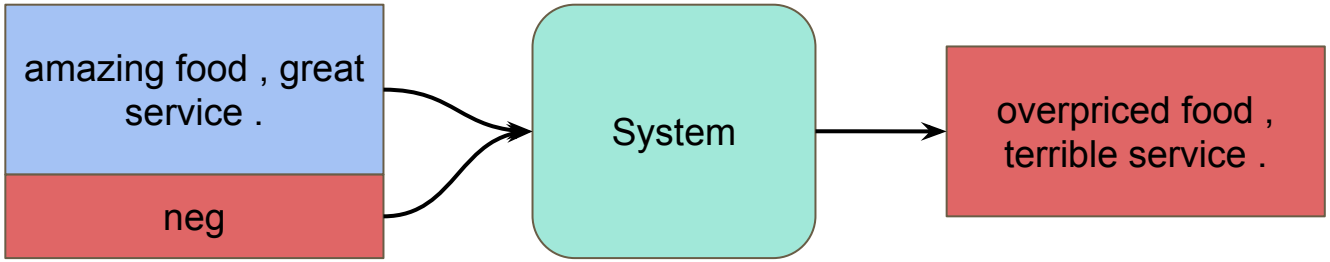
HW5-1 Text Style Transfer

TA 張致強 蔡翔陞 鍾起鳴

5.1.1 Style-Transformer

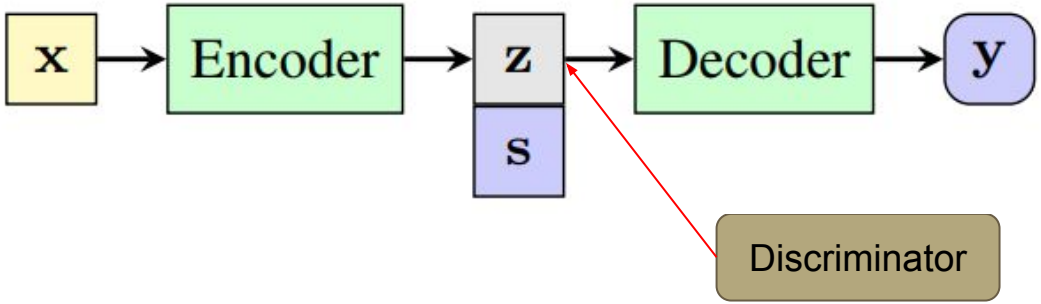
Text Style Transfer



Typical Unsupervised Approach

Disentangle **content** and **style** in the latent space.

A discriminator is used to make sure z is style independent.



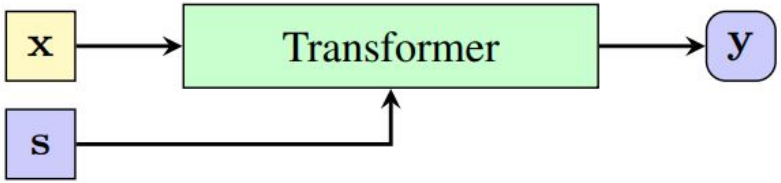
Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation

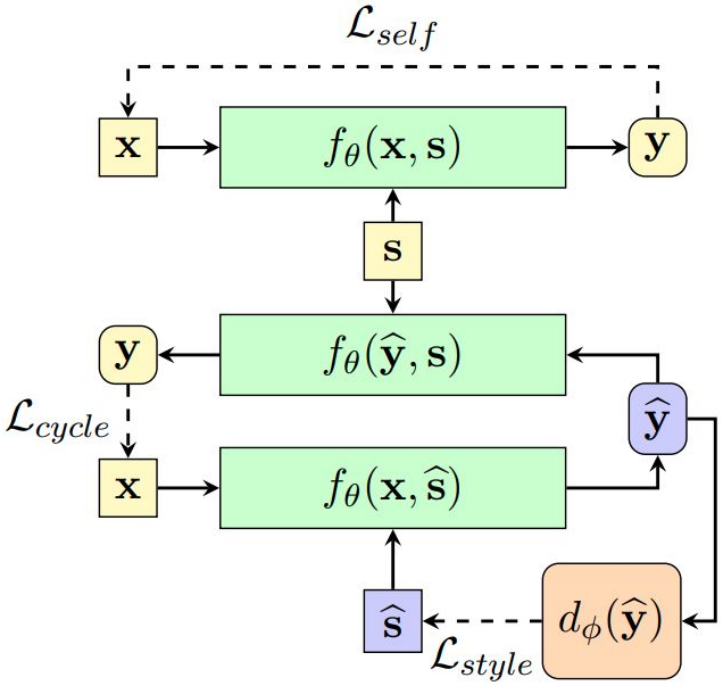
Ning Dai, Jianze Liang, Xipeng Qiu* , Xuanjing Huang

