Revision in Continuous Space: Unsupervised Text Style Transfer without Adversarial Learning

Dayiheng Liu¹, Jie Fu², Yidan Zhang¹, Chris Pal², Jiancheng Lv¹
¹College of Computer Science, Sichuan University
²Mila, IVADO, Polytechnique Montreal

Introduction

Unsupervised Text Style Transfer:
1. Converting some attributes of a sentence (e.g., negative sentiment) to other attributes (e.g., positive sentiment)
2. Preserving attribute-independent content
3. Accessing non-parallel, but style labeled sentences

Previous works: (1) seeking the explicit disentanglement of the content and the attributes. (2) troublesome adversarial learning

This paper:
- **Easily Training**: The method can be easily trained on the non-parallel dataset, avoiding the problem of training difficulties caused by adversarial learning and achieving higher performance
- **Diverse, controllable, and interpretable**: Our method revises the original sentence with gradient information for several steps during inference, which explicitly presents the process of the style transfer and can easily provide us multiple results with tuning the gradients. Therefore, the proposed method has higher interpretability and is more controllable
- **Control multiple fine-grained attributes**: Our approach is more generic in the sense that it naturally has the ability to control multiple fine-grained attributes, such as sentence length and the existence of specific words

Experiments

Three Datasets:
- Amazon
- Yelp (sentiment)
- Yelp (Gender)

Metrics:
- Accuracy
- PPL
- Overlap Noun BLEU

<table>
<thead>
<tr>
<th>Methods</th>
<th>Accuracy</th>
<th>PPL</th>
<th>Overlap</th>
<th>Noun/BLEU</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>0.1</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Delta, Retrieve, &amp; Generate [Liu et al. 2018]:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TemplateBased</td>
<td>81.3</td>
<td>43.6</td>
<td>67.2</td>
<td>83.3</td>
<td>28.9</td>
</tr>
<tr>
<td>DeltaOnly</td>
<td>85.6</td>
<td>43.1</td>
<td>67.5</td>
<td>74.9</td>
<td>24.7</td>
</tr>
<tr>
<td>DeltaAndRetrieve</td>
<td>89.5</td>
<td>39.4</td>
<td>70.4</td>
<td>74.0</td>
<td>24.9</td>
</tr>
<tr>
<td>Delta &amp; Retrieve [Liu et al. 2018]:</td>
<td>98.4</td>
<td>37.7</td>
<td>75.7</td>
<td>90.6</td>
<td>6.7</td>
</tr>
<tr>
<td>StyleInversion [Pan et al. 2018]:</td>
<td>7.2</td>
<td>93.9</td>
<td>75.4</td>
<td>54.2</td>
<td>0.3</td>
</tr>
<tr>
<td>MultiDecoder [Pan et al. 2018]</td>
<td>48.6</td>
<td>106.3</td>
<td>51.5</td>
<td>52.2</td>
<td>23.1</td>
</tr>
<tr>
<td>BTS [Pathak et al. 2018]</td>
<td>94.8</td>
<td>32.8</td>
<td>72.3</td>
<td>73.5</td>
<td>6.3</td>
</tr>
<tr>
<td>CrossAligned [Senn et al. 2017]:</td>
<td>72.6</td>
<td>72.0</td>
<td>41.1</td>
<td>42.9</td>
<td>18.4</td>
</tr>
<tr>
<td>Over (content-strength)</td>
<td>98.5</td>
<td>25.3</td>
<td>70.2</td>
<td>93.0</td>
<td>21.8</td>
</tr>
<tr>
<td>Over (style-strength)</td>
<td>92.3</td>
<td>18.3</td>
<td>38.9</td>
<td>69.5</td>
<td>18.8</td>
</tr>
<tr>
<td>Over (style-content balance)</td>
<td>99.7</td>
<td>26.6</td>
<td>39.7</td>
<td>65.5</td>
<td>17.9</td>
</tr>
</tbody>
</table>

Methods:
- Accuracy | PPL | Overlap | Noun/BLEU |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>88.1</td>
<td>62.9</td>
<td>60.5</td>
</tr>
</tbody>
</table>

Variational auto-encoder:

\[
\mathcal{L}_{\text{VAE}}(\theta_{\text{ENC}}, \theta_{\text{REC}}) = \mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{KL}} = -E_{q(z|x)}[\log p(z|x)] + D_{KL}(q(z|x)\|p(z))
\]

Content predictor:

\[
\begin{align*}
\hat{f}_{\text{bow}}(z) &= \text{MLP}_{\text{bow}}(z) = p(x_{\text{bow}}|z) \\
\log p(x_{\text{bow}}|z) &= \log \prod_{t=1}^{T} \sum_{s} \mathbf{e}_{s}(z) \\
\mathcal{L}_{\text{bow}}(\theta_{\text{bow}}, \theta_{\text{rec}}) &= -E_{q(z|x)}[\log p(x_{\text{bow}}|z)]
\end{align*}
\]

Attribute predictors:

\[
\begin{align*}
\mathcal{L}_{\text{Attr,}s_j}(\theta_{s_j}, \theta_{\text{rec}}) &= -E_{q(z|x)}[\log f_j(z)] \\
\mathcal{L}_{\text{Attr,}s_j}(\theta_{s_j}, \theta_{\text{enc}}) &= -E_{q(z|x)}[\log f_j(z)] \\
\mathcal{L}_{\text{Attr,}s_j}(\theta_{s_j}) &= -E_{p(z|x)}[\log p(\text{CNN}(x)|z)] \\
\mathcal{L}_{\text{Attr,}s_j}(\theta_{s_j}) &= -E_{p(z|x)}[\log p(x_{\text{bow}}|z)]
\end{align*}
\]

Total loss:

\[
\mathcal{L} = \mathcal{L}_{\text{VAE}} + \lambda_{\text{bow}} \mathcal{L}_{\text{bow}} + \sum_{s=1}^{k} \mathcal{L}_{\text{Attr,}s_j}
\]

Inference:

\[
\hat{z} = z - \eta \sum_{j=1}^{k} \nabla_{\hat{z}} \mathcal{L}_{\text{Attr,}s_j} + \lambda_{\text{bow}} \nabla_{\hat{z}} \mathcal{L}_{\text{bow}}
\]

Our code and data are available at https://github.com/dayihengliu/Fine-Grained-Style-Transfer