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

Dayiheng Liu[1], Jie Fu[2], Yidan Zhang[1], Chris Pal[2], Jiancheng Lv[1]
[1]College of Computer Science, Sichuan University
[2]Mila, 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

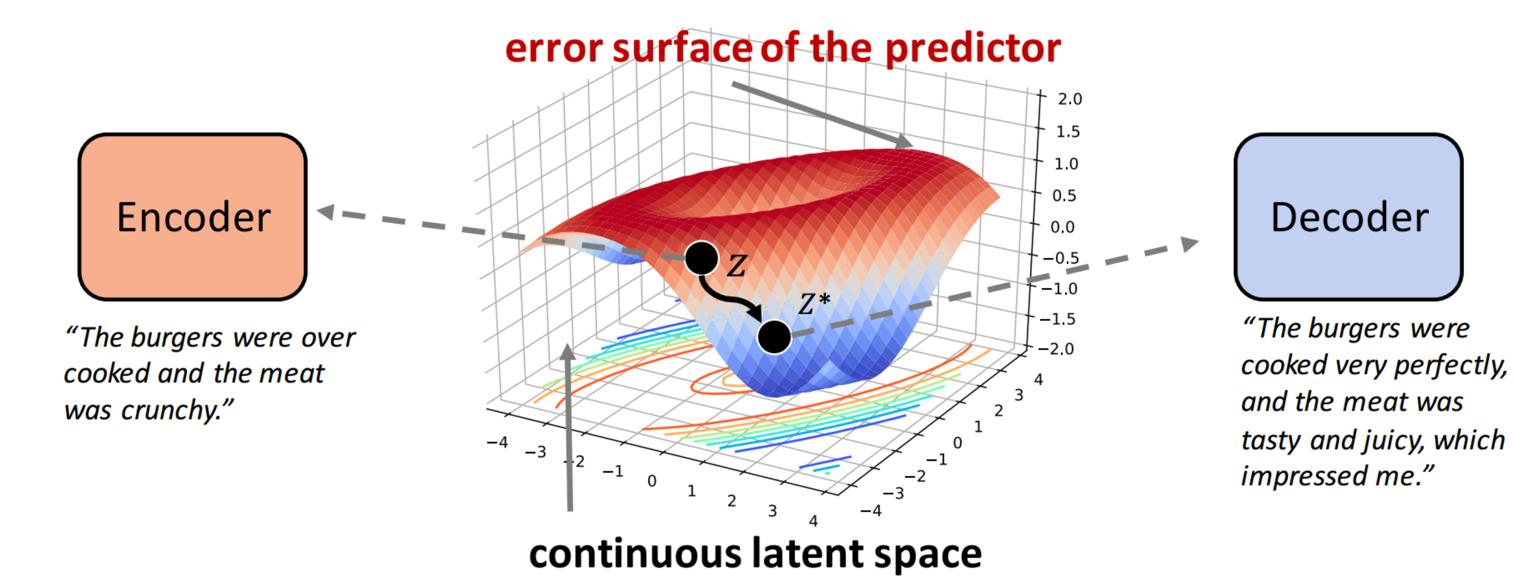
Methods	Accuracy↑	PPL↓	Overlap↑	Noun%↑	BLEU↑
Original	0.1	22.9	100.0	100.0	42.4
Human	91.8	76.9	47.2	78.5	100.0
Delete, Retrieve, & Generate (Li et al. 2018):					
TemplateBased	81.3	183.6	55.6	**83.3**	28.9
DeleteOnly	85.8	81.4	49.5	74.9	24.7
DeleteAndRetrieve	89.5	96.1	49.4	74.0	24.9
RetrievalOnly	**98.4**	25.7	15.8	39.6	4.7
StyleEmbedding (Fu et al. 2018)	7.2	93.9	**75.4**	74.2	**31.9**
MultiDecoder (Fu et al. 2018)	48.8	166.5	51.5	52.2	23.1
BTS (Prabhumoye et al. 2018)	94.8	32.8	21.5	23.5	6.8
CrossAligned (Shen et al. 2017)	73.6	72.0	41.1	42.9	18.4
Ours (content-strengthen)	88.2	26.5	46.6	77.4	21.8
Ours (style-content balance)	92.3	**18.3**	38.9	69.3	18.8
Ours (style-strengthen)	95.7	20.6	39.7	61.5	17.9

