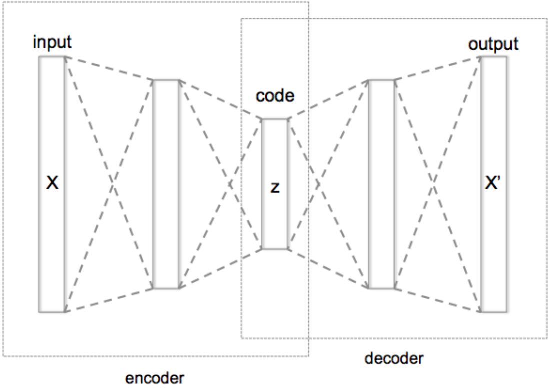
Toward Controlled Generation of Text

Hu, Yang, Liang, Salakhutdinov & Xing

ICML 2017

Motivation



[1]

[1] https://commons.wikimedia.org/wiki/File:Autoencoder_structure.png

Motivation

Distentangled Latent Representations

I love this movie $\rightarrow$

I hate this movie $\rightarrow$

0.21	0.32	0.74	0.43
0.45	0.78	0.97	0.17

Motivation

Distentangled Latent Representations

I love this movie	$\rightarrow$	0.21	0.32	0.74	0.43
I hate this movie	$\rightarrow$	0.45	0.78	0.97	0.17

$$\Downarrow$$

I love this movie	$\rightarrow$	0.68	0.12	0.33	1.00
I hate this movie	$\rightarrow$	0.68	0.12	0.33	0.00

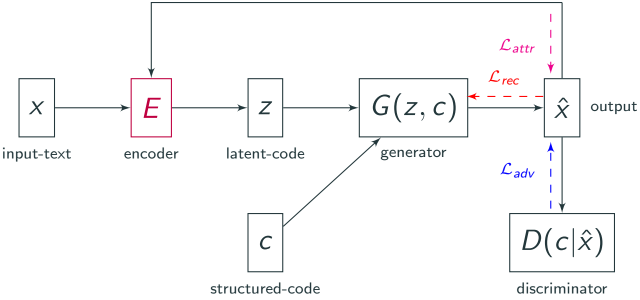
Problem Statement

Generate fake samples similar to the source distribution by conditioning their generation on a tunable set of attributes.

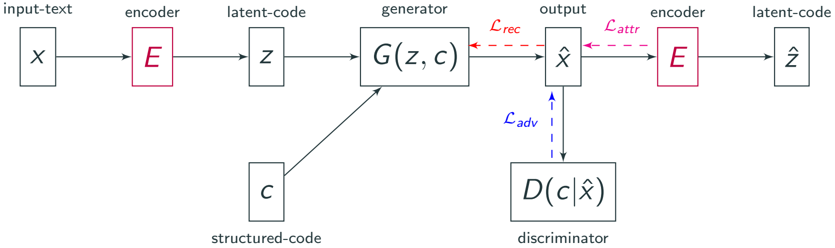
Notation

$x \Rightarrow$ **source documents**
$\hat{x} \Rightarrow$ **output documents**
$c \Rightarrow$ **structured latent code**
$z \Rightarrow$ **unstructured latent code**

$E \Rightarrow$ **encoder**, parameterized to generate z
$G \Rightarrow$ **decoder/generator**, produces $\hat{x}$ conditioned on (z, c)
$D \Rightarrow$ **discriminator**, predicts c given x or $\hat{x}$.
$\tau \Rightarrow$ **softmax temperature**, for decoder word prediction


x
input-text
E
encoder
z
latent-code
$G(z, c)$
generator
$\hat{x}$
output
c
structured-code
$D(c|\hat{x})$
discriminator
$\mathcal{L}_{attr}$
$\mathcal{L}_{rec}$
$\mathcal{L}_{adv}$

Unrolled Architecture



Optimization Objectives

Discriminator Optimization

- Maximize the likelihood of predicting the correct distribution of the structured code c given the labeled examples X_L.

$$\mathcal{L}_s(\theta_D) = -\mathbb{E}_{X_L}[log q_D(c_L|x_L)]$$

- Maximize the likelihood of predicting the correct distribution of the structured code c given the generated sentences $\hat{x}$. Also minimize the empirically observed Shannon entropy of the discriminator predictions $q_D(c'|\hat{x})$.

