ACL 2022

22ND – 27TH MAY | 60TH MEETING | DUBLIN

# So Different Yet So Alike! Constrained Unsupervised Text Style Transfer


*Abhinav Ramesh Kashyap*, Devamanyu Hazarika*, Min-Yen Kan, Roger Zimmermann, Soujanya Poria*

NUS National University of Singapore

School of Computing

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

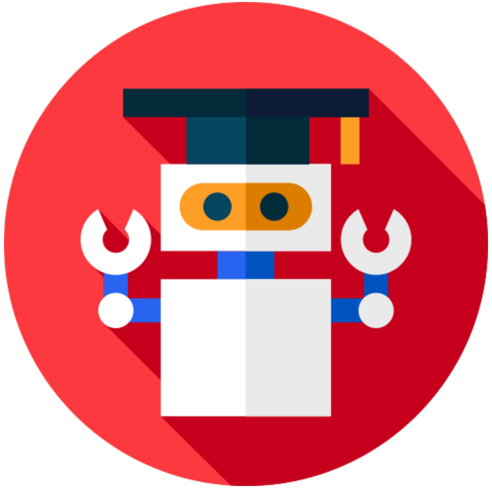
## *Source Domain*

$\mathcal{S}$

## *Target Domain*

$\mathcal{T}$

***Welcome to ACL 2022***



***Bienvenue à l’ACL, 2022***

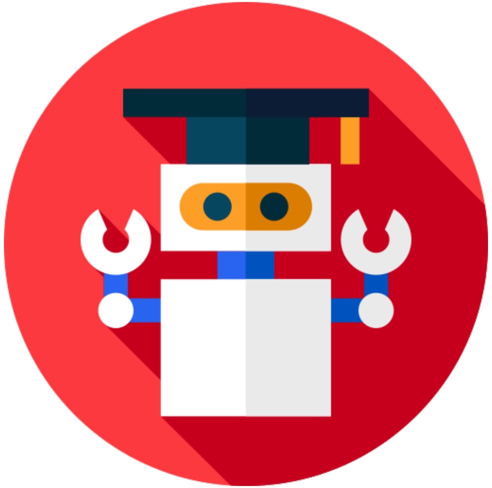
## Source Domain

$\mathcal{S}$

## Target Domain

$\mathcal{T}$

***Just open the door!***



***Could you please open the door for me***

## Source Domain

$\mathcal{S}$

## Target Domain

$\mathcal{T}$

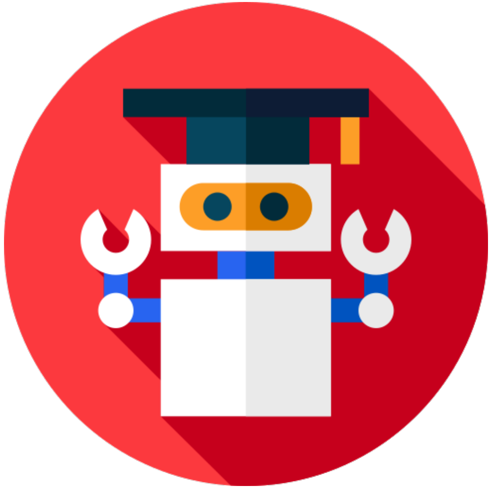


Image from toonify.photos

# Definition of Text Style Transfer [1]

$x$ ***Just open the door!***

$a$ formal

$p(x'|x, a)$

$x'$ ***Could you please open the door***

$a'$ informal

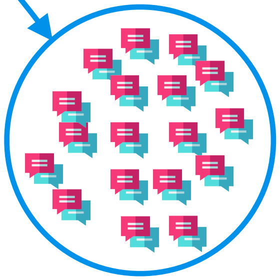
# Data-Driven Definition of Style [1]

Link an attribute
the corpus

*Topic*

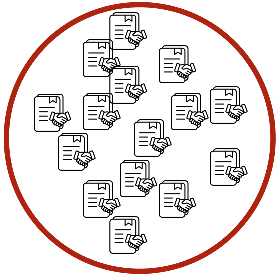
*Content*

*Meta Information*



***Style 1***

$\mathcal{S}$

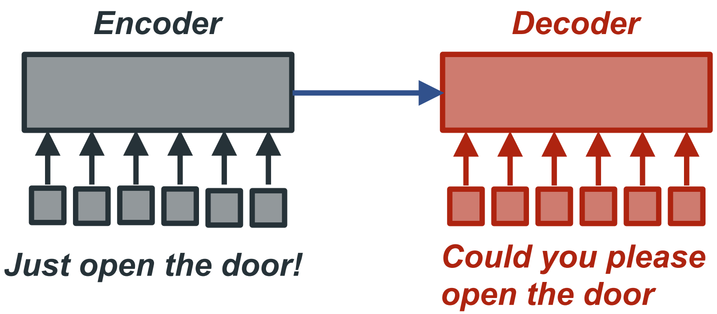


***Style 2***

$\mathcal{T}$

*[1]: Di Jin et al, Deep Learning For Text Style Transfer, Computational Linguistics Journal 2022*

