Language Style Transfer

Tianxiao Shen
Different Language Styles

King James Bible

In the beginning God created the heaven and the earth.
And God saw the light, that it was good:
and God divided the light from the darkness.

Bible in basic English

At the first God made the heaven and the earth.
And God, looking on the light, saw that it was good:
and God made a division between the light and the dark.

Simplicity, formality, politeness, personal styles…
Language Style Transfer

King James Bible

In the beginning God created the heaven and the earth.
And God saw the light, that it was good:
and God divided the light from the darkness.

Bible in basic English

At the first God made the heaven and the earth.
And God, looking on the light, saw that it was good:
and God made a division between the light and the dark.

- Towards real language understanding
- Personalized chatbots, appropriately convey a message according to different social contexts…
Parallel Translation

$\mathbf{x}_1 \rightarrow h_1 \rightarrow h_2 \rightarrow h_3 \rightarrow \mathbf{y}_1$

$\mathbf{x}_2 \rightarrow \mathbf{y}_2$

$\mathbf{x}_3 \rightarrow \text{Awesome}$
To collect parallel data is very costly or even impossible in the style transfer scenario.
Non-Parallel Transfer

To be, or not to be, that is the question:
Whether 'tis nobler in the mind to suffer
The slings and arrows of outrageous fortune,
Or to take Arms against a Sea of troubles,
And by opposing end them: to die, to sleep

They’re bringing drugs, they’re bringing crime, they're rapists, and some, I assume, are good people

Obama, and all others, have been so weak, and so politically correct, that terror groups are forming and getting stronger! Shame.
Image Style Transfer

photograph + artwork → after transfer

Monet  Van Gogh  Cezanne  Ukiyo-e

[Zhu et al. 2017]
Challenges in Language Style Transfer

• Style and content interact in subtle ways
• Content must be preserved
• Discreteness
Our Approach

• Style and content interact in subtle ways
• Content must be preserved
• Discreteness

- Map between sentences and continuous latent representations
- Decompose latent representations into style and content
- Modify the latent style component to realize style transfer
Generative Assumption

A latent style variable \( y \sim p(y) \)

A latent content variable \( z \sim p(z) \)

A sentence \( x \sim p(x|y, z) \)

We observe two corpora in different styles:

\[
X_1 = \{x_1^{(1)}, \cdots, x_1^{(n)}\} \text{ consisting of samples from } p(x|y_1)
\]

\[
X_2 = \{x_2^{(1)}, \cdots, x_2^{(m)}\} \text{ consisting of samples from } p(x|y_2)
\]
Model Overview

Encoder \( E : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{Z} \)

to infer the content for a given sentence and style

Generator \( G : \mathcal{Y} \times \mathcal{Z} \rightarrow \mathcal{X} \)

to generate a sentence from a given style and content

excellent chinese food and superb service

chinese food, service

bland chinese food and horrible service
Model Overview

$E$ and $G$ form an auto-encoder when applying to the same style

$$G(y_1, \cdot) \circ E(\cdot, y_1) = \text{id}_{\mathcal{X}_1} \quad G(y_2, \cdot) \circ E(\cdot, y_2) = \text{id}_{\mathcal{X}_2}$$
Model Overview

$E$ and $G$ form an auto-encoder when applying to the same style

$$G(y_1, \cdot) \circ E(\cdot, y_1) = \text{id}_{\mathcal{X}_1} \quad G(y_2, \cdot) \circ E(\cdot, y_2) = \text{id}_{\mathcal{X}_2}$$

$E$ and $G$ form a transfer model when applying to different styles

$$G(y_2, \cdot) \circ E(\cdot, y_1) : \mathcal{X}_1 \to \mathcal{X}_2 \quad G(y_1, \cdot) \circ E(\cdot, y_2) : \mathcal{X}_2 \to \mathcal{X}_1$$
Model Architecture

Encoder $E : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{Z}$

to infer the content for a given sentence and style
Model Architecture

Generator $G : \mathcal{Y} \times \mathcal{Z} \rightarrow \mathcal{X}$

to generate a sentence from a given style and content
Reconstruction Loss

$E$ and $G$ form an auto-encoder when applying to the same style

$$G(y_1, \cdot) \circ E(\cdot, y_1) = \text{id}_{\mathcal{X}_1} \quad G(y_2, \cdot) \circ E(\cdot, y_2) = \text{id}_{\mathcal{X}_2}$$

$$\mathcal{L}_{\text{rec}}(\theta_E, \theta_G) = \mathbb{E}_{x_1 \sim \mathcal{X}_1} \left[ - \log p_G(x_1 | y_1, E(x_1, y_1)) \right] + \mathbb{E}_{x_2 \sim \mathcal{X}_2} \left[ - \log p_G(x_2 | y_2, E(x_2, y_2)) \right]$$
Good to Go?