$$\mathcal{L}_u(\theta_D) = -\mathbb{E}_{p_G(\hat{x}|z,c)p(z)p(c)}[log q_D(c|\hat{x}) + \beta\mathcal{H}(q_D(c'|\hat{x}))]$$

Optimization Objectives

Generator Optimization

- Maximize the likelihood of predicting the original document x, given the latent spaces and the generator $G(z, c)$.

$$\begin{aligned}\mathcal{L}_{VAE}(\theta_G, \theta_E; x) &= -\mathbb{E}_{q_E(z|x)q_D(c|x)}[log p_G(x|z, c)] \\ &\quad + KL(q_E(z|x)||p(z))\end{aligned}$$

- Maximize the likelihood of generating the output documents with the correct structured code c

$$\mathcal{L}_{attr,c}(\theta_G) = -\mathbb{E}_{p(z)p(c)}[log q_D(c|\tilde{G}_\tau(z, c))]$$

Training Objectives

Additional Generator Optimization: Independency Constraint

The encoder is re-used to regenerate the latent distribution z devoid of the structured code c, from the output distribution $\tilde{G}_{\tau}(z, c)$.

$$\mathcal{L}_{attr,z}(\theta_G) = -\mathbb{E}_{p(z)p(c)}[log q_E(z|\tilde{G}_{\tau}(z, c))]$$

Training Objectives

Generator:

$$\begin{aligned}\min_{\theta_G} \mathcal{L}_G = \quad & \mathcal{L}_{VAE} && \text{(reconstruction loss)} \\ & + \lambda_c \mathcal{L}_{attr,c} && \text{(style entanglement)} \\ & + \lambda_z \mathcal{L}_{attr,z} && \text{(independency constraint)}\end{aligned}$$

Discriminator:

$$\begin{aligned}\min_{\theta_D} \mathcal{L}_D = \quad & \mathcal{L}_s && \text{(labeled example classification)} \\ & + \lambda_u \mathcal{L}_{attr,u} && \text{(synthesized example classification)}\end{aligned}$$

Algorithm

Require: A large corpus of unlabeled sentences $\mathcal{X} = \{x\}$

A few sentence attribute labels $\mathcal{X}_L = \{(x_L, c_L)\}$

Parameters: $\lambda_c, \lambda_z, \lambda_u, \beta$ – balancing parameters

Initialize the base VAE by minimizing $\mathcal{L}_{VAE}$ on $\mathcal{X}$

repeat

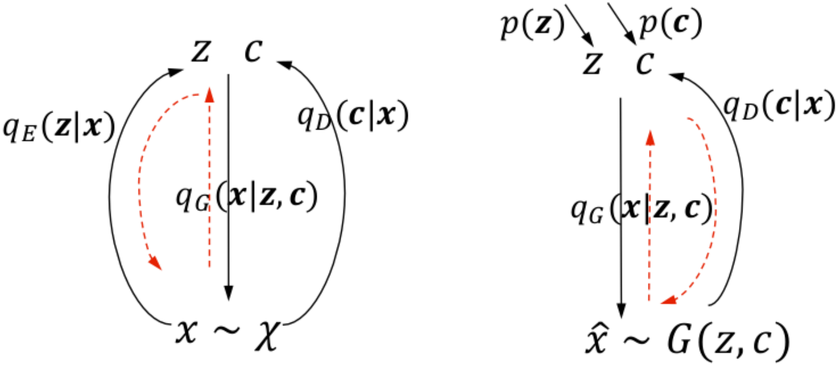
  Train the discriminator D by minimizing $\mathcal{L}_u$

  Train the generator G and the encoder E by minimizing $\mathcal{L}_G$ and $\mathcal{L}_{VAE}$, respectively.

until convergence

Ensure: Sentence generator G conditioned on disentangled representation (z, c)

Wake-Sleep Algorithm



Experimental Setup

Experimental Setup

Simple Reconstruction

- $350K$ IMDB movie reviews
- Sentence length ≤ 15
- Total sentence count $= 1.4M$
- Vocab size $= 16K$