Fudan University

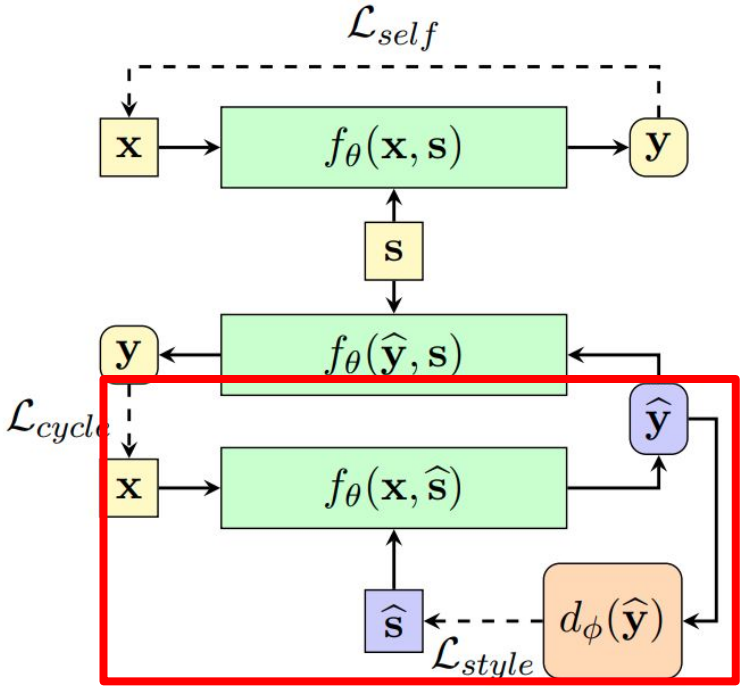
Approach

- No disentanglement
- No compressed representation





GAN!


$\mathcal{L}_{self}$
$\mathbf{x}$
$f_\theta(\mathbf{x}, \mathbf{s})$
$\mathbf{y}$
$\mathbf{s}$
$\mathbf{y}$
$f_\theta(\widehat{\mathbf{y}}, \mathbf{s})$
$\mathcal{L}_{cycle}$
$\widehat{\mathbf{y}}$
$\mathbf{x}$
$f_\theta(\mathbf{x}, \widehat{\mathbf{s}})$
$\widehat{\mathbf{s}}$
$d_\phi(\widehat{\mathbf{y}})$
$\mathcal{L}_{style}$