$E$ and $G$ form a transfer model when applying to different styles

$$
G(y_2, \cdot) \circ E(\cdot, y_1) : \mathcal{X}_1 \rightarrow \mathcal{X}_2 \quad G(y_1, \cdot) \circ E(\cdot, y_2) : \mathcal{X}_2 \rightarrow \mathcal{X}_1
$$

excellent chinese food
and superb service

? 

17
$E$ and $G$ form a transfer model when applying to different styles

$$G(y_2, \cdot) \circ E(\cdot, y_1) : \mathcal{X}_1 \rightarrow \mathcal{X}_2 \quad G(y_1, \cdot) \circ E(\cdot, y_2) : \mathcal{X}_2 \rightarrow \mathcal{X}_1$$

excellent chinese food and superb service
Just Copy, No Transfer

![Diagram of hope and reality in Chinese food and service]

Hope:

excellent chinese food and superb service

Reality:

0 reconstruction loss
Shared Content Distribution

Constrained optimization problem:

\[
\theta^* = \arg \min_{\theta} \mathcal{L}_{rec}(\theta_E, \theta_G)
\]

s.t. \( E(x_1, y_1) \overset{d}{=} E(x_2, y_2) \) \( x_1 \sim X_1, x_2 \sim X_2 \)
Aligned Auto-Encoder

\[
\theta^* = \arg\min_{\theta} \mathcal{L}_{\text{rec}}(\theta_E, \theta_G)
\]

\[
\text{s.t. } E(x_1, y_1) \overset{d}{=} E(x_2, y_2) \quad x_1 \sim X_1, x_2 \sim X_2
\]

Introduce \( D \) to distinguish \( Z_1 \) and \( Z_2 \):

\[
\mathcal{L}_{\text{adv}}(\theta_E, \theta_D) = \mathbb{E}_{x_1 \sim X_1}[- \log D(E(x_1, y_1))] + \mathbb{E}_{x_2 \sim X_2}[- \log (1 - D(E(x_2, y_2)))]
\]

\( Z_1 \overset{d}{=} Z_2 \) when they’re indistinguishable to \( D \)

Overall training objective:

\[
\min_{E,G} \max_D \mathcal{L}_{\text{rec}} - \lambda \mathcal{L}_{\text{adv}}
\]
Aligned Auto-Encoder

Results:

great!
horrible!

mediocre dim sum if you 're from southern california.
dim sum if you can not choose from california.

i would n't bother.
i would n't bother.

i would never go back for the food.
i would definitely go back for the food.

- 48.3% sentiment accuracy as measured by a classifier 😐
$z_1$ and $z_2$’s initial misalignment could propagate through the recurrent generating process

As a result the transferred sentence may end up somewhere far from the target domain
Cross Alignment

Transferred sentences from one style should match example sentence from the other style as a population

Introduce two discriminators:

\[ D_1 \text{ tries to distinguish } x_1 \text{ and transferred } x_2 \]

\[ D_2 \text{ tries to distinguish } x_2 \text{ and transferred } x_1 \]

\[
\min_{E,G} \max_{D_1,D_2} \mathcal{L}_{\text{rec}} - \lambda (\mathcal{L}_{\text{adv}1} + \mathcal{L}_{\text{adv}2})
\]
Cross Alignment

\[ E(\cdot, y_1) \quad G(y_2, \cdot) \]

\[ x_1 \quad z_1 \quad \tilde{x}_1 \quad D_2 \]

\[ x_2 \]

\[ 0 \quad 1 \]
Cross Alignment

discrete sampling process hinders gradients back-propagation

$$\begin{array}{c}
x_1 \\ \rightarrow \\ E(\cdot, y_1) \\ \rightarrow \\ z_1 \\ \leftarrow \\ G(y_2, \cdot) \\ \rightarrow \\ \tilde{x}_1 \\ \leftarrow \\ D_2 \\ \rightarrow \\ 0 \\ \rightarrow \\ 1 \\ \leftarrow \\ x_2 \\
\end{array}$$
Continuous Relaxation

