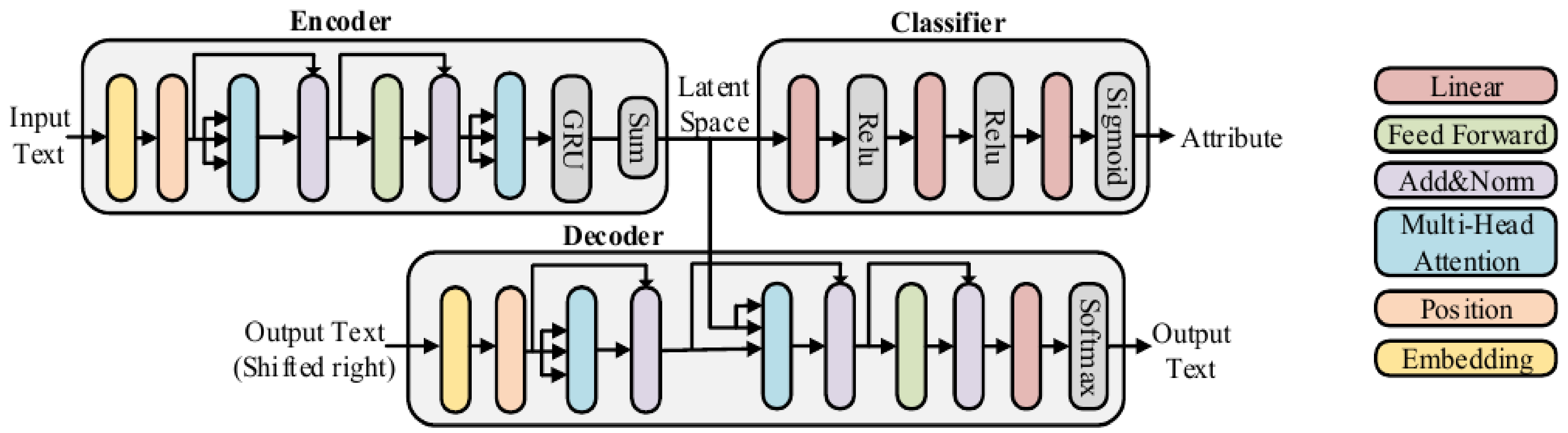
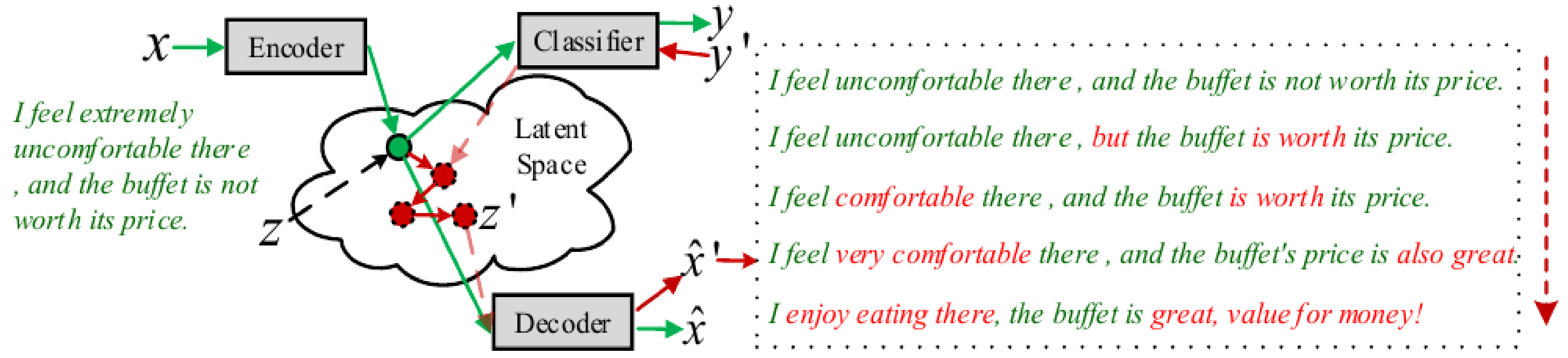
Controllable Unsupervised Text Attribute Transfer via Editing Entangled Latent Representation