Methods	Accuracy↑	PPL↓	Overlap↑	Noun%↑	BLEU↑
Original	23.4	24.4	100.0	100.0	57.2
Human	88.1	62.9	60.5	85.0	100.0
Delete, Retrieve, & Generate (Li et al. 2018):					
TemplateBased	69.6	108.9	73.3	87.9	42.8
DeleteOnly	51.6	49.3	**74.4**	**95.1**	**44.7**
DeleteAndRetrieve	55.2	48.2	69.1	92.6	41.8
RetrievalOnly	87.2	28.7	21.0	44.5	6.7
StyleEmbedding (Fu et al. 2018)	40.5	87.7	42.2	41.8	22.1
MultiDecoder (Fu et al. 2018)	66.5	80.8	30.6	30.4	14.3
BTS (Prabhumoye et al. 2018)	82.6	25.3	24.7	22.5	9.2
CrossAligned (Shen et al. 2017)	69.6	18.3	19.3	20.4	5.0
Ours (content-strengthen)	81.9	35.0	37.7	76.0	11.5
Ours (style-content balance)	85.1	21.8	49.3	49.8	21.5
Ours (style-strengthen)	**90.0**	**15.9**	39.5	41.4	16.3

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

Methodology

Core idea:
The proposed model consists of three components: (1) a variational auto-encoder (VAE), (2) attribute predictors, and (3) a content predictor. Predictors takes the continuous representation of a sentence as input and predicts its Bag-of-words content and other attributes. With the gradients obtained from these predictors, we can revise the continuous representation of the original sentence by gradient-based optimization to find a target sentence with the desired fine-grained attributes, and achieve the content-preserving text style transfer.



Variational auto-encoder:

$$\mathcal{L}_{\text{VAE}}(\theta_{\text{enc}}, \theta_{\text{dec}}) = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}}$$
$$= -\mathbb{E}_{q_E(z|x)}\left[\log p_G(x|z)\right] + \mathcal{D}_{\text{KL}}(q_E(z|x)\|p(z)),$$

Content predictor:

$$f_{\text{bow}}(z) = \text{MLP}_{\text{bow}}(z) = p(x_{\text{bow}}|z).$$

$$\log p(x_{\text{bow}}|z) = \log \prod_{t=1}^{|x|} \frac{e^{f_{\text{bow}}^{(x_t)}}}{\sum_j^{\mathcal{V}} e^{f_{\text{bow}}^{(x_j)}}}$$

$$\mathcal{L}_{\text{BOW}}(\theta_{\text{bow}}, \theta_{\text{enc}}) = -\mathbb{E}_{q_E(z|x)} \log\left[p(x_{\text{bow}}|z)\right]$$

Attribute predictors:

$$\mathcal{L}_{\text{Attr},s_j}(\theta_{s_j}, \theta_{\text{enc}}) = -\mathbb{E}_{q_E(z|x)} \log\left[f_j(z)\right]$$

$$\mathcal{L}_{\text{Attr},s_j}(\theta_{s_j}, \theta_{\text{enc}}) = -\mathbb{E}_{q_E(z|x)} \log\left[f_j(z)\right]$$

$$\mathcal{L}'_{\text{Attr},s_j}(\theta_{s_j}) = -\mathbb{E}_{p(z)p_G(\hat{x}|z)} \log\left[p(\text{CNN}(\hat{x})|z)\right],$$
$$\mathcal{L}'_{\text{Attr},s_j}(\theta_{s_j}) = \mathbb{E}_{p(z)p_G(\hat{x}|z)}\left[(\hat{s}_j - f_j(z))^2\right].$$

Total loss:

$$\mathcal{L} = \mathcal{L}_{\text{VAE}} + \lambda_b \mathcal{L}_{\text{BOW}} + \lambda_s \sum_{i=1}^{k} \mathcal{L}_{\text{Attr},s_j}$$

Inference:

$$\hat{z} = z - \eta\left(\sum_{j=1}^{k} \nabla_z \mathcal{L}_{\text{Attr},s_j} + \lambda_c \nabla_z \mathcal{L}_{\text{BOW}}\right)$$