Text Style Transfer

Presenter: Sasha Lew

October 14, 2021

Presentation Outline

Task Overview

Paper #1: Style Transfer Through Back-translation

Paper #2: A Dual Reinforcement Learning Framework for Unsupervised Text Style Transfer

Papers #3-4: Style Transformer, A Recipe for Arbitrary Text Style Transfer with Large Language Models

Resources and Conclusion

What is style transfer?

Natural language has semantic ***content*** and controlled ***style***

Example text: “I am happy”

Changing the content (what is said): “I am tall”

Changing the style (how to say): “I am overjoyed”

Style transfer is the task of transforming text to achieve a specific style while preserving the semantic content.

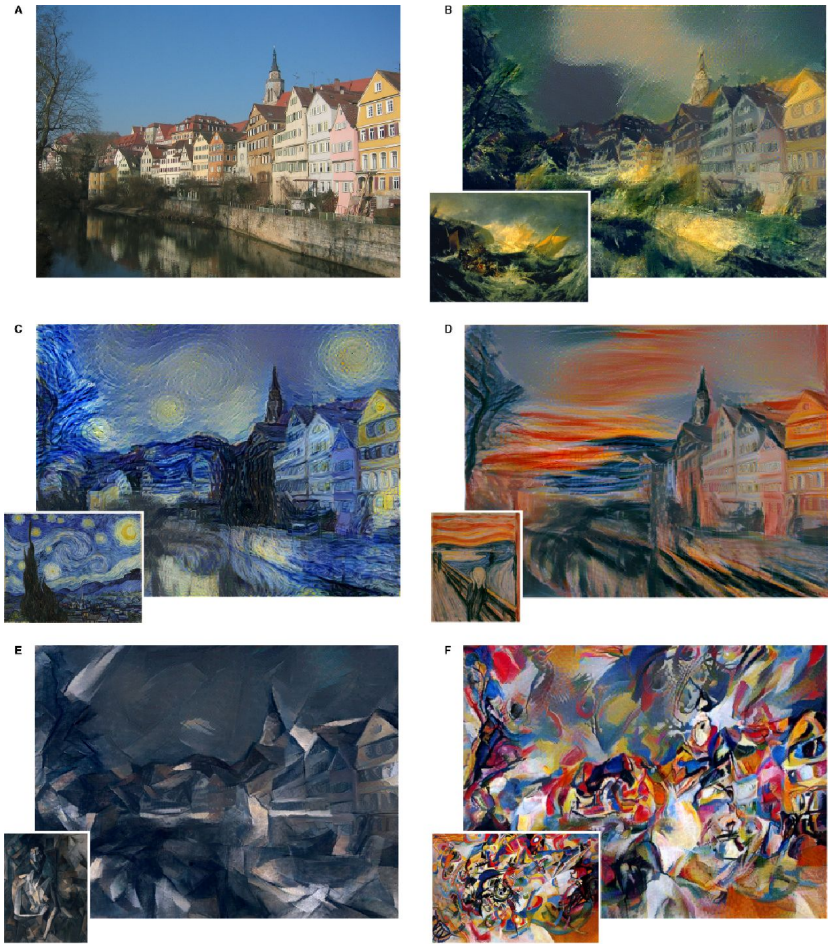
Style Transfer in Visual Domain

Historical background:

Neural style transfer proposed by Gatys et al., 2016: “A neural algorithm of artistic style.”