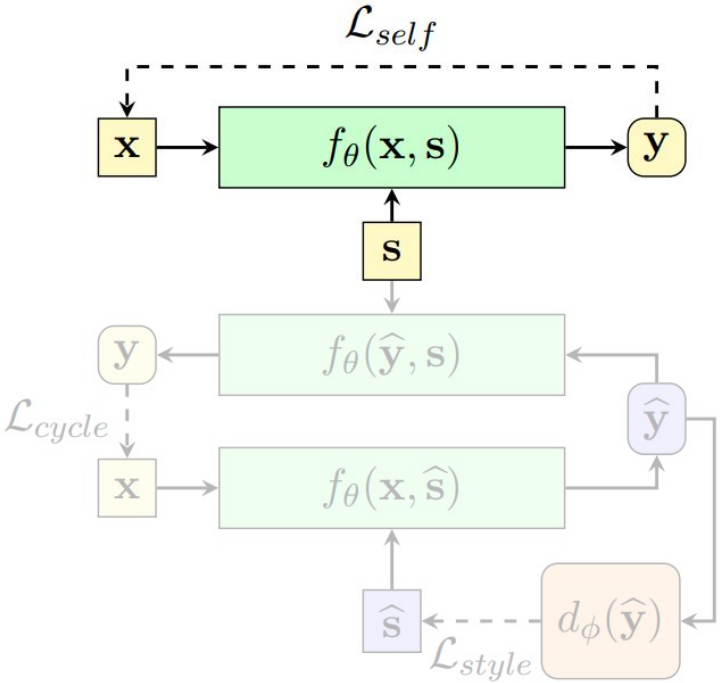
Training Style-Transformer

- Self Reconstruction

  $\mathcal{L}_{self}(\theta) = -p_\theta(\mathbf{y} = \mathbf{x}|\mathbf{x}, \mathbf{s})$
- Cycle Reconstruction

  $\mathcal{L}_{cycle}(\theta) = -p_\theta(\mathbf{y} = \mathbf{x}|f_\theta(\mathbf{x}, \widehat{\mathbf{s}}), \mathbf{s})$
- Style Controlling

  $\mathcal{L}_{style}(\theta) = -p_\phi(\mathbf{c} = \widehat{\mathbf{s}}|f_\theta(\mathbf{x}, \widehat{\mathbf{s}}))$



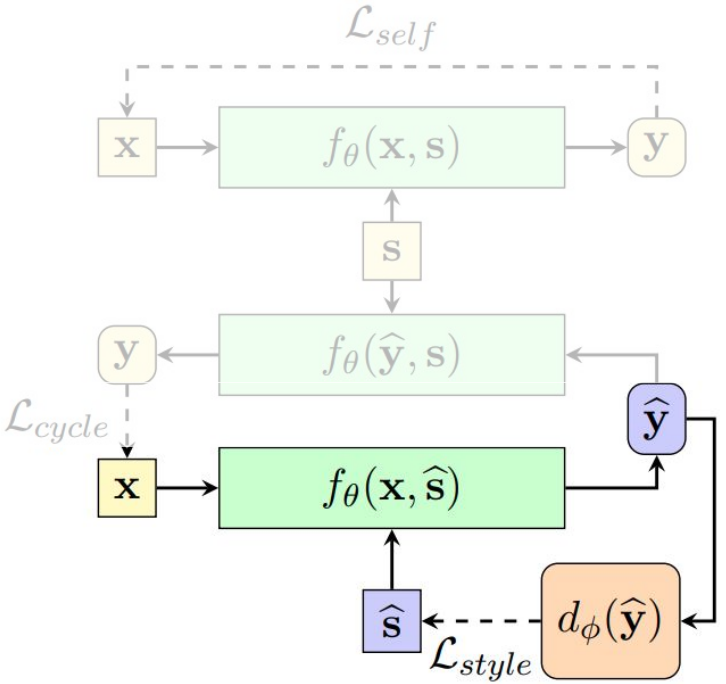
Training Style-Transformer

- Self Reconstruction

  $\mathcal{L}_{self}(\theta) = -p_\theta(\mathbf{y} = \mathbf{x}|\mathbf{x}, \mathbf{s})$
- Cycle Reconstruction

  $\mathcal{L}_{cycle}(\theta) = -p_\theta(\mathbf{y} = \mathbf{x}|f_\theta(\mathbf{x}, \widehat{\mathbf{s}}), \mathbf{s})$
- Style Controlling

  $\mathcal{L}_{style}(\theta) = -p_\phi(\mathbf{c} = \widehat{\mathbf{s}}|f_\theta(\mathbf{x}, \widehat{\mathbf{s}}))$



