1 \cup \boldsymbol{\phi}_2.$$

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Method

## 2. Constraining the Posterior

- $\mathbf{z}$ 의 Gaussian Posterior Mean이 아래와 같은 구조를 갖도록 constrain
- $\boldsymbol{e}$ - relevant factors, $\boldsymbol{p}$ - $\boldsymbol{i}$ 번째 element의 비율.
- $\mathbf{z}$ 의 Gaussian posterior의 평균이 K-dimension probability simplex ($\mathrm{R}^{\mathrm{N}}$ space - orthogonal basis vector $\boldsymbol{e}_i$, $i$=1, ⋯, K)안으로 제한됨.
- h는 MLP 인코더의 output, 이를 $\pi$ 가 constrained posterior space로 mapping.

$$\boldsymbol{\mu} = \sum_{i=1}^{K} p_i \boldsymbol{e}_i, \sum_{i=1}^{K} p_i = 1, \langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle = 0, i \neq j, K \leq N$$

$$\boldsymbol{\mu} = \pi(\boldsymbol{h}) = \boldsymbol{E} \cdot \mathrm{softmax}(\boldsymbol{W}\boldsymbol{h} + \boldsymbol{b}),$$

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Method

## 2. Constraining the Posterior

- Mapping function만으로는 posterior collapse(학습중 latent code를 모델이 무시하는 VAE의 문제)가 발생할 수 있음.
- Regularizer – Enforce **orthogonality** in the learnt basis vectors.

$$D_{\mathrm{KL}}(q_{\boldsymbol{\phi}_1}(\boldsymbol{z}^{(1)}|\boldsymbol{x})\|p(\boldsymbol{z}^{(1)})) = \frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\mu} + \frac{1}{2}\left(\boldsymbol{\sigma}^\top\boldsymbol{\sigma} - \log\boldsymbol{\sigma}^\top\boldsymbol{\sigma} - 1\right).$$

$$\boldsymbol{\mu}^\top\boldsymbol{\mu} = \left(\sum_i p_i\boldsymbol{e}_i\right)^\top\left(\sum_i p_i\boldsymbol{e}_i\right) = \sum_i p_i^2\boldsymbol{e}_i^\top\boldsymbol{e}_i.$$

$$\mathcal{L}_{\mathrm{REG}}(\boldsymbol{x};\boldsymbol{\phi}_1) = \|\boldsymbol{E}^\top\boldsymbol{E} - \alpha\boldsymbol{I}\|, \quad \text{When } \mathcal{L}_{\mathrm{REG}} = 0, \boldsymbol{e}_i^\top\boldsymbol{e}_i = \alpha.$$

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Method

## 3. Filling the Constrained Space

- 특정 공간으로 posterior를 제한하는 것이 그 공간을 완전히 채운다는 것을 의미하지 않음.
- 확률 분포가 해당 공간을 최대한 cover할 수 있도록 해야함.
- MLP output h(raw latent code)를 restructured latent code $\boldsymbol{\mu}$ 와 비슷해지게 , negative sample 과는 달라지게 학습해야 함.
- M개의 negative sample을 뽑아 $\boldsymbol{\mu}_i^{(-)}$ 로 만듦, margin loss로 학습.

$$\mathcal{L}_{\text{S-REC}}(\boldsymbol{x};\boldsymbol{\phi}_1) = \mathbb{E}_{\boldsymbol{z}^{(1)}\sim q_{\boldsymbol{\phi}_1}(\boldsymbol{z}^{(1)}|\boldsymbol{x})}\left[\frac{1}{m}\sum_{i=1}^{m}\max(0, 1-\boldsymbol{h}\cdot\boldsymbol{\mu}+\boldsymbol{h}\cdot\boldsymbol{\mu}_i^{(-)})\right].$$

$$\mathcal{L}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi}) = \mathcal{L}_{\text{VAE}} + \mathcal{L}_{\text{REG}} + \mathcal{L}_{\text{S-REC}}.$$

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Method

## 3. Filling the Constrained Space

- 특정 공간으로 posterior를 제한하는 것이 그 공간을 완전히 채운다는 것을 의미하지 않음.
- 확률 분포가 해당 공간을 최대한 cover할 수 있도록 해야함.