Generator \( G: [y, z] \) → 

\[ \begin{align*}
\text{output logits} & \quad v_1 \\
<\text{go}> & \quad \text{softmax}(v_1/\gamma) \\
& \quad \vdots \\
& \quad \text{softmax}(v_{t-1}/\gamma) \\
& \quad \text{temperature}
\end{align*} \]
Professor Forcing

Match hidden states instead of output words
- contain all the information, smoothly distributed

Generator $G$: $[y, z]$  

Output logits

$v_1$  $v_2$  $v_t$

<go>  $\text{softmax}(v_1/\gamma)$  $\text{softmax}(v_{t-1}/\gamma)$

temperature

[Lamb et al. 2016]
Cross-Aligned Auto-Encoder

Enhances aligned auto-encoder, where only the first hidden states $z_1$ and $z_2$ are aligned
Cross-Aligned Auto-Encoder

Training procedure:

Take two mini-batches \( \{x_1^{(i)}\}_{i=1}^k \) from \( X_1 \) and \( \{x_2^{(i)}\}_{i=1}^k \) from \( X_2 \)

- Encode with \( E \rightarrow z_1^{(i)}, z_2^{(i)} \)
- Unroll \( G \) from \( (y_1, z_1^{(i)}), (y_2, z_2^{(i)}) \rightarrow h_1^{(i)}, h_2^{(i)} \)
  (reconstruction, teacher-forced by \( x^{(i)} \))
- Unroll \( G \) from \( (y_2, z_1^{(i)}), (y_1, z_2^{(i)}) \rightarrow \tilde{h}_1^{(i)}, \tilde{h}_2^{(i)} \)
  (style transfer, self-fed by previous output logits)

Update \( D_1 \) (and symmetrically \( D_2 \)) by gradient descent on loss:

\[
L_{adv_1} = -\frac{1}{k} \sum_{i=1}^{k} \log D_1(h_1^{(i)}) - \frac{1}{k} \sum_{i=1}^{k} \log(1 - D_1(\tilde{h}_2^{(i)}))
\]

Update \( E, G \) by gradient descent on loss \( L_{rec} - \lambda(L_{adv_1} + L_{adv_2}) \)
Cross-Aligned Auto-Encoder

Results:

great !
horrible !

mediocre dim sum if you 're from southern california .
good dim sum if you have korean friends .

i would n't bother .
i would recommend !

i would never go back for the food .
i would definitely go back for the food .

- 78.4% sentiment accuracy as measured by a classifier
Variational Auto-Encoder

Impose a prior $p(z) \sim \mathcal{N}(0, I)$

Maximize variational lower bound of data likelihood $-(\mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}})$

$$\mathcal{L}_{\text{KL}}(\theta_E) = \mathbb{E}_{x_1 \sim X_1} [D_{KL}(p_E(z|x_1, y_1) \| p(z))] + \mathbb{E}_{x_2 \sim X_2} [D_{KL}(p_E(z|x_2, y_2) \| p(z))]$$

Align both posteriors to the prior

[Kingma and Welling 2013]
Variational Auto-Encoder

\[ E(\cdot, y_1) \quad Z_1 \quad G(y_2, \cdot) \]

\[ \mathcal{X}_1 \quad \mathcal{Z}_1 \quad \mathcal{X}_2 \]

fast
slow
Variational Auto-Encoder

\[ E(\cdot, y_1) \quad Z_1 \quad G(y_2, \cdot) \quad Z_2 \quad \mathcal{X}_2 \]

Distributional alignment \( ? \) instance-level matching

Limiting \( z \) to a simple and even distribution is detrimental to content preservation
Sentiment Transfer Results

Model Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hu et al. (2017)</td>
<td>83.5</td>
</tr>
<tr>
<td>Variational auto-encoder</td>
<td>23.2</td>
</tr>
<tr>
<td>Aligned auto-encoder</td>
<td>48.3</td>
</tr>
<tr>
<td>Cross-aligned auto-encoder</td>
<td>78.4</td>
</tr>
</tbody>
</table>

Human Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>sentiment</th>
<th>fluency</th>
<th>overall transfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hu et al. (2017)</td>
<td>70.8</td>
<td>3.2</td>
<td>41.0</td>
</tr>
<tr>
<td>Cross-align</td>
<td>62.6</td>
<td>2.8</td>
<td>41.5</td>
</tr>
</tbody>
</table>

"Is the transferred sentence semantically equivalent to the source sentence with an opposite sentiment?"