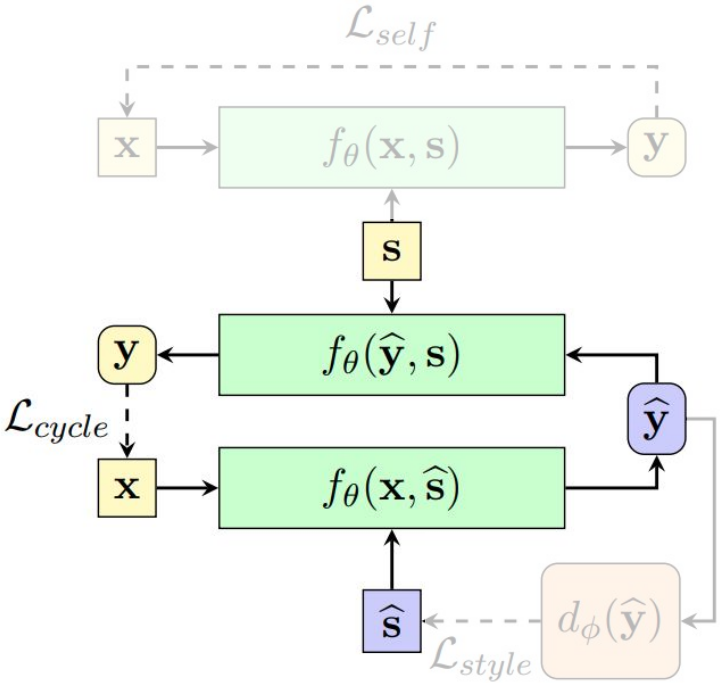
Training Style-Transformer

- Self Reconstruction

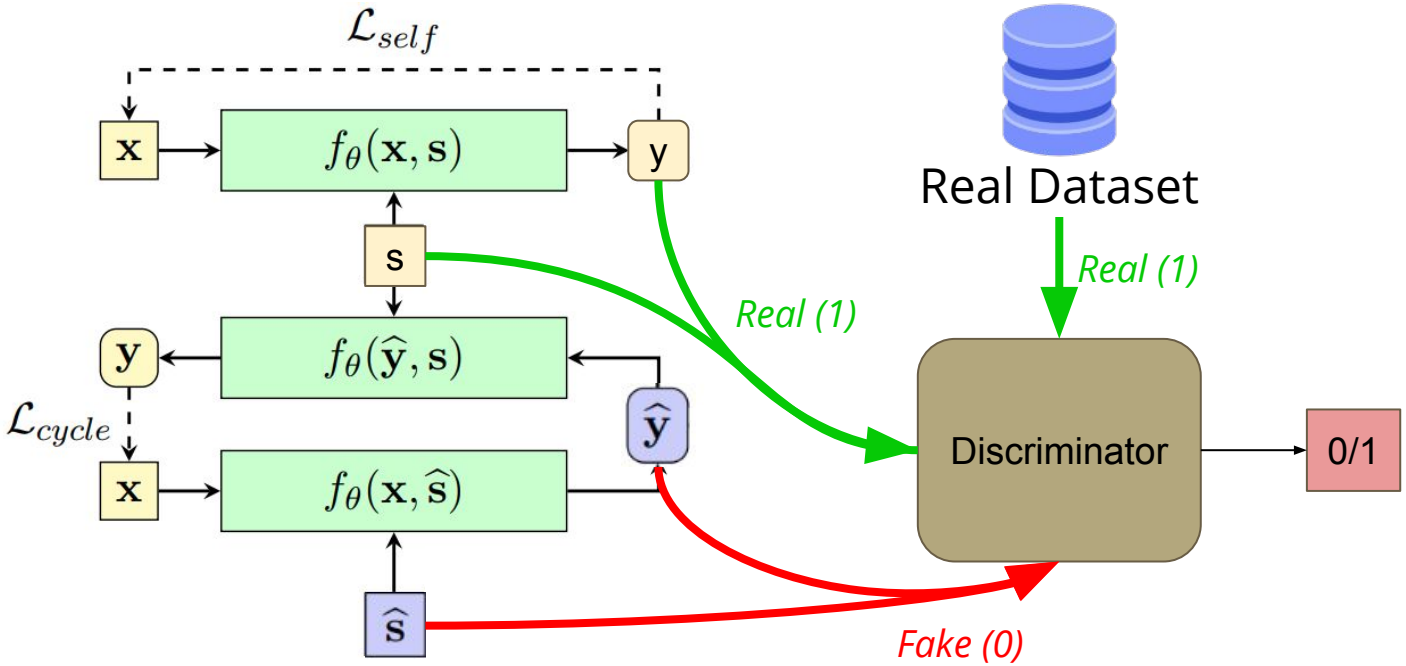
  $\mathcal{L}_{self}(\theta) = -p_\theta(\mathbf{y} = \mathbf{x}|\mathbf{x}, \mathbf{s})$
- Cycle Reconstruction