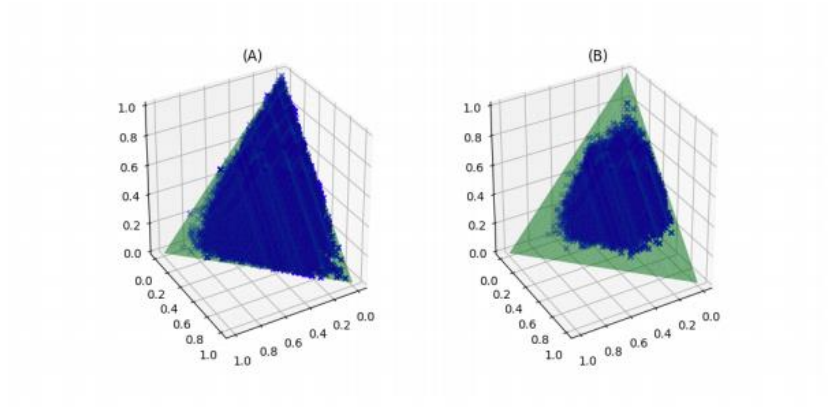



*Figure 4.* Visualization of all training samples in the probability simplex: (A) With $\mathcal{L}_{\text{S-REC}}$ ;(B) Without $\mathcal{L}_{\text{S-REC}}$.

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Experiments

## 1. Comparisons with Unsupervised Baselines

*Table 2.* Comparisons with unsupervised baselines on Yelp dataset.

| Model | Accuracy (AC) ↑ | BLEU (BL) ↑ |
|---|---|---|
| $\beta$-VAE ($\pm\sigma$) | $50.98 \pm 2.89$ | $4.02 \pm 0.77$ |
| $\beta$-VAE ($\pm 2 * \sigma$) | $78.44 \pm 4.84$ | $1.49 \pm 0.29$ |
| $\beta$-VAE (extremum) | $98.18 \pm 1.56$ | $0.56 \pm 0.40$ |
| $\beta$-VAE w. aggr training ($\pm\sigma$) | $26.76 \pm 6.44$ | $27.91 \pm 4.39$ |
| $\beta$-VAE w. aggr training ($\pm 2 * \sigma$) | $57.46 \pm 14.47$ | $11.73 \pm 6.74$ |
| $\beta$-VAE w. aggr training (extremum) | $88.08 \pm 14.95$ | $4.57 \pm 4.63$ |
| CP-VAE w. GloVe | $60.22 \pm 4.57$ | $33.69 \pm 1.47$ |
| without $\mathcal{L}_{\text{REG}}$ | $10.82 \pm 0.91$ | $33.27 \pm 2.84$ |
| without $\mathcal{L}_{\text{S-REC}}$ | $12.28 \pm 3.69$ | $49.34 \pm 2.65$ |

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Experiments

## 2. Comparisons to Supervised Approaches on Text Style Transfer

*Table 3.* Comparisons with supervised approaches on Yelp and Amazon dataset.