Development of appropriate evaluation measures is crucial
consistently slow.
consistently good.
consistently fast.

my goodness it was so gross.
my husband's steak was phenomenal.
my goodness was so awesome.

i love the ladies here!
i avoid all the time!
i hate the doctors here!

came here with my wife and her grandmother!
came here with my wife and hated her!
came here with my wife and her son.

first line—input, second—Hu et al. (2017), third—Cross-align
Decipherment

Non-parallel transfer
Access only to the cipher text, want to transfer it into plain text
Keep the meaning, vary its style
Word Substitution Decipherment

Map every word to a cipher token according to a 1-to-1 substitution key

<table>
<thead>
<tr>
<th>cipher text</th>
<th>plain text</th>
</tr>
</thead>
<tbody>
<tr>
<td>eht azzip saw ton doog</td>
<td>the pizza was not good</td>
</tr>
<tr>
<td>ew lliw ton eb kcab</td>
<td>we will not be back</td>
</tr>
<tr>
<td>doog remotsuc ecivres</td>
<td>good customer service</td>
</tr>
<tr>
<td>os ytsan</td>
<td>so nasty</td>
</tr>
<tr>
<td>ym ssendoog ti saw os ssorg</td>
<td>my goodness it was so gross</td>
</tr>
<tr>
<td>ym etirovaf azzip</td>
<td>my favorite pizza</td>
</tr>
</tbody>
</table>
Word Substitution Decipherment

Non-parallel training, parallel evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>Substitution decipher</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20%</td>
</tr>
<tr>
<td>No transfer (copy)</td>
<td>56.4</td>
</tr>
<tr>
<td>Unigram matching</td>
<td>74.3</td>
</tr>
<tr>
<td>Variational auto-encoder</td>
<td>79.8</td>
</tr>
<tr>
<td>Aligned auto-encoder</td>
<td>81.0</td>
</tr>
<tr>
<td>Cross-aligned auto-encoder</td>
<td>83.8</td>
</tr>
<tr>
<td>Parallel translation</td>
<td>99.0</td>
</tr>
</tbody>
</table>

Bleu score between plain text and transferred cipher text
Randomly shuffle a sentence, recover its original word order

**Bag of words**

! 'm impressed so

was even it how i
gross handle n't

really . is which they
have good and daily
also ice specials cream

**Grammatical sentence**

i 'm so impressed !

i ca n't even handle
how gross it was .

they also have daily specials and
ice cream which is really good .
Word Ordering

Non-parallel training, parallel evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>Order recover</th>
</tr>
</thead>
<tbody>
<tr>
<td>No transfer (copy)</td>
<td>5.1</td>
</tr>
<tr>
<td>Variational auto-encoder</td>
<td>5.3</td>
</tr>
<tr>
<td>Aligned auto-encoder</td>
<td>5.2</td>
</tr>
<tr>
<td>Cross-aligned auto-encoder</td>
<td><strong>26.1</strong></td>
</tr>
<tr>
<td>Parallel translation</td>
<td>64.6</td>
</tr>
</tbody>
</table>

Bleu score between grammatical sentences and transferred shuffled sentences
Conclusion

- Non-parallel style transfer
  keep the content, vary the style

- Cross-aligned auto-encoder
  transferred sentences from one style should match example sentence from the other style

- Distributional alignment \( \rightarrow \) instance-level matching

- Applications
  sentiment transfer, decipherment, word ordering
Future Work

• *Real* language style transfer
  
critic $\leftrightarrow$ general audience movie reviews
  
Shakespeare $\leftrightarrow$ Trump, CNN $\leftrightarrow$ Fox news

• Evaluation
  
how to measure the transferred sentence preserves the content?
how to measure it has the target style?

• Better model
  
attention, specific constraints…


Code & data: https://github.com/shentianxiao/language-style-transfer