Ke Wang, Hang Hua and Xiaojun Wan
Wangxuan Institute of Computer Technology, Peking University

Introduction:

- Unsupervised text attribute transfer automatically transforms a text to alter a specific attribute (e.g. sentiment) without using any parallel data, while simultaneously preserving its attribute-independent content. The dominant approaches are trying to model the content-independent attribute separately, e.g., learning different attributes' representations or using multiple attribute-specific decoders.
- In this work, we present a controllable unsupervised text attribute transfer framework, which can edit the entangled latent representation instead of modeling attribute and content separately. Specifically, we first propose a Transformer-based autoencoder to learn an entangled latent representation for a discrete text, then we transform the attribute transfer task to an optimization problem and propose the Fast-Gradient-Iterative-Modification algorithm to edit the latent representation until conforming to the target attribute.




x
Encoder
Classifier
y
y'
I feel extremely uncomfortable there , and the buffet is not worth its price.
Latent Space
z
z'
Decoder
$\hat{x}'$
$\hat{x}$

I feel uncomfortable there , and the buffet is not worth its price.
I feel uncomfortable there , but the buffet is worth its price.
I feel comfortable there , and the buffet is worth its price.
I feel very comfortable there , and the buffet's price is also great.
I enjoy eating there, the buffet is great, value for money!

Problem:

- Given a source text with an attribute (e.g., positive sentiment), the goal of the task is to generate a new text with a different attribute (e.g., negative sentiment). The generated text should meet the requirements:
  a. Maintaining the attribute-independent content as the source text;
  b. Conforming to the target attribute;
  c. Maintaining the linguistic fluency.

Model:

- The whole framework can be divided into three sub-models: an encoder **E** which encodes the text **x** into a latent representation z, a decoder **D** which decodes text **x** from z, and an attribute classifier **C** that classifies attribute of the latent representation *z*.

$$\boldsymbol{z} = E_{\theta_e}(\boldsymbol{x});\ \boldsymbol{y} = C_{\theta_c}(\boldsymbol{z});\ \hat{\boldsymbol{x}} = D_{\theta_d}(\boldsymbol{z})$$

- We formulate the text attribute transfer task as an optimization problem:

$$\hat{\boldsymbol{x}}' = D_{\theta_d}(\boldsymbol{z}')\ where\ \boldsymbol{z}' = argmin_{\boldsymbol{z}^*}||\boldsymbol{z}^* - E_{\theta_e}(\boldsymbol{x})||\ s.t.\ C_{\theta_c}(\boldsymbol{z}^*) = \boldsymbol{y}'$$

- We propose the Fast-Gradient-Iterative-Modification algorithm (FGIM), which modifies *z* based on the gradient of back-propagation by linearizing the attribute classifier's loss function on *z*.

Algorithm 1 Fast Gradient Iterative Modification Algorithm.

```
Input: Original latent representation $\boldsymbol{z}$; Well-trained attribute classifier $C_{\theta_c}$; A set of weights $\boldsymbol{w} = \{w_i\}$;
    Decay coefficient $\lambda$; Target attribute $\boldsymbol{y}'$; Threshold $t$;
Output: An optimal modified latent representation $\boldsymbol{z}'$;
for each $w_i \in \boldsymbol{w}$ do
    $\boldsymbol{z}^* = \boldsymbol{z} - w_i \nabla_{\boldsymbol{z}} \mathcal{L}_c(C_{\theta_c}(\boldsymbol{z}), \boldsymbol{y}')$;
    for s-steps do
        if $|\boldsymbol{y}' - C_{\theta_c}(\boldsymbol{z}^*)| < t$ then $\boldsymbol{z}' = \boldsymbol{z}^*$ ; Break;
        end if
        $w_i = \lambda w_i$;
        $\boldsymbol{z}^* = \boldsymbol{z}^* - w_i \nabla_{\boldsymbol{z}^*} \mathcal{L}_c(C_{\theta_c}(\boldsymbol{z}^*), \boldsymbol{y}')$;
    end for
end for
return $\boldsymbol{z}'$;
```

Results:

- We use datasets provided in Li et al. for sentiment and style transfer experiments, where the test sets contain human-written references.