# *Supervised Method*

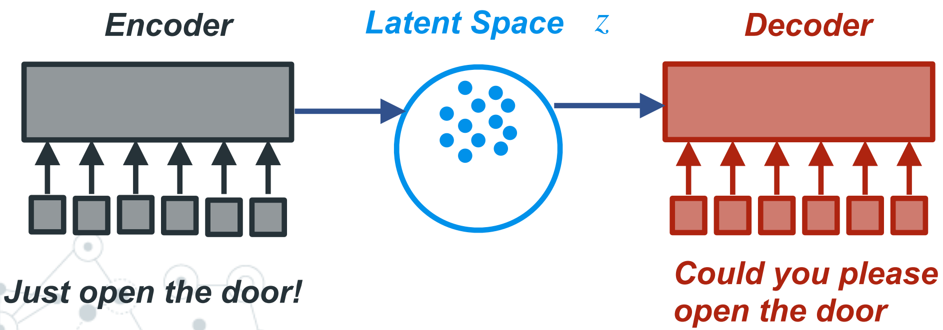



Requires parallel data

Hard to obtain and not scalable

Sequence to Sequence Neural Network Models

# *Unsupervised Method*




Does not require parallel data

Uses ***Data-Driven Definition of Style***

Manipulate the ***latent space*** to disentangle content and style*

** Major approaches disentangle the style and content. There are other methods that do not disentangle*

## DIALOG

I would like to book a restaurant for 2 at 7:00pm

I would like to book a movie for 2 at 7:00pm

**Restaurant** → **Movie**

## TWITTER

Trump loses 2022 elections

Trump is ousted in 2022 elections

**Newspaper** → **Social Media**

DIALOG

I would like to book a restaurant for 2 at 7:00pm

I would like to book a movie for 2 at 7:00pm

Restaurant

Movie

# MAINTAINING CONSTRAINTS IS IMPORTANT BUT IGNORED

Trump loses 2022 ele

Trump is ousted in 2022 elections

Newspaper

Social Media

I really loved Murakami's book

**Text Style Transfer**

Loved the movie

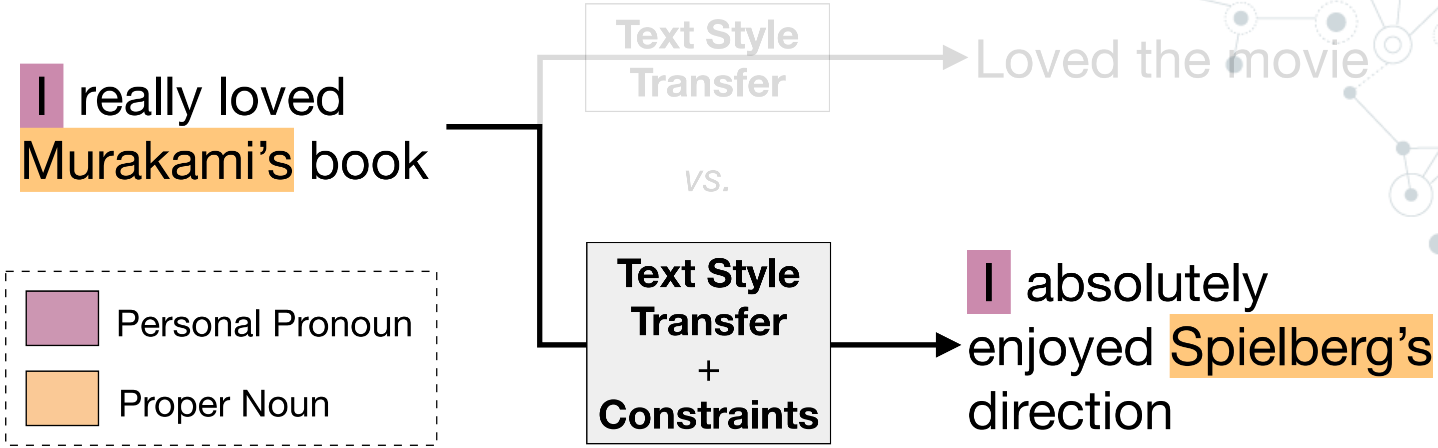
*vs.*

**Text Style Transfer + Constraints**

I absolutely enjoyed Spielberg's direction

Personal Pronoun

Proper Noun

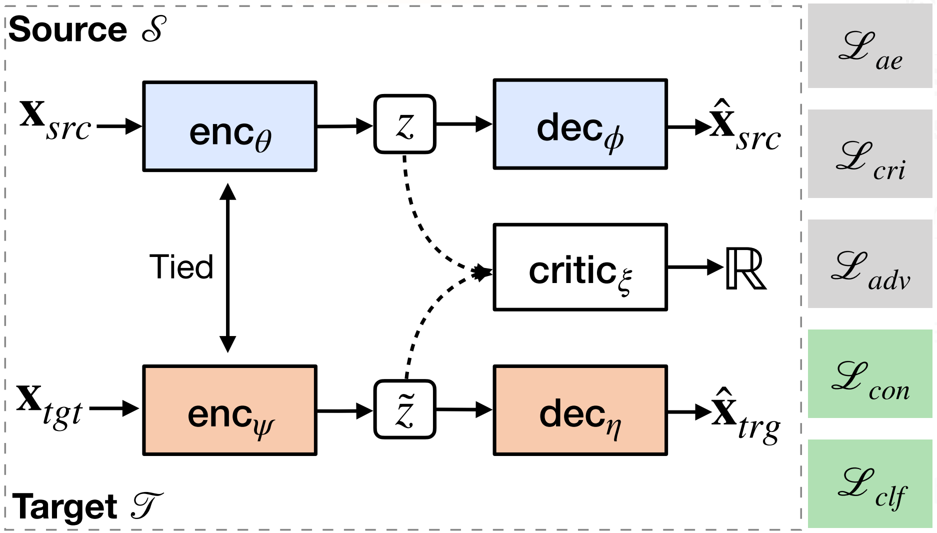



Constraints need to be maintained after transfer

The personal pronoun I is maintained

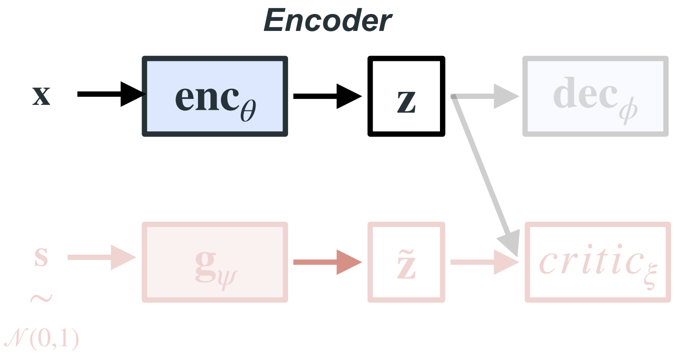
The number of proper nouns are maintained Murakami & Spielberg