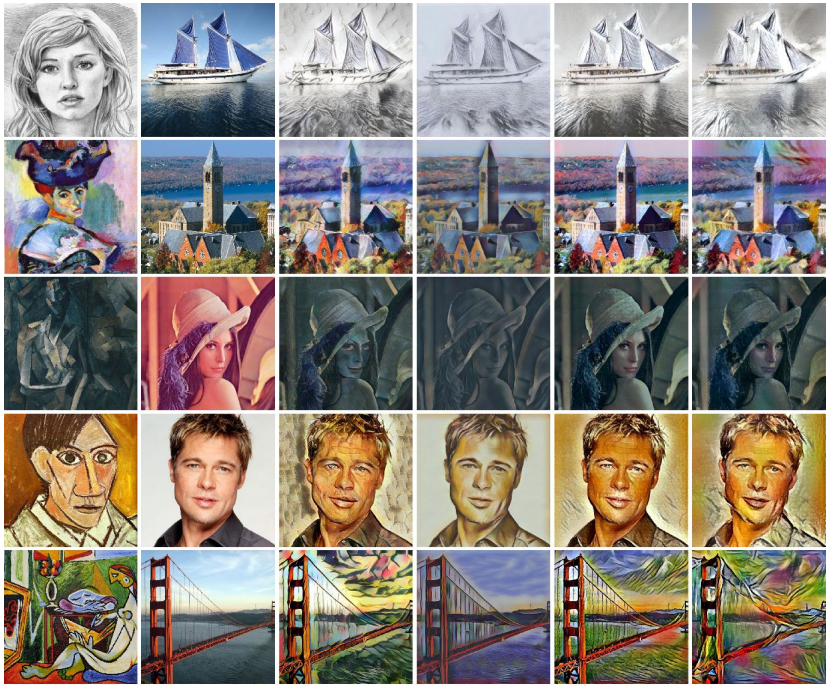
Used GANs “to separate and recombine content and style of arbitrary images”

Example: **content** of the scene stays the same, but **style** of representation is modulated



https://arxiv.org/pdf/1508.06576.pdf

Adaptive Style Transfer



Huang and Belongie, 2017: “Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization
https://arxiv.org/pdf/1703.06868.pdf

Text Style Transfer Applications

- Controlling attributes of generated text
  - “That book sucks” → “That book has some flaws”
- Inverting sentiment
  - “I hate Mozart” → “I love Mozart”
- Debiasing human-written text
  - Reddy and Knight, 2016: “Obfuscating gender in social media writing.” [https://www.aclweb.org/anthology/W16-5603.pdf]
  - Protecting identity-based characteristics inherent in writing
- Automatic editing of existing text for specific quality (e.g., dialect)
  - “I am happy” → “I’m chuffed”

“Style” does not have a monolithic definition.

Text Style Transfer Examples

FORMALITY

Gotta see both sides of the story. *(informal)*
You have to consider both sides of the story. *(formal)*

SENTIMENT

The screen is just the right size. *(positive)*
The screen is too small. *(negative)*

AUTHOR IMITATION

Bring her out to me. *(modern)*
Call her forth to me. *(shakespearean)*

Briakou et al., 2021: "A Review of Human Evaluation for Style Transfer." [https://arxiv.org/pdf/2106.04747.pdf]

Source	Speaker	Input	Output
Romeo & Juliet	Benvolio	He killed your relative, brave Mercutio, and then young Romeo killed him.	he slew thy kinsman , brave mercutio , and then young romeo kill him .
Romeo & Juliet	Romeo	I can read my own fortune in my misery.	i can read mine own fortune in my woes .
Star Wars	Palpatine	If you will not be turned, you will be destroyed!	if you will not be turn 'd , you will be undone !
Star Wars	Luke	Father, please! Help me!	father , i pray you , help me !
The Matrix	Agent Smith	Good bye, Mr. Anderson.	fare you well , good master anderson .
The Matrix	Morpheus	I'm trying to free your mind, Neo. But I can only show you the door. You're the one that has to walk through it.	i 'll to free your mind , neo. but i can but show you the door. you 're the one that hath to tread it .
Raiders of the Lost Ark	Belloq	Good afternoon, Dr. Jones.	well met , dr. jones .
Raiders of the Lost Ark	Jones	I ought to kill you right now.	i should kill thee straight .