Methods	Yelp			Amazon			Captions			Yelp			Amazon			Captions		
	Acc	BLEU	PPL↓	Acc	BLEU	PPL↓	Acc	BLEU	PPL↓	Att	Con	Gra	Att	Con	Gra	Att	Con	Gra
CrossAlign [28]	72.3%	9.1	50.8	70.3%	1.9	66.2	78.3%	1.8	69.8	2.5	2.8	3.3	2.7	2.7	3.1	2.1	2.5	3.0
MultiDec [5]	50.2%	14.5	84.5	67.3%	9.1	60.3	68.3%	6.6	60.2	2.3	3.1	2.7	2.6	2.9	2.9	2.5	2.6	2.9
StyleEmb [5]	10.2%	21.1	47.9	43.6%	15.1	60.1	56.2%	8.8	57.1	2.6	3.0	2.9	3.1	2.8	3.2	2.3	3.1	3.0
CycleRL [38]	53.6%	18.8	98.2	52.3%	14.4	183.2	45.2%	5.8	50.3	2.9	3.0	3.2	3.2	3.1	3.2	2.5	2.9	2.8
BackTrans [26]	93.4%	2.5	49.5	84.6%	1.5	48.3	78.3%	1.6	68.3	2.0	2.4	2.9	2.6	2.8	3.4	2.4	2.8	2.8
RuleBase [17]	80.3%	22.6	66.6	67.8%	33.6	52.1	85.3%	**19.2**	35.6	3.4	3.2	3.4	3.6	3.7	3.8	2.6	3.1	3.0
DelRetrGen [17]	88.8%	16.0	49.6	51.2%	29.3	55.4	90.4%	12.0	33.4	3.2	2.9	3.0	3.7	3.6	3.4	2.5	2.9	3.2
UnsupMT [41]	95.2%	22.8	53.9	84.2%	33.9	57.9	**95.5%**	12.7	31.2	3.2	3.3	3.5	3.7	4.0	3.7	2.8	2.8	3.3
Ours	**95.4%**	**24.6**	**46.2**	**85.3%**	**34.1**	**47.4**	92.3%	17.6	**23.7**	**3.6**	**3.5**	**3.8**	**4.0**	**4.2**	**4.1**	**3.5**	**3.4**	**3.5**