# CONTRARAE




- We introduce a GAN-based seq2seq network that explicitly enforces such constraints
- Two ***cooperative losses (the discriminator and the generator reduce the same loss)***
  - ***Contrastive Loss*** - Brings sentences with similar constraints closer together and pushes sentences with different constraints far away
  - ***Classifier loss*** - A discriminative classifier identifies the constraints from latent space

# ADVERSARIALLY REGULARIZED AUTOENCODER (ARAE)




The aim is to generate natural sentences

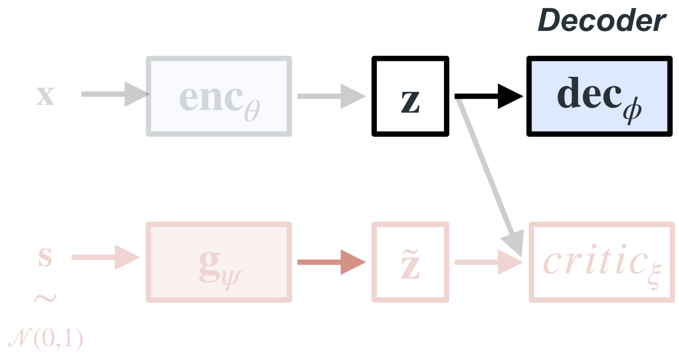
Learn a representation space over a ***prior distribution($\mathcal{N}$)*** that mimics ***real distribution***

Encodes sentences $x$ ***(real)***

$$\mathrm{enc}_\theta : \mathcal{X} \rightarrow \mathcal{Z}$$

$$\mathrm{z} \sim \mathrm{P_z}$$

# ADVERSARIALLY REGULARIZED AUTOENCODER (ARAE)




The aim is to generate natural sentences

Learn a representation space over a ***prior distribution($\mathcal{N}$)*** that mimics ***real distribution***

Encodes sentences $x$ ***(real)***

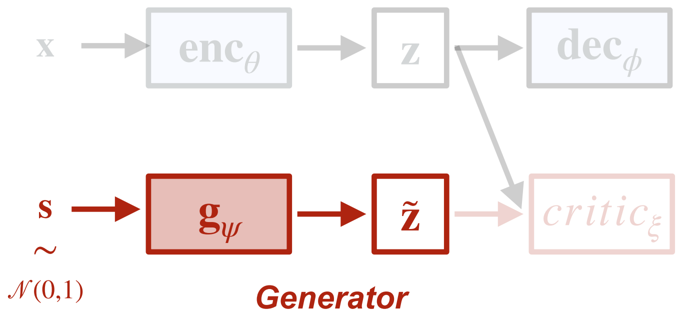
$$\mathrm{enc}_\theta : \mathcal{X} \rightarrow \mathcal{Z}$$

$$z \sim P_z$$

Reconstructs sentences from the latent

$$p_\phi(x \mid z)$$

# ADVERSARIALLY REGULARIZED AUTOENCODER (ARAE)




The aim is to generate natural sentences

Learn a representation space over a ***prior distribution($\mathcal{N}$)*** that mimics ***real distribution***

Encodes sentences $x$ ***(real)***

$$\mathrm{enc}_\theta : \mathcal{X} \rightarrow \mathcal{Z}$$

$$z \sim P_z$$

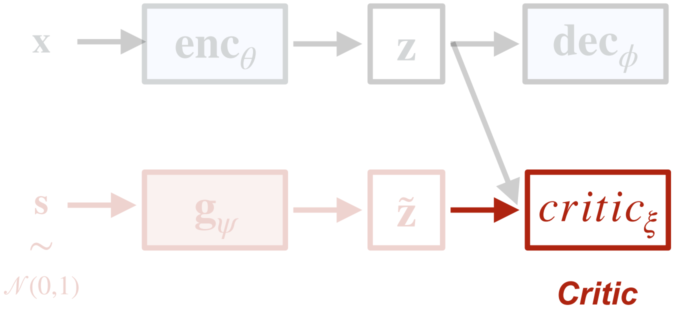
Reconstructs sentences from the latent

$$p_\phi(x \mid z)$$

Maps noise samples to a latent space

$$g_\psi : \mathcal{N}(0,1) \rightarrow \tilde{\mathcal{Z}}$$

# ADVERSARIALLY REGULARIZED AUTOENCODER (ARAE)




The aim is to generate natural sentences

Learn a representation space over a ***prior distribution($\mathcal{N}$)*** that mimics ***real distribution***