| | | | Yelp | | | | Amazon | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | Supervised | GPT-2 | AC ↑ | BL ↑ | GL ↑ | PL ↓ | AC ↑ | BL ↑ | GL ↑ | PL ↓ |
| Source | - | - | 1.8 | 100.0 | 8.4 | 26.6 | 16.3 | 100.0 | 22.8 | 34.5 |
| Human | - | - | 70.1 | 25.3 | 100.0 | 63.7 | 41.2 | 45.7 | 100.0 | 68.6 |
| CA | ✓ | × | 74.0 | 20.7 | 6.0 | 103.6 | **75.5** | 0.0 | 0.0 | **39.3** |
| SE | ✓ | × | 8.2 | **67.4** | 6.9 | 65.4 | 40.2 | 0.4 | 0.0 | 125.0 |
| MD | ✓ | × | 49.5 | 40.1 | 6.6 | 164.1 | 70.1 | 0.3 | 0.0 | 138.8 |
| D&R | ✓ | × | **88.1** | 36.7 | 7.9 | 85.5 | 49.2 | 0.6 | 0.0 | 46.3 |
| **CP-G** | × | × | 66.7 | 35.5 | 7.5 | 67.8 | 60.1 | **35.4** | **11.5** | 109.1 |
| **CP-B** | × | × | 55.4 | 48.4 | **9.6** | **47.6** | 40.0 | **39.7** | **12.7** | 97.3 |
| B-GST | ✓ | ✓ | 85.6 | 45.2 | 12.7 | 49.6 | 55.2 | 52.3 | 18.1 | 48.2 |

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Experiments

## 2. Comparisons to Supervised Approaches on Text Style Transfer

*Table 4.* Samples of generated sentences. SRC is the input sentence.

| **Yelp** | *Positive to Negative* | *Negative to Positive* |
|---|---|---|
| SRC | this place is super yummy ! | but it probably sucks too ! |
| B-GST | this place is super bad ! | but it tastes great too ! |
| **CP-G** | this place is super slow and watered down . | but it 's truly fun and insanely delicious . |
| **CP-B** | this place is super greasy and gross ! | but it 's probably wonderful when you ! |
| **Amazon** | *Positive to Negative* | *Negative to Positive* |
| SRC | because it s made of cast iron , scorching is minimized . | they are cheerios, afterall, and we love the original kind . |
| B-GST | because it s cheaply made of cast iron , is useless . | they are sturdy, afterall, sturdy and we love the original . |
| **CP-G** | because it s made of cast iron , vomitting . | they are ripe, tastier , and we love them . |
| **CP-B** | because it s made of cast iron , limp . | they are divine, fluffier , and we love them . |

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Conclusion

- Variational autoencoder(VAE): 연속적인 latent space에서 interpolation, extrapolation을 통해 이미지안의 manifold를 학습.
- Text(discrete data)에서 unsupervised learning으로 controllable manipulation이 가능한 latent space를 학습할 수 있는지는 불확실.
- Sequence VAE(trained on text)의 aggregated posterior latent space안에 빈(vacant) 공간이 많기 때문에 manipulated latent code로 decode를 잘 하지 못한다는 것을 발견.
- Posterior mean을 probability simplex로 제한하여 해결.
- The first success in unsupervised learning of controllable representations for text.

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# 1.3 CoCon - Content Conditioner

ICLR 2021

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# CoCon

## Motivation

- 일반적인 auto-regressive 방식의 text generation

$$p(x_t, \ldots, x_l | x_1, \ldots, x_{t-1}) = \prod_{i=t}^{l} p(x_i | x_1, \ldots, x_{i-1}). \tag{1}$$

- a를 target attribute 혹은 control code로 가지는 text generation:

  Controlled at a global attribute level (sentiment/topic)

$$p(x_t, \ldots, x_l | x_1, \ldots, x_{t-1}) = \prod_{i=1}^{l} p(x_i | \mathbf{a}, \{x_1, \ldots, x_{i-1}\}), \tag{2}$$

이전 시퀀스, target attribute와 잘 어울리는 possible candidates가 너무 많기 때문에 sampling과정에서 전혀 다른 content가 들어갈 수 있다는 문제점 발생.