  $\mathcal{L}_{cycle}(\theta) = -p_\theta(\mathbf{y} = \mathbf{x}|f_\theta(\mathbf{x}, \widehat{\mathbf{s}}), \mathbf{s})$
- Style Controlling

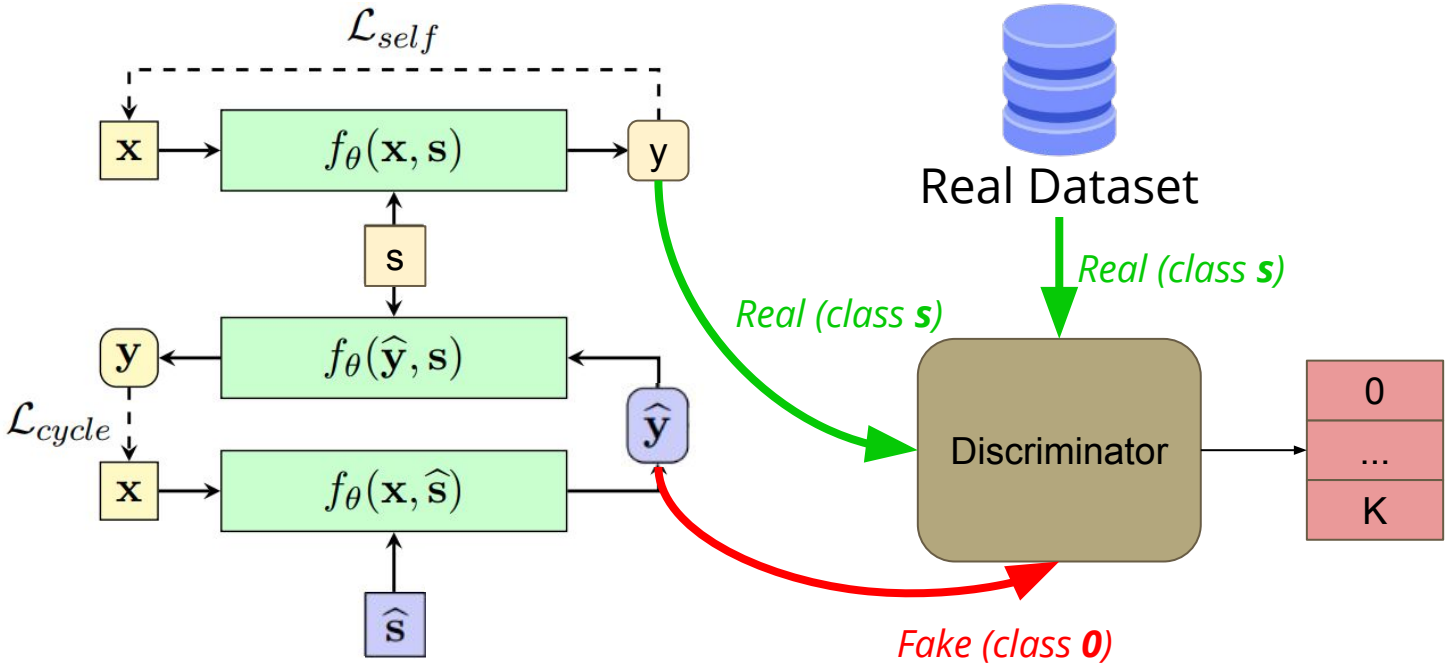
  $\mathcal{L}_{style}(\theta) = -p_\phi(\mathbf{c} = \widehat{\mathbf{s}}|f_\theta(\mathbf{x}, \widehat{\mathbf{s}}))$



Training the Discriminator (Conditional)



Training the Discriminator (Multiclass)



Evaluating Text Style Transfer System

- Accuracy: Classification accuracy of transferred sentence.
- BLEU: Content preservation.
  - self-BLEU: score by source sentence
  - ref-BLEU: score by reference (target) sentence
- Perplexity: Sentence fluency.
- Human Evaluation

5.1.2 Tasks

Code

Github - https://github.com/MarvinChung/HW5-TextStyleTransfer

See README.md for instructions to run.