Experimental Setup

Binary Sentiment Classification and Conditioned Generation

- **Stanford Sentiment Treebank-2**: Movie reviews
  - 2837 training examples
  - Sentence length ≤ 15
- **Sentiment Lexicon**: Words used as sentences
  - 2700 words
  - Sentence length ≤ 15
- **IMDB**: Movie reviews
  - $16K$ training examples

Experimental Setup

Tense Classification and Conditioned Generation

- TimeBank[2] Lexicon
- 5250 words and phrases labeled with one of past, present, future
- Verbs in different tenses (e.g., was, will be) as well as time expressions (e.g., in the future)

[2] http://timeml.org

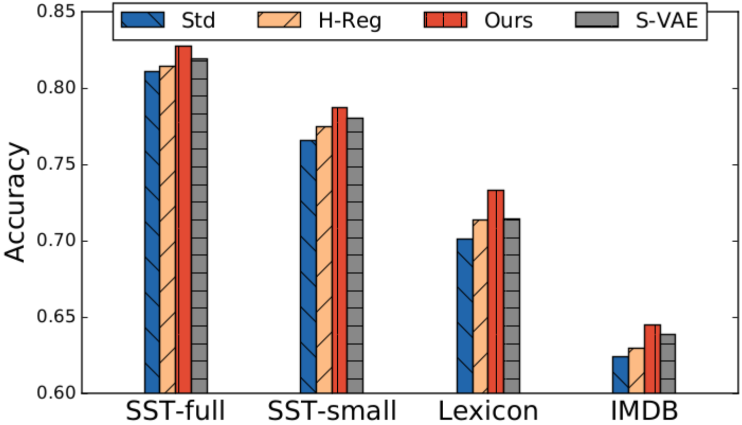
Experimental Setup

Parameters

- The generator and encoder are set as single-layer LSTM RNNs with input/hidden dimension of 300 and max decoding time-step count of 15
- Discriminators are set as Convolutional Nets
- VAE KL-Divergence weight linearly annealed from 0 to 1 during training
- Softmax temperature τ annealed from 1 to 0
- Balancing λ weights all set to 0.1, β is selected on the dev set

Results

Results


Std
H-Reg
Ours
S-VAE
0.85
0.80
0.75
0.70
0.65
0.60
Accuracy
SST-full
SST-small
Lexicon
IMDB

Without 'Independency Constraint'

the acting is bad	$\Rightarrow$	the movie is so much fun
none of this is very original	$\Rightarrow$	highly recommended viewing for its courage , and ideas
too bland	$\Rightarrow$	highly watchable
i can analyze this movie without more than three words	$\Rightarrow$	i highly recommend this film to anyone who appreciates music

Results

With ‘Independency Constraint’

the film is strictly routine	$\Rightarrow$	the film is full of imagination
after watching this movie , i felt that disappointed	$\Rightarrow$	after seeing this film , i ’m a fan
the acting is uniformly bad either	$\Rightarrow$	the performances are uniformly good
this is just awful	$\Rightarrow$	this is pure genius

Results

Varying the unstructured code z

("negative", "past") the acting was also kind of hit or miss .	**("positive", "past")** his acting was impeccable
("negative", "present") the era seems impossibly distant	**("positive", "present")** i 've always been a big fan of the smart dialogue .
("negative", "future") and that would be devastating !	**("positive", "future")** i will definitely be buying this on dvd

Commentary

- Non-parallel attribute-controlled generation
- Encoder independency constraint

- No quantifiable results for content preservation

Related Work

- Shen, Tianxiao, et al. ‘Style transfer from non-parallel text by cross-alignment.’ Advances in Neural Information Processing Systems. 2017.
- Kim, Yoon, et al. ‘Adversarially regularized autoencoders for generating discrete structures.’ arXiv preprint arXiv:1706.04223 (2017).
- Fu, Zhenxin, et al. ‘Style Transfer in Text: Exploration and Evaluation.’ arXiv preprint arXiv:1711.06861 (2017).
- Melnyk, Igor, et al. ‘Improved Neural Text Attribute Transfer with Non-parallel Data.’ arXiv preprint arXiv:1711.09395 (2017).

Questions?