Encodes sentences $x$ ***(real)***

$$\mathrm{enc}_\theta : \mathcal{X} \rightarrow \mathcal{Z}$$

$$z \sim P_z$$

Reconstructs sentences from the latent

$$p_\phi(x \mid z)$$

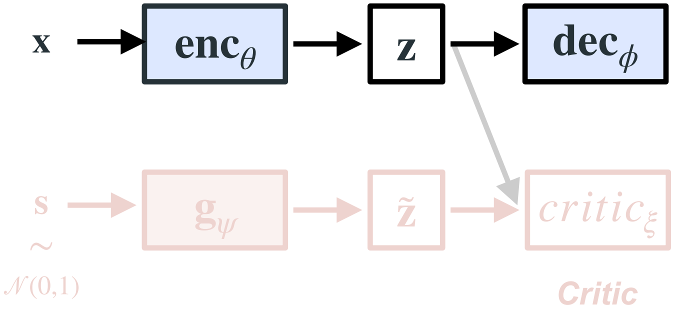
Maps noise samples to a latent space

$$g_\psi : \mathcal{N}(0,1) \rightarrow \tilde{\mathcal{Z}}$$

Distinguishes the real vs generated representations

$$\min_\psi \max_\xi \mathop{\mathbb{E}}_{\mathbf{z}\sim P_z}[crc_\xi(\mathbf{z})] - \mathop{\mathbb{E}}_{\bar{\mathbf{z}}\sim P_{\bar{z}}}[crc_\xi(\bar{\mathbf{z}})]$$

# ADVERSARIALLY REGULARIZED AUTOENCODER (ARAE)



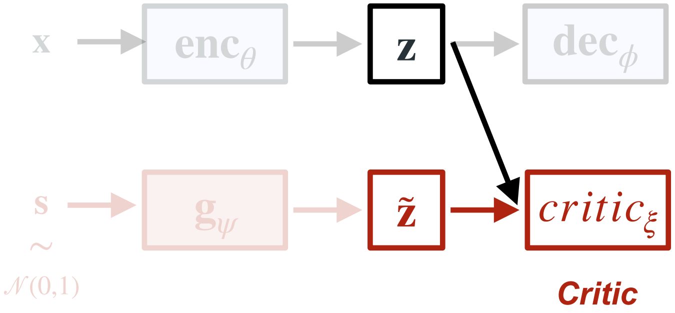

$$\mathcal{L}_{ae}(\theta, \phi) = \mathop{\mathbb{E}}_{\mathbf{z} \sim P_z} [-\log \ p_\phi(\mathbf{x} | \mathbf{z})]$$

*Loss to reconstruct sentences*

*encourage copying behaviour*

*Maintains Semantic Similarity*

# ADVERSARIALLY REGULARIZED AUTOENCODER (ARAE)




$$\mathcal{L}_{ae}(\theta, \phi) = \mathop{\mathbb{E}}_{\mathbf{z} \sim P_z} [-\log \ p_\phi(\mathbf{x} | \mathbf{z})]$$

*Loss to reconstruct sentences*

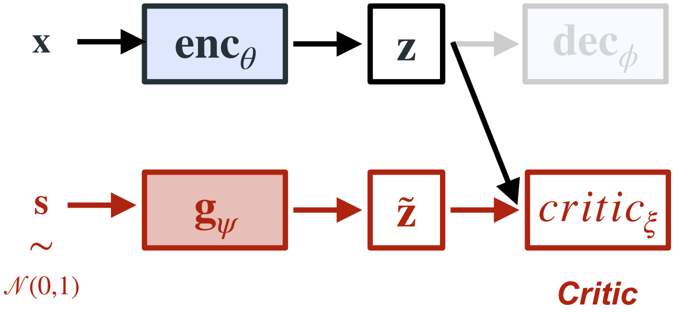
*Encourage copying behaviour*

*Maintains Semantic Similarity*

$$\mathcal{L}_{crc}(\xi) = -\mathop{\mathbb{E}}_{\mathbf{z} \sim P_z} [crc_\xi(\mathbf{z})] + \mathop{\mathbb{E}}_{\bar{\mathbf{z}} \sim P_{\bar{z}}} [crc_\xi(\bar{\mathbf{z}})]$$

*The Critic should succeed in Distinguishing **real** from **fake***

# ADVERSARIALLY REGULARIZED AUTOENCODER (ARAE)




$$\mathcal{L}_{ae}(\theta, \phi) = \mathop{\mathbb{E}}_{\mathbf{z} \sim P_z} [-\log \ p_\phi(\mathbf{x} | \mathbf{z})]$$

*Loss to reconstruct sentences*

*Encourage copying behaviour*

*Maintains Semantic Similarity*

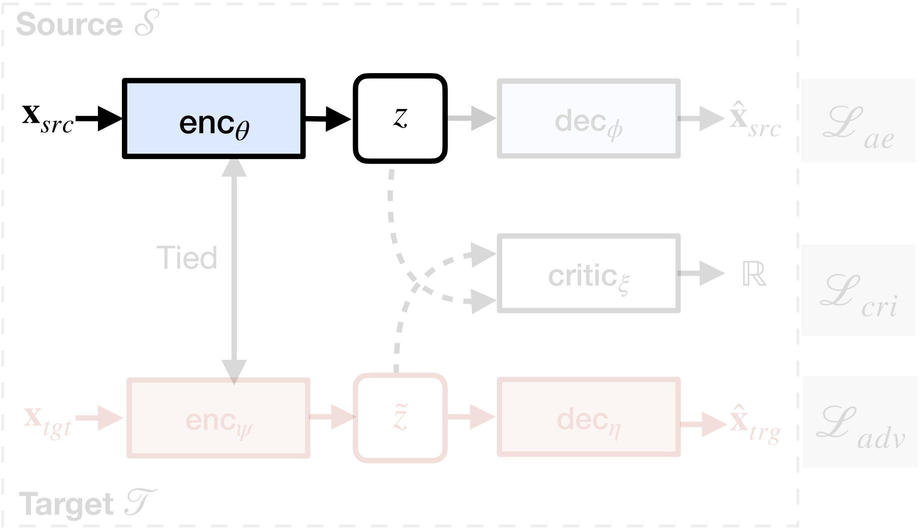
$$\mathcal{L}_{crc}(\xi) = -\mathop{\mathbb{E}}_{\mathbf{z} \sim P_z} [crc_\xi(\mathbf{z})] + \mathop{\mathbb{E}}_{\bar{\mathbf{z}} \sim P_{\bar{z}}} [crc_\xi(\bar{\mathbf{z}})]$$