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# CoCon

## Motivation

- Input target content c를 이용하여 더 높은 수준의 컨트롤 가능

$$p(x_t, \ldots, x_l | x_1, \ldots, x_{t-1}) = \prod_{i=1}^{l} p(x_i | \mathbf{c}, \{x_1, \ldots, x_{i-1}\}), \tag{3}$$

- c는 target text sequence

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# CoCon

## Model Architecture

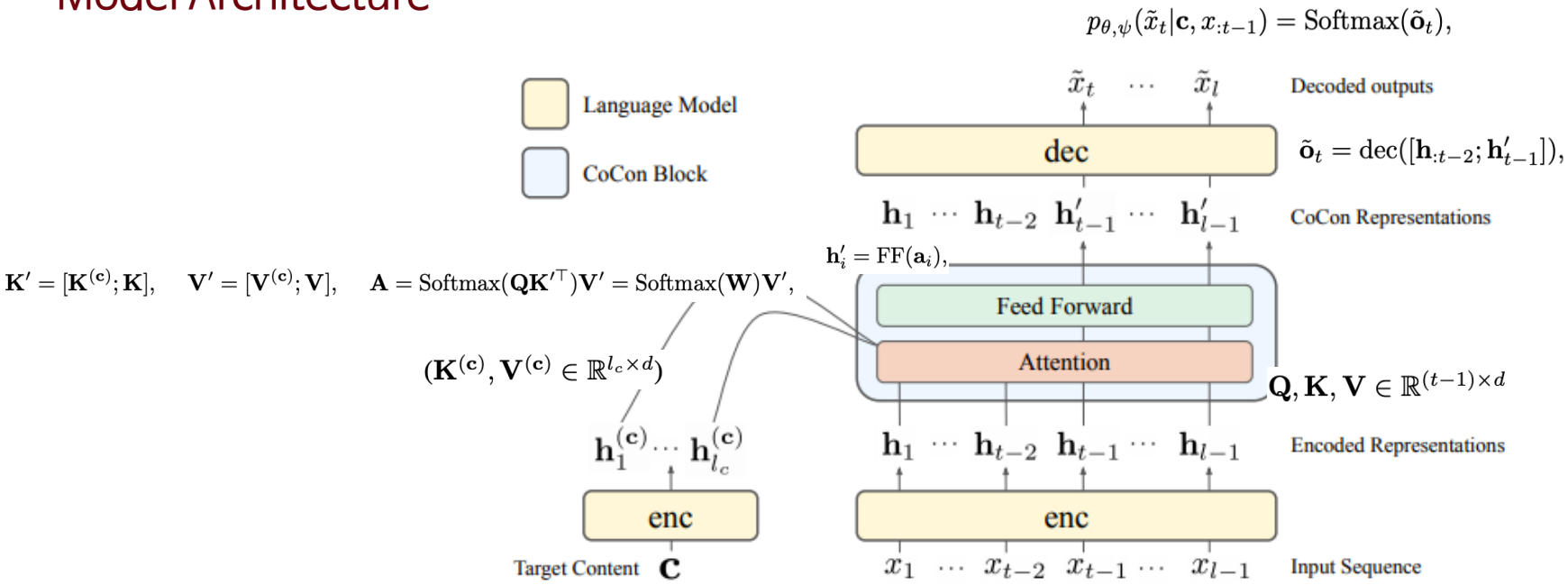



Figure 1: Model architecture of Content-Conditioner (CoCon).

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# CoCon

## Model Architecture

- Multiple Content input: **N content inputs** for a single generation

$$\mathbf{K}' = [\mathbf{K}^{(\mathbf{c}^1)} \dots \mathbf{K}^{(\mathbf{c}^N)}; \mathbf{K}], \quad \mathbf{V}' = [\mathbf{V}^{(\mathbf{c}^1)} \dots \mathbf{V}^{(\mathbf{c}^N)}; \mathbf{V}], \quad \mathbf{A} = \mathrm{Softmax}(\mathbf{Q}\mathbf{K}'^{\top})\mathbf{V}'.$$

- Strength of Content Conditioning: target content c의 strength 조정 - $\tau_{\text{content}}$

  content attention weights에 positive bias term으로 적용하여 $\mathbf{V}^{(\mathbf{c})}$ 의 영향력 조정

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# CoCon

## Self-Supervised Learning

- text sequence를 two contiguous segments로 나눔.

$$\mathbf{x} = \{x_1, \ldots, x_{t-1}, x_t, \ldots, x_l\}$$

$$\mathbf{x}^a = \{x_1, \ldots, x_{t-1}\} \qquad \mathbf{x}^b = \{x_t, \ldots, x_l\}$$

- 실제로는 $\mathrm{X}^a$가 주어졌을 때 $\mathrm{X}^b$ 가 될 수 있는 것은 무수히 많음 – $\mathrm{X}^b$ 에 대한 정보 없이, $\mathrm{X}^a$만으로 full X 를 reconstruct하기 어려움.