Xu et al., 2012: "Paraphrasing for Style." [https://www.aclweb.org/anthology/C12-1177.pdf]

Formalization of Text Style Transfer

A dataset ***X*** = {*x1*, *x2*, …, *xn*} is composed of natural language sentences with the style *s*

Style is the unifying characteristic of dataset ***X***

Style Transfer is the task of generating samples of ***X***, such that they belong to style *s*□, which is the unifying characteristic of a dataset ***Y*** = {*y1*, *y2*, …, *yn*}

Or conversely, generating samples of ***Y*** such that they belong to style *s*

Hence, the task is to generate sentences that fit the desired target style while preserving the meaning of the input sentence

Common Approaches and Challenges

- Treat as sequence-to-sequence problem
  - Xu et al., 2012: “Paraphrasing for style.” [http://aclweb.org/anthology/C12-1177]
  - Translate from Shakespearean English to modern English
  - Needs parallel corpus for source and target styles
- Control or rewrite generated text directly
  - Li et al., 2018: “Delete, retrieve, generate: A simple approach to sentiment and style transfer.” [https://doi.org/10.18653/v1%2FN18-1169]
  - Identify stylistic keywords and generate replacements
  - Style can be implicit
- Learn disentangled latent representation
  - Shen et al., 2017: “Style transfer from non-parallel text by cross-alignment.” [https://arxiv.org/pdf/1705.09655.pdf]
  - Train autoencoder to separate and recombine styles and content
  - Disentanglement is hard

Paper #1: Style Transfer Through Back-translation

ACL 2018 (Melbourne, Australia)
Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, Alan W. Black

Motivation and Key Insights

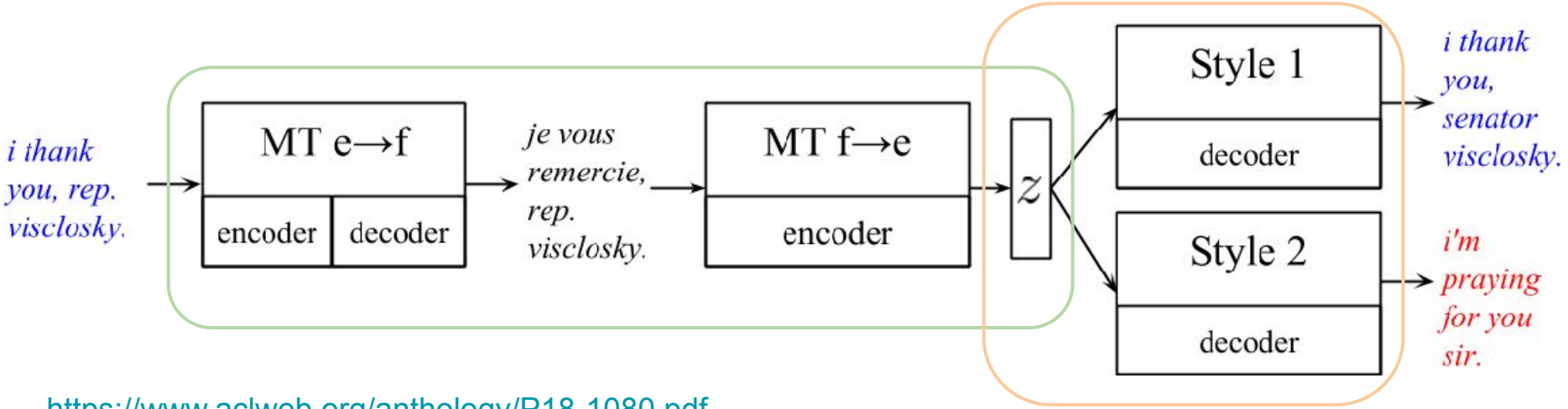
- Learn a latent representation that disentangles content and style…
- … via **back-translation**
- Rabinovich et al., 2016: “Personalized machine translation: Preserving original author traits.”
  - Back-translation strips away author characteristics
  - Could the same technique also remove style from an input sequence, while preserving its semantic content?
- Modular pipeline: separate translation (disentanglement) and style generation components
  - Better control over the latent representation
  - Potentially more explainable and adaptable