*The Critic should succeed in Distinguishing **real** from **fake***

$$\mathcal{L}_{adv}(\theta, \psi) = \mathop{\mathbb{E}}_{\mathbf{z} \sim P_z} [crc_\xi(\mathbf{z})] - \mathop{\mathbb{E}}_{\bar{\mathbf{z}} \sim P_{\bar{z}}} [crc_\xi(\bar{\mathbf{z}})]$$

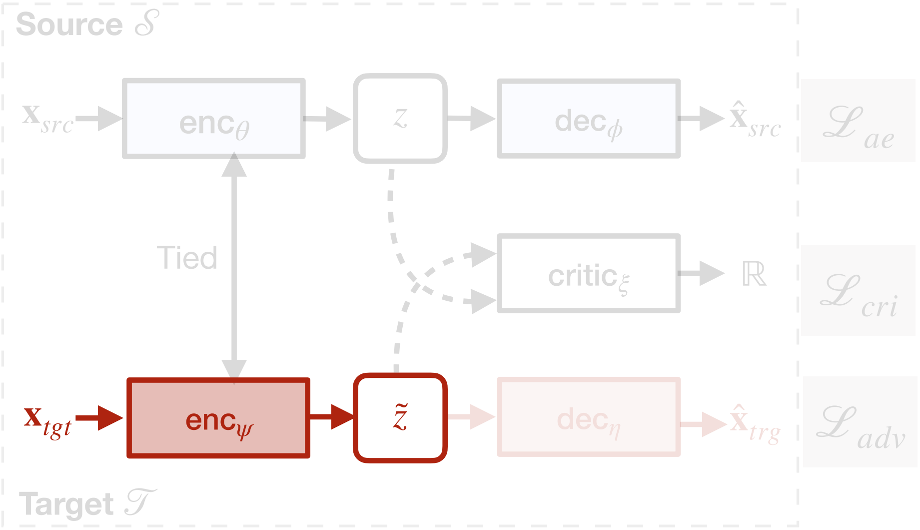
*The **Generator** and the encoder should fool the Critic*

# ADVERSARIALLY REGULARIZED AUTOENCODER ( $ARAE_{seq2seq}$ )



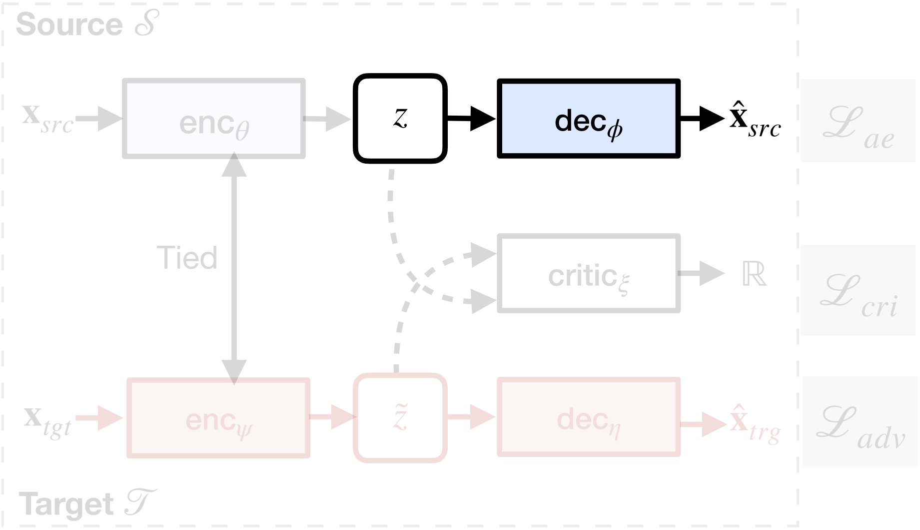

- Encodes sentences from domain $\mathcal{S}$

# ADVERSARIALLY REGULARIZED AUTOENCODER ( ARAE$_{seq2seq}$ )



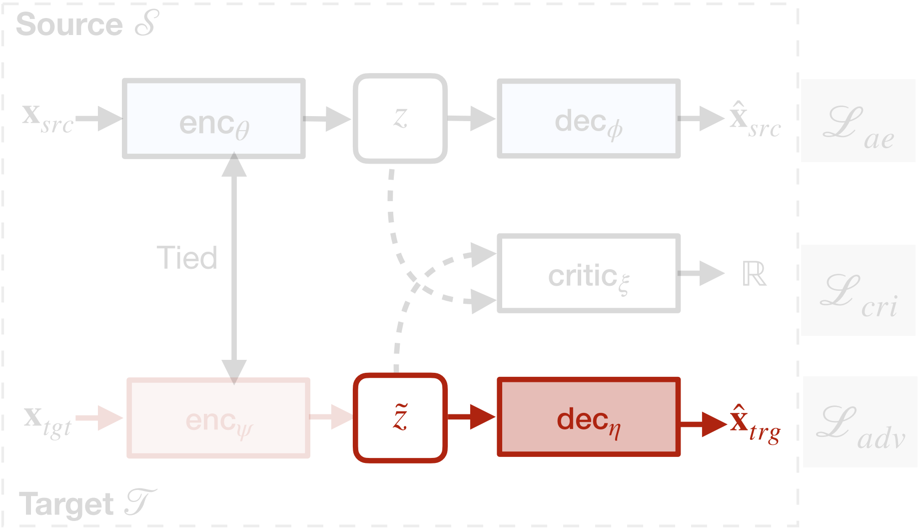

- Encodes sentences from domain $\mathcal{S}$
- Encodes sentences from domain $\mathcal{T}$