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# CoCon

## Self-Supervised Learning

- Self-reconstruction Loss

$$\mathcal{L}_{\text{self}} = -\sum_{i=t}^{l} \log p_{\theta,\psi}\left(x_i | (\mathbf{c} = \mathbf{x}^b), \{x_1, \ldots, x_{i-1}\}\right)$$

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# CoCon

## Self-Supervised Learning

- Null Content Loss

$\mathrm{X}^a$가 $\mathrm{X}^b$ 와 상관 없이 fluently generated 되도록 하기 위함, target content c가 null.

$$\mathcal{L}_{\text{null}} = -\sum_{i=t}^{l} \log p_{\theta,\psi}\left(x_i | (\mathbf{c} = \varnothing), \{x_1, \ldots, x_{i-1}\}\right).$$

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# CoCon

## Self-Supervised Learning

- Cycle Reconstruction Loss

target content c가 initial prompt text p가 다른 source에서 온 경우를 위해, 서로 다른 training sample $(\mathbf{x}, \mathbf{x}')$ 과 two CoCon forward step을 이용.

Content c가 잘 반영된 fluent sequence $[\mathbf{p}; \mathbf{y}]$ , where $\mathbf{y} = f_{\theta,\psi}(\mathbf{c}, \mathbf{p})$, is the output of the auto-regressive generation.

First step: $\mathbf{y}_{\mathbf{x},\mathbf{x}'} = f_{\theta,\psi}((\mathbf{c} = \mathbf{x}^b), (\mathbf{p} = \mathbf{x}'^a))$,

Second step: $\mathbf{y}_{\text{cycle}} = f_{\theta,\psi}((\mathbf{c} = \mathbf{y}_{\mathbf{x},\mathbf{x}'}), (\mathbf{p} = \mathbf{x}^a))$,

$\mathbf{x} = [\mathbf{x}^a; \mathbf{x}^b]$ and $\mathbf{x}' = [\mathbf{x}'^a; \mathbf{x}'^b]$

$$\mathcal{L}_{\text{cycle}} = -\sum_{i=t}^{l} \log p_{\theta,\psi}\left(\mathbf{y}_{\text{cycle}} = \mathbf{x}^b | (\mathbf{c} = \mathbf{y}_{\mathbf{x},\mathbf{x}'}), (\mathbf{p} = \mathbf{x}^a)\right).$$

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# CoCon

## Self-Supervised Learning

- Adversarial Loss
- Output texts' representation enc(y) to match those of training samples enc(x)

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{\mathbf{x}}[\log f_{\text{disc}}(\text{enc}(\mathbf{x}))] + \mathbb{E}_{\mathbf{y}}[\log(1 - f_{\text{disc}}(\text{enc}(\mathbf{y}))],$$

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# CoCon

## Self-Supervised Learning

- Full training

$$\theta^* = \arg\min_{\theta}(\lambda_{\text{self}}\mathcal{L}_{\text{self}} + \lambda_{\text{null}}\mathcal{L}_{\text{null}} + \lambda_{\text{cycle}}\mathcal{L}_{\text{cycle}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}),$$

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Experiments

Table 1: CoCon samples with single or *multiple* content targets, given same prompt text (underlined), exhibiting control over generations. More samples are in the Appendix (Table 14 and 15).