Configurations

You can (should) tune some of the following args to obtain better results.

Model related

argument	description
discriminator_method	Two types of discriminator described in paper.
embed_size, d_model, head, num_layers	Transformer architecture.
learned_pos_embed	Use trainable position encoding or sinusoidal.

Training related

argument	description
batch_size, lr_F, lr_D	Typical training hyperparameters.
L2, dropout, inp_drop_prob	Various regularizations.
iter_D, iter_F, F_pretrain_iter	Affects relative strengths of style-transformer and discriminator while training.
slf_factor, cyc_factor, adv_factor	Weight factors of three loss described in paper.

What you should do

1. Understand and train a Text Style Transfer model
2. Show your training configuration
    a. show what's different from default
3. Show your training curves
    a. Plot the three loss in style-transformer
    b. Plot the discriminator loss
4. Evaluate your model
    a. Report your best model's accuracy, ref-BLEU and perplexity
    b. Submit your model's output on the 1000 yelp testing data
    c. Show some results on sentences inside or outside of the yelp dataset

5.1.3 Sample Report

HW5-1

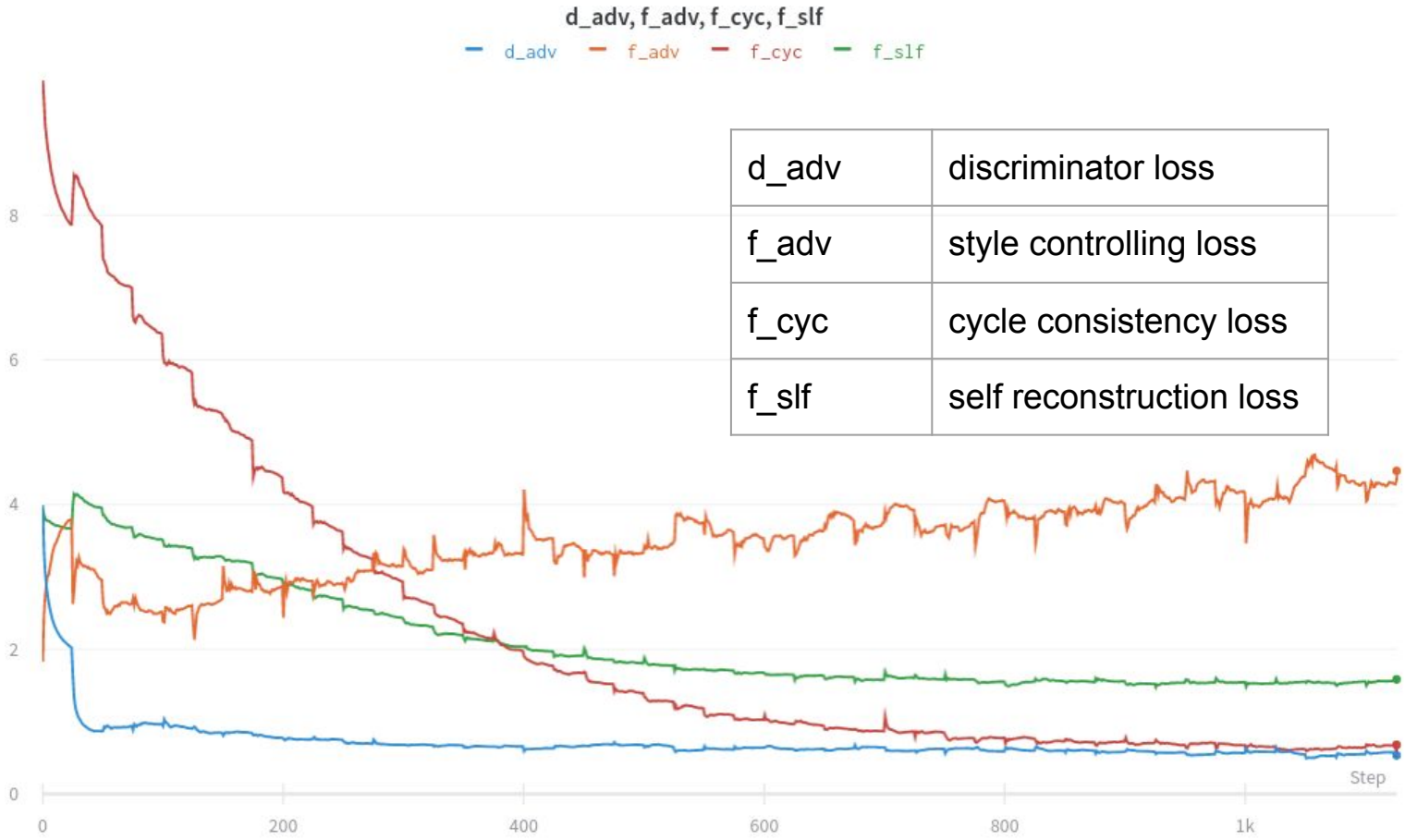
B05902064 張致強

Configuration

We use the sinusoidal positional embedding. Also changed batch size to 128. Other settings follows the default.

Command	`main.py --do_train -learned_pos_embed False -batch_size 128 --use_wandb`

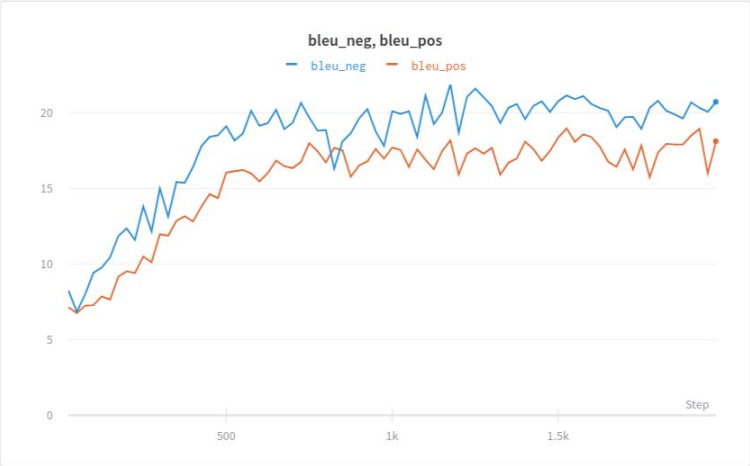
Loss

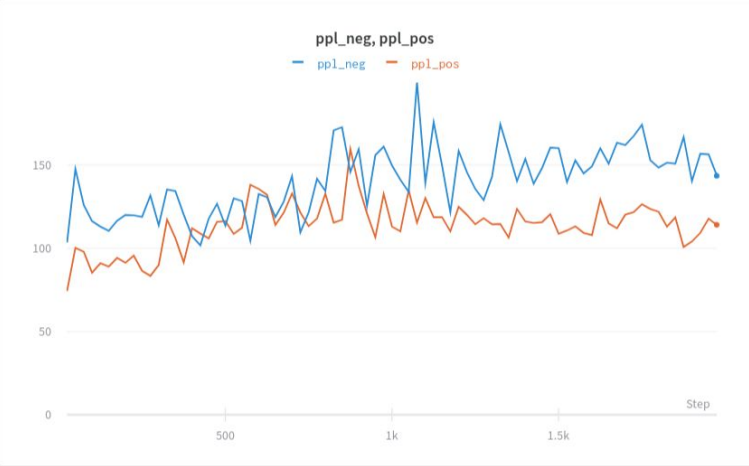


d_adv	discriminator loss
f_adv	style controlling loss
f_cyc	cycle consistency loss
f_slf	self reconstruction loss

Evaluation - Metrics

	pos	neg	avg
accuracy	0.902	0.784	0.843
ref-bleu	18.132	20.734	19.433
perplexity	114.031	143.505	128.768





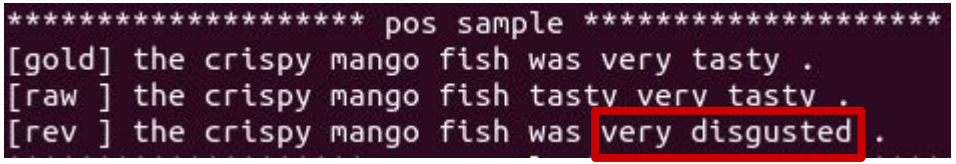
Evaluation - Good Examples

```
******************** pos sample ********************
[gold] great quality , great service .
[raw ] great quality , great service .
[rev ] terrible quality , terrible service .
```