# ADVERSARIALLY REGULARIZED AUTOENCODER ( $ARAE_{seq2seq}$ )



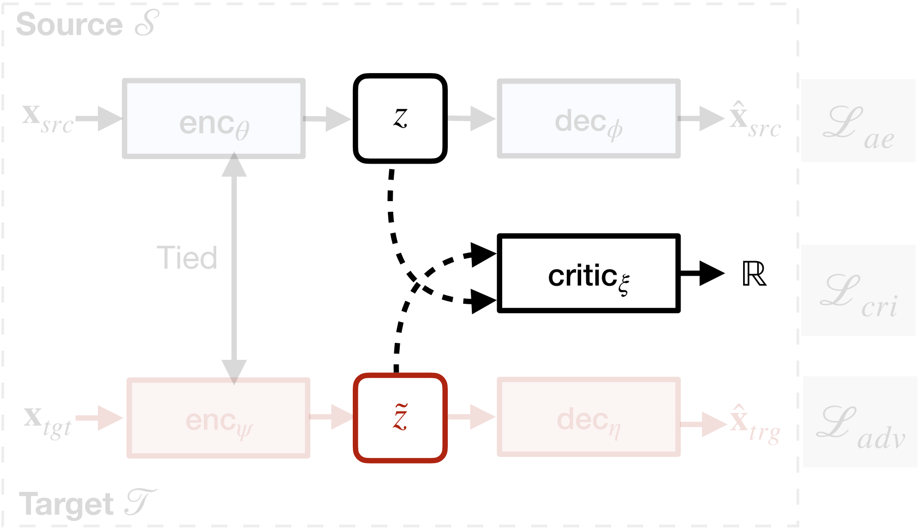

- Encodes sentences from domain $\mathscr{S}$
- Encodes sentences from domain $\mathscr{T}$
- Decodes sentence into domain $\mathscr{S}$

# ADVERSARIALLY REGULARIZED AUTOENCODER ( ARAE$_{seq2seq}$ )



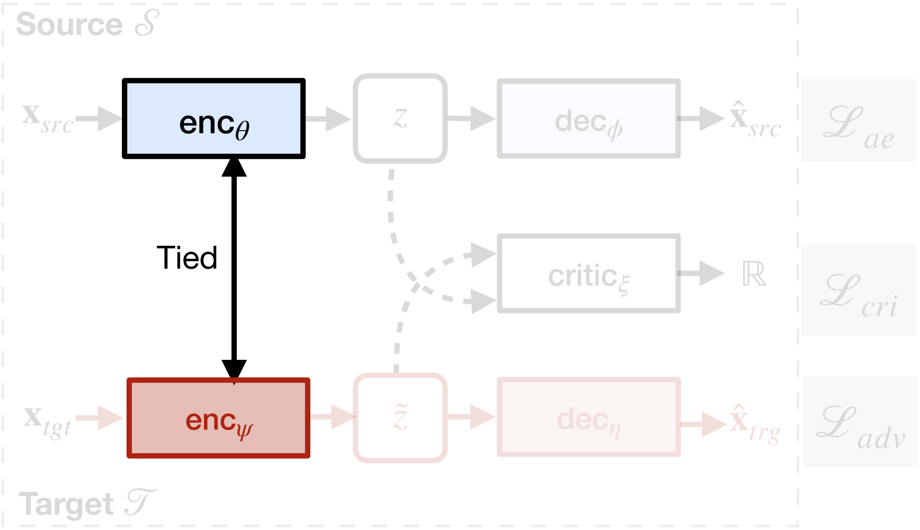

- Encodes sentences from domain $\mathscr{S}$
- Encodes sentences from domain $\mathscr{T}$
- Decodes sentence into domain $\mathscr{S}$
- Decodes sentences into domain $\mathscr{T}$

# ADVERSARIALLY REGULARIZED AUTOENCODER ( $ARAE_{seq2seq}$ )



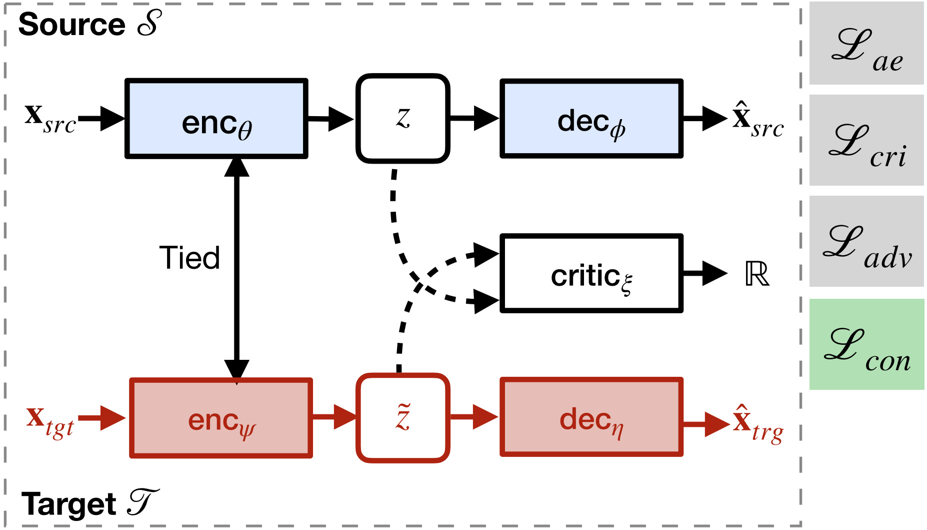

- Encodes sentences from domain $\mathscr{S}$
- Encodes sentences from domain $\mathscr{T}$
- Decodes sentence into domain $\mathscr{S}$
- Decodes sentences into domain $\mathscr{T}$
- Critic distinguishes between $\mathscr{S}$ and $\mathscr{T}$