| |
|---|
| Target Topic: SCIENCE, Content Input (**c**): **Scientist** |
| The movie historian Thomas A. Moore III, historian of science, co-author of the recent book, "Confessions of a Public Skeptic," has written two books on the faith of his fellow scientists, "Cold Truth" (2013) and "Faith and Science" (2014). As Moore described it in an interview with SPACE.com: But in truth.. |
| Target Sentiment: Positive, Content Input (**c**): **is perfect** |
| The movie EXPRESS is perfect. Perfection is perfect! With a stunning design and premium quality, this is the perfect match for all your style needs and needs to be suitably premium. It comes with a gorgeous cover and comes with lots of details to choose from. You get to choose your exact style and accessories.. |
| Content Input ($\mathbf{c}^1$): **officials predict there could be 5,800 submerged**<br>+ Target Topic: SCIENCE, Content Input ($\mathbf{c}^2$): **Scientist**<br>+ Target Sentiment: Positive, Content Input ($\mathbf{c}^3$): **is perfect** |
| The movie makers speculate there's a perfect match. Expectations there could be up to 500 kilograms of clay could be thrown onto the surface of the ocean. The BBC reported that it could have taken up to a year and a half to add clay to the ocean floor, though experts believe it could be done within several days.. |

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Experiments

## Setup

- GPT-2 medium 345M model - 24 layers of transformer block
- encoder: first 7 blocks, decoder: remaining 17 blocks
- CoCon block: single GPT-2 transformer block with 1024-dim
- split at position $8^{th}$~$12^{th}$ of 30-BPE long segments

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Experiments

## Generation

Table 2: Content similarity and quality of generated content-conditioned samples. BLEU, NIST and METEOR values are reported in scale of ($\times 10^{-2}$).

| Model | BLEU-4 (↑ better) | NIST-4 (↑ better) | METEOR (↑ better) | Perplexity (↓ better) | Dist-1 (↑ better) | Dist-2 (↑ better) | Dist-3 (↑ better) |
|---|---|---|---|---|---|---|---|
| GPT-2 | 0.22 | 7.09 | 6.14 | 105.7 | 0.057 | 0.49 | 0.82 |
| CoCon | 2.76 | 22.9 | 21.5 | 70.8 | 0.048 | 0.39 | 0.70 |
| ∟ w/o $\mathcal{L}_{\text{cycle}}$ | 3.30 | 25.1 | 23.9 | 150.8 | 0.050 | 0.42 | 0.74 |
| ∟ w/o $\mathcal{L}_{\text{null}}$ | 4.44 | **28.3** | 26.8 | 73.2 | 0.046 | 0.37 | 0.68 |
| ∟ w/o $\mathcal{L}_{\text{adv}}$ | **4.47** | 28.2 | **27.2** | **68.7** | 0.047 | 0.38 | 0.69 |
| CoCon-Webtext | 2.90 | 24.6 | 23.0 | 112.5 | 0.054 | 0.44 | 0.74 |
| Prompt-Content | – | – | – | 442.2 | – | – | – |
| Webtext | – | – | – | 185.8 | – | – | – |

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Experiments

## Topic Relevance

Table 3: Evaluation of topic-controlled generations. Topic accuracy report ratio of samples that were classified as their target topic.

| Model | Topic % (↑ better) | Perplexity (↓ better) | Dist-1 (↑ better) | Dist-2 (↑ better) | Dist-3 (↑ better) |
|---|---|---|---|---|---|
| GPT-2 | 22.5 | 84.7 | 0.23 | 0.74 | 0.91 |
| PPLM | 42.5 | **32.4** | 0.15 | 0.54 | 0.78 |
| CTRL | 86.7 | 60.5 | 0.14 | 0.56 | 0.77 |
| CoCon | **90.4** | 52.4 | 0.17 | 0.60 | 0.86 |

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Experiments

## Sentiment Control

Table 4: Evaluation of sentiment-controlled generations. Sentiment accuracy report ratio of samples that were classified as their target sentiment.

| Model | Sentiment % (↑ better) | Perplexity (↓ better) | Dist-1 (↑ better) | Dist-2 (↑ better) | Dist-3 (↑ better) |
|---|---|---|---|---|---|
| GPT-2 | 50.0 | 101.2 | 0.38 | 0.82 | 0.92 |
| PPLM | 68.9 | **35.5** | 0.24 | 0.63 | 0.82 |
| CTRL | 81.1 | 44.1 | 0.21 | 0.62 | 0.80 |
| CoCon | **98.9** | 50.3 | 0.20 | 0.61 | 0.80 |