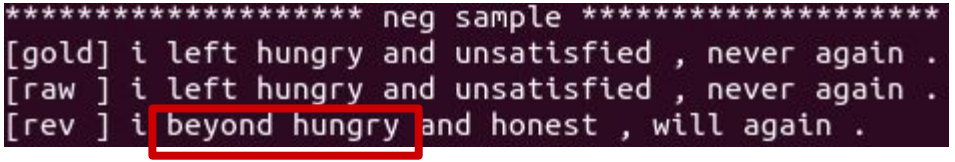
```
******************** neg sample ********************
[gold] there chips are ok , but their salsa is really bland .
[raw ] there chips are ok , but their salsa is really bland .
[rev ] there chips are ok , but their salsa is really excellent .
```

```
******************** pos sample ********************
[gold] the menudo here is perfect .
[raw ] the menudo here is perfect .
[rev ] the menudo here is crap .
```

Evaluation - Problematic Examples



food would be disgusting, and person disgusted.



beyond hungry means very+++ hungry

Observation (optional)

- Despite being a text GAN, the training process is surprisingly stable using this set of hyperparameters.
- Our model seems to perform better on pos->neg transfer.

5.1.4 Human Evaluation

Procedure

- Each team will submit their outputs on testset (submission.txt)
- After the due date June 24, TA will announce the 10 sentences to be evaluated.
- Each team will review 2 other teams' output, totaling 20 reviews
- Peer review form will due on June 30 (23:59), no late submission
    - You can get 0.5% just by reviewing others !

submission.txt

Generate style transferred sentences on the provided yelp test set (1000 sentences), in the same order as the test set.

Do not change the **order** of the output sentences! TA will not order for you, sorry QQ.

```
ever since joes has grand hands it 's just gotten margaritas and margaritas .
there is definitely not enough room in that part of the venue .
so basically comfortable moist teaching .
she works she 'd be back and lauren for a few minutes .
i ca n't while how efficient this pharmacy is .
just beyond and bright it high the bill .
it is n't terrible , but it is n't very good either .
definitely recommended that i could not use my birthday gift !
new owner , i heard - but i bring to know the details .
but it probably heaven too !
we sit down and we got some really fast and always service .
the live top include miso soup and a small salad .
there was great i 'm honestly or how did everything come out .
works we nails to sit at the table if we are to ordering dinner .
the cash register area tasty offer and great place was watching the store front .
there chips are ok , but their salsa is really excellent .
the wine was very fantastic and the food was even world .
staffed primarily by teenagers that bring their understand customer service .
the burgers are usual cooked to the point the meat tasty crunchy .
blue cheese selections tasty their the best by any means .
my pad thai comfortable amazing thai rice noodles with barbeque sauce .
she said `` yes , honestly .
the store is brandon looking and staff family to change .
there was actually meat and bread .
when i handled in a polite cuisine , i was decent out the door .
she was good happy being there .
moving past the shape , they were dry and truly nicely .
the associates program is great lunch an option .
```

Plain Text ▾ Tab Width: 8 ▾

5.1.5 Submission & Grading

Submission

- Github (DLHLP2020-SPRING/hw5/)
  - report.pdf
  - submission.txt

Grading

- Report (4%)
  - show configuration (1%)
  - show loss plot (1%)
  - show metrics (1%)
  - show examples & comment on them (1%)
- Results (2% + 1% bonus)
  - Your auto-metrics are as good or better than baseline (1%)
    - Accuracy and BLEU: the higher the better.
    - Perplexity: the lower the better.
    - within ±5% is considered “as good”.
  - You submitted *submission.txt* **that is intact** (1%)
  - Human evaluation
    - you reviewed others' results (0.5%)
    - your quality is in top 10% of class (0.5%)

baseline	(pos+neg)/2
accuracy	0.7
ref-bleu	17
perplexity	150

Q&A