# *ADVERSARIALLY REGULARIZED AUTOENCODER ( ARAE$_{seq2seq}$ )*




- Encodes sentences from domain $\mathscr{S}$
- Encodes sentences from domain $\mathscr{T}$
- Decodes sentence into domain $\mathscr{S}$
- Decodes sentences into domain $\mathscr{T}$
- Critic distinguishes between $\mathscr{S}$ and $\mathscr{T}$
- We tie the source and the target encoders to encourage them to learn domain invariant representations

# ADVERSARIALLY REGULARIZED AUTOENCODER ( $ARAE_{seq2seq}$ )




$$\mathscr{L}_{con}(\theta, \psi, \xi) = -\frac{1}{|P|} \log \left( \sum_{j=1}^{P} \frac{e^{(\mathbf{z}_i \cdot \mathbf{z}_j)}}{\sum_{k=1}^{B \setminus \{i\}} e^{(\mathbf{z}_i \cdot \mathbf{z}_k)}} \right)$$

*Given a sentence $s \in Src$*

*Mine P sentences each from $Src, Trg$*

*All other sentences in the batch are negatives*

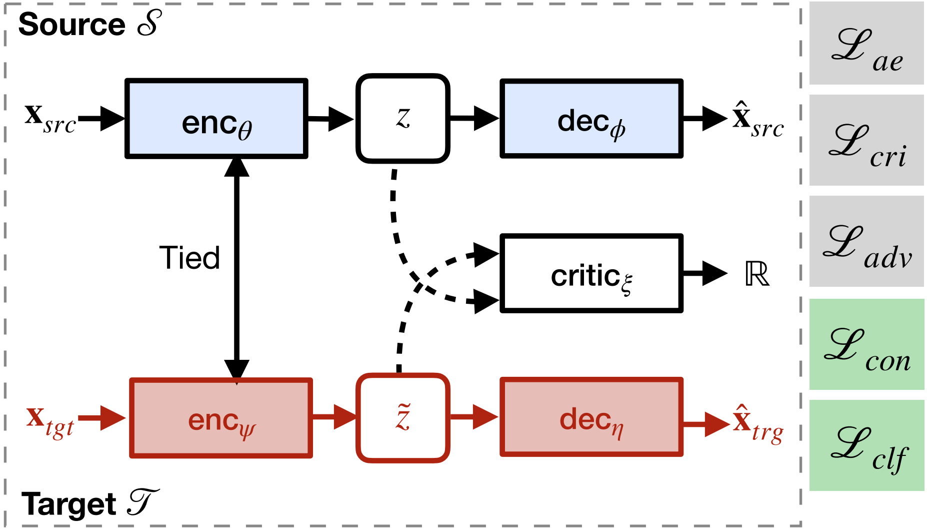
*We add it to both the encoder and the critic*

*$\mathbf{z}$ are representations from the encoders or the last layer of the critic*

*Similar Ideas in Kang et al, 2020*

*Kang et al, 2020 - Contragan: Contrastive Learning for Conditional Image Generation, NEURIPS*

# ADVERSARIALLY REGULARIZED AUTOENCODER ( $ARAE_{seq2seq}$ )




$$\mathscr{L}_{clf}(\theta, \phi, \xi, \delta) = -\sum_{c=1}^{|\mathscr{C}|} \log \left( \sigma\left(l_c\right)^{y_c} \left(1 - \sigma\left(l_c\right)\right)^{1-y_c} \right)$$

*It might be hard to mine positive and negative instances*

*We encourage the encoders and the critic to instead Reduce a classification loss*

$|\mathscr{C}|$ *Number of constraints per sentences*

$l_c$ *Logits for the class c*

$\sigma(.)$ *Sigmoid function*

*Similar Ideas in ACGAN (Odena et al, 2017)*

*Odena et al 2017: Conditional Image Synthesis with Auxiliary Classifier GANs, ICML*

# *DATASETS*

| | |
|---|---|
| **YELP** | *Business reviews labelled as either positive and negative* |
| **IMDB** | *Movie reviews labelled as either positive or negative* |
| **POLITICAL** | *Facebook posts labelled with either the Republican or Democratic slant* |

# *METRICS*

| | |
|---|---|
| **ACC** | How well the sentence adheres to target domain ? |
| **FL** | How fluent is the sentence ? |
| **SIM** | How semantically similar is the sentence to the source domain? |
| **AGG** | Joint Metric at instance level |

# OVERALL RESULTS

| | | Yelp | | | | IMDB | | | | POLITICAL | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Sampling | ACC | FL | SIM | AGG | ACC | FL | SIM | AGG | ACC | FL | SIM | AGG |
| DRG | Greedy | 67.4 | 54.5 | 43.6 | 16.7 | 56.5 | 44.3 | 54.1 | 14.4 | 61.3 | 35.7 | 38.7 | 8.8 |
| ARAE | Greedy | 93.1 | 67.9 | 31.2 | 19.8 | 95.0 | 76.3 | 26.4 | 19.9 | 63.0 | 72.1 | 17.3 | 11.0 |
| ARAE $_{seq2seq}$ +CLF +CONTRA | Greedy | 89.3 | 69.2 | 32.9 | **20.6** | 97.8 | 84.0 | 33.5 | **28.1** | 99.0 | 56.8 | 41.8 | **24.9** |
| | nucleus (p=0.6) | 89.4 | 68.6 | 32.8 | 20.4 | 97.1 | 82.6 | 33.6 | 27.4 | 99.0 | 56.0 | 41.6 | 24.4 |