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Experiments

Table 11: Generated CoCon samples with varying degree of topic content-conditioning.

| Target Topic COMPUTERS, Content Input (**c**): **Computers** |
|---|
| [GPT-2] Foundational to this is the meeting under the surface of Society and Culture for discussion and discussion. Don't use society to discuss the few, take it on an other level. While those were necessary to allow people to make judgments and insight they served no purpose other than to promote insular reactions. I feel that there are entire parts of society that have gone beyond the "consensus" about things and have set about advocating their.. |
| [$\tau_{\text{content}} = -100$] Foundational to this is that the concept of independence was something everyone wanted to have and the belief in the importance of that freedom for the future. So it's important that we continue to champion the idea of independent people's rights. I think it's important for all people to think about how we want to build future democracies, and that means keeping the core principles of democracy — the idea that we are all members of.. |
| [$\tau_{\text{content}} = -10$] Foundational to this is real-world computers — built by IBM and others. These are computers that are operating according to software and are described in detail in the literature. For example, IBM's No. 2.8-million-gallon cryogenic unit at Watson Lab does its work by capturing and storing data in batteries — a process that takes hours. The bottom line is that this is not a very useful design for.. |
| [$\tau_{\text{content}} = 0$] Foundational to this is computers, computer systems. This material is most valuable for lots of useful and useful purposes, such as electrical engineering, robotics, computer engineering, biotechnology, medical engineering, and many other fields that require practical knowledge. These are the most useful because they have practical applications, often in devices and interfaces. Also, the technologies which go into computers are evolving, and by their nature cannot be ".. |
| [$\tau_{\text{content}} = 2$] Foundational to this is computers, computers that generate computers. Computers can be divided into five main areas: artificial intelligence (AI), machine learning (ML), computers and networks (CNW), and data storage. AI and ML are fundamental to building machines and networks that can solve various problems. ML is where computers work together with a human to solve problems. In computers, systems work with a target and a target set of.. |
| [$\tau_{\text{content}} = 10$] Foundational to this is computers, computers or computers software - computers (computer) programs (program) specialised (specialised) (specialised) the (computer) computer-part (computer-part) specialised (specialised) Computer-Part computer-specialised (specialised) specialised (specialised.. |

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# 2. Persona-based Dialogue

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# 2.1 Methods & Experiments

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Persona-based Dialogue

Task & datasets

## Task Information – ConvAI2 challenge (Neurips 2018)

- 일반적인 chit-chat 모델에서 너무 일반화된 답변만을 생성하고, personality가 일정하게 나타나지 않는 문제점 존재.
- 사용자에게 어떠한 페르소나가 존재한다고 가정, 3-5개의 페르소나 문장이 각 사용자에게 있고 이를 기반으로 두 사용자가 서로의 페르소나를 알아가며 대화.
- 사용자의 프로필 정보에 기반하여 대화를 함으로써 더 적극적인 대화 모델링.

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Persona-based Dialogue

Task & datasets

## Persona-chat datasets

<Train set>
131,438 utterances
17,878 dialogs
1,155 personas

<Validation set>
7,801 utterances
1,000 dialogs
100 personas

<Test set>
6,634 utterances
1,015 dialogs
100 personas

persona

| Persona 1 | Persona 2 |
|---|---|
| I like to ski | I am an artist |
| My wife does not like me anymore | I have four children |
| I have went to Mexico 4 times this year | I recently got a cat |
| I hate Mexican food | I enjoy walking for exercise |
| I like to eat cheetos | I love watching Game of Thrones |

dialog

At least 6-8 turns

[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ? ← utterance
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.

# Q&A

감사합니다!

Copyright © 2020 Natural Language Processing & Artificial Intelligence Lab.