- Latent Representation Modification Study

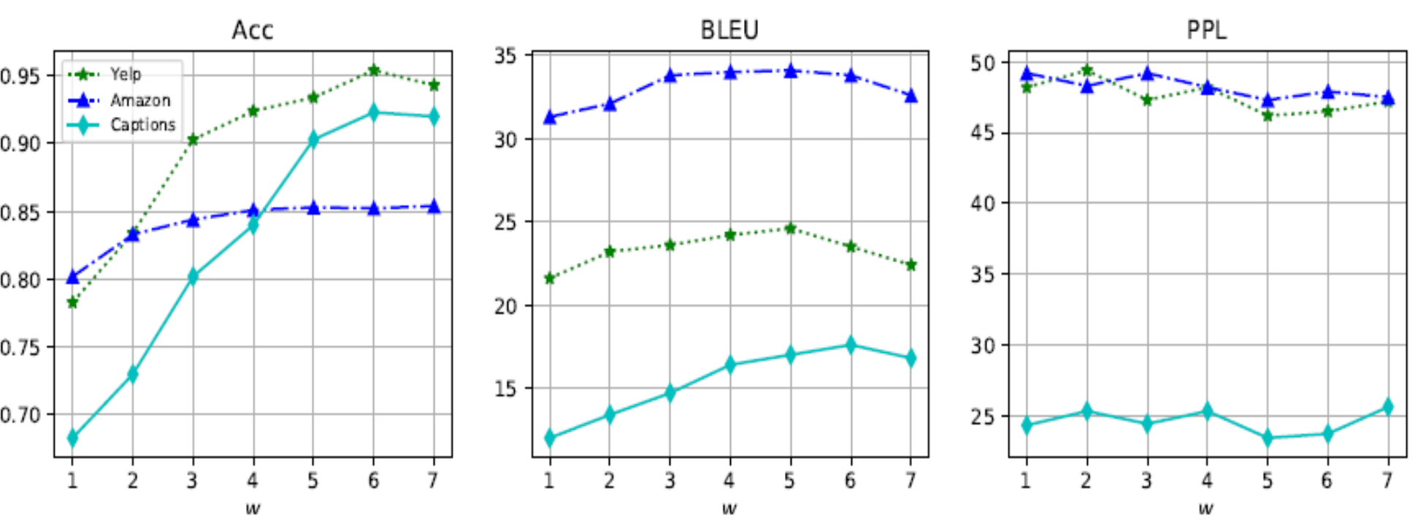


Figure 2: Influence of the modification weight w.

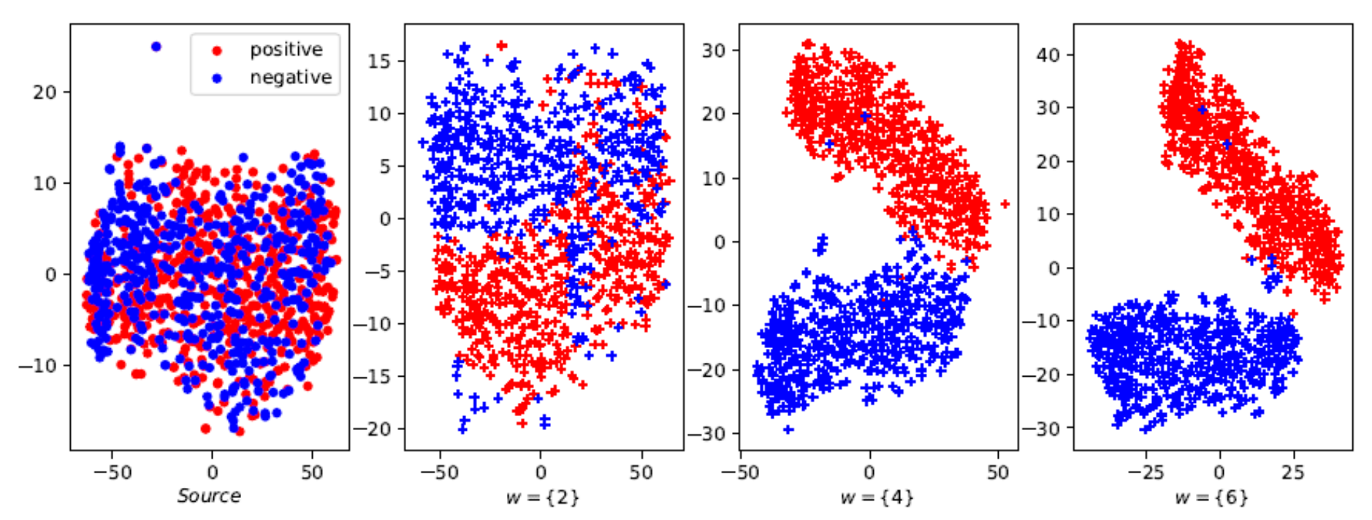


Figure 3: Visualization of representations with different modification weight w.

- Transfer Degree Control

Table 5: Examples of generation with different modification weight w.

	Positive ->Negative	Negative ->Positive
Source:	really good service and food .	it is n't terrible , but it is n't very good either .
Human:	the service was bad	it is n't perfect , but it is very good .
$w = \{1\}$	really good service and food .	it is n't terrible , but it is n't very good either .
$w = \{2\}$	very good service and food .	it is n't terrible , but it is n't very good delicious either .
$w = \{3\}$	very good food but service is terrible !	it is n't terrible , but it is very good delicious either .
$w = \{4\}$	not good food and service is terrible !	it is n't terrible , but it is very good and delicious .
$w = \{5\}$	bad service and food !	it is n't terrible , but it is very good and delicious appetizer .
$w = \{6\}$	very terrible service and food !	it is excellent , and it is very good and delicious well .

Conclusion:

- We present a controllable unsupervised text attribute transfer framework, which can edit the entangled latent representation instead of modeling attribute and content separately.
- To the best of our knowledge, this is the first one that can not only control the degree of transfer freely but also perform sentiment transfer over multiple aspects at the same time.

Code is available here!