Compared to DRG(Li et al.) and ARAE (Zhao et al.), our method has better aggregate for 3 different datasets

Regularizing the latent space, brings advantages to the overall quality of generated sentences

*Li et al., Delete, Retrieve, Generate: a Simple Approach to Sentiment and Style Transfer, NAACL*

*Zhao et al., Adversarially Regularized Autoencoders, ICML*

# *REMOVING LOSS ON GENERATOR AND CRITIC*

| Model | ACC | FL | SIM | AGG |
|---|---|---|---|---|
| $\text{ARAE}_{seq2seq}$ + CLF | 95.0 | 83.2 | **34.2** | **27.5** |
| -generator | **96.2** | **87.2** | 31.3 | 26.7 |
| -critic | 94.9 | 84.4 | 30.8 | 25.5 |

Adding the ***CLF*** loss improves the over all **AGG** score

Mostly improves the ***SIM*** score

# REMOVING LOSS ON GENERATOR AND CRITIC

| Model | ACC | FL | SIM | AGG |
|---|---|---|---|---|
| ARAE $_{seq2seq}$ + CLF | 95.0 | 83.2 | **34.2** | **27.5** |
| -generator | **96.2** | **87.2** | 31.3 | 26.7 |
| -critic | 94.9 | 84.4 | 30.8 | 25.5 |

Adding the ***CLF*** loss improves the over all **AGG** score

Mostly improves the ***SIM*** score

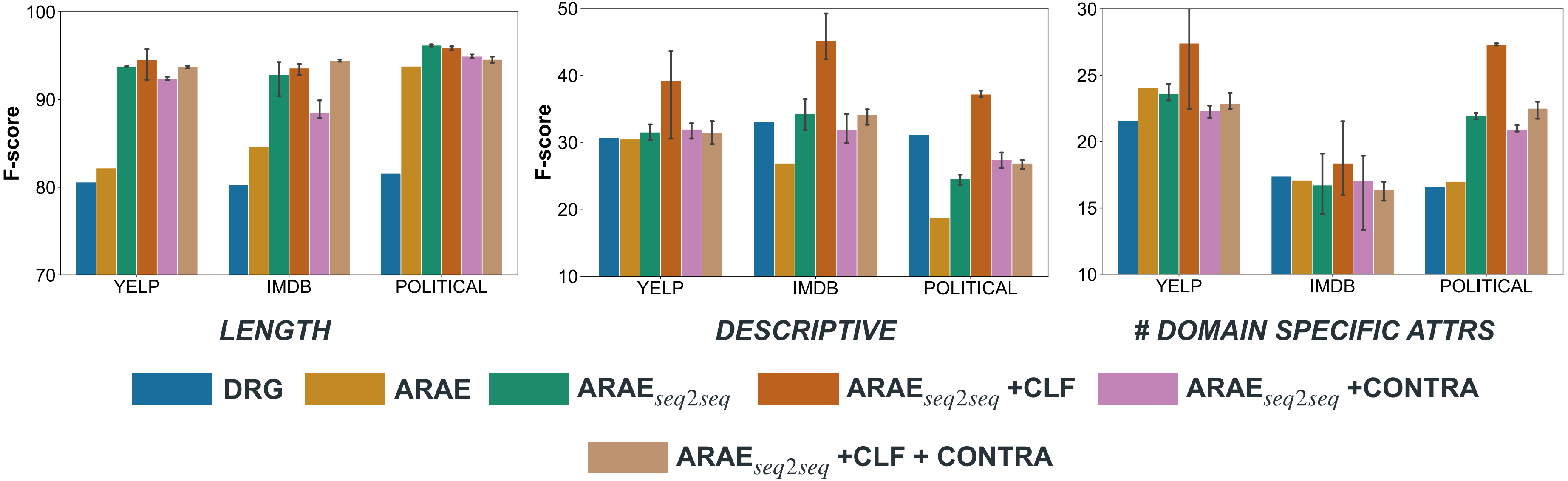
| Model | ACC | FL | SIM | AGG |
|---|---|---|---|---|
| ARAE $_{seq2seq}$ + CONTRA | **96.1** | **80.6** | 36.0 | 28.6 |
| -generator | 93.5 | 78.8 | 34.0 | 26.0 |
| -critic | 90.1 | 67.8 | 39.5 | 24.9 |

Adding the ***CONTRA*** loss improves the over all **AGG** score

Adding the ***CONTRA*** loss improves the ***ACC*** and ***FL*** score

# CLF and CONTRA losses are complementary and necessary on CRITIC and GENERATOR to improve the AGG score

# MAINTAINING CONSTRAINTS




Adding Cooperative losses helps in maintaining constraints

***LENGTH*** is an easier constraint to maintain for most methods

Syntactic attributes like ***DESCRIPTIVENESS*** and ***#DOMAIN SPECIFIC ATTRS*** are harder to maintain

Improvements in ***AGG*** score does not mean, constraints are maintained.

# CONCLUSION

Unsupervised Style Transfer Methods do not explicit define what is maintained between two domains

We introduced two cooperative losses to ARAE to further regularize the latent space

We improve the general quality of translating sentences from one domain to another

*In addition, we maintain the constraints between the domains in a better manner*


Abhinav Ramesh Kashyap
abhinavkashyap.io

Devamanyu Hazarika
devamanyu.com

Min-Yen Kan
www.comp.nus.edu.sg/~kanmy

Roger Zimmermann
www.comp.nus.edu.sg/~rogerz

Soujanya Poria
sporia.info/