Architecture Overview

Back-translation: Input sentence is translated, then returned to source language

Yields latent representation, *z*

Style-specific decoder takes *z* as input, generates re-styled output sequence

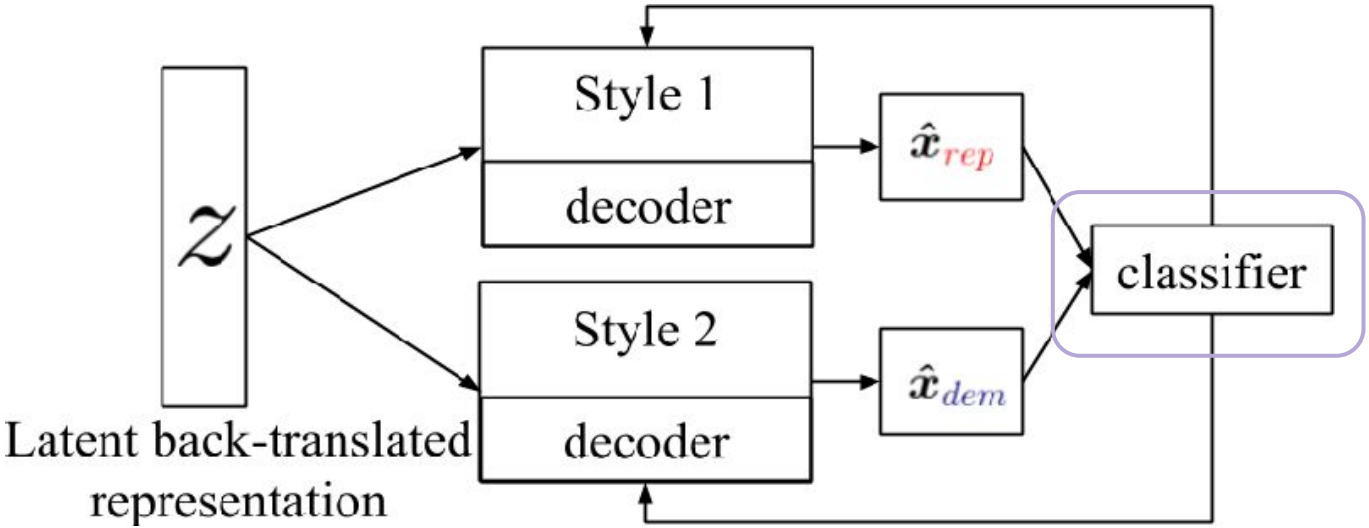


https://www.aclweb.org/anthology/P18-1080.pdf

Training the Style Generators

Pre-train CNN classifier whose predictions contribute to training style generators

Style generators also try to “reconstruct” original sample from *z*



https://www.aclweb.org/anthology/P18-1080.pdf

Style Generator Loss Function

Two competing terms: semantic preservation *versus* style transfer accuracy

Reconstruction loss: cross entropy between original input ***x*** and latent representation ***z***

$$\mathcal{L}_{recon}(\boldsymbol{\theta}_G; \boldsymbol{x}) = \mathbb{E}_{q_E(\boldsymbol{z}|\boldsymbol{x})}[\log p_{gen}(\boldsymbol{x}|\boldsymbol{z})]$$

Classifier loss : minimize mapping same style ***s*** to original input ***x***

$$\mathcal{L}_{class}(\boldsymbol{\theta}_C) = \mathbb{E}_{\boldsymbol{X}}[\log q_C(\boldsymbol{s}|\boldsymbol{x})]$$

Overall loss: sum of reconstruction and classifier losses, with balancing parameter

$$\min_{\theta_{gen}} \mathcal{L}_{gen} \quad = \quad \mathcal{L}_{recon} + \lambda_c \mathcal{L}_{class}$$

Experiment Setup

- Style generators and encoders are two-layer, bidirectional LSTM
- Training data for machine translation: Europarl v7, news commentary v10, WMT15 common crawl corpora for French-English
- 3 styles to transfer:
  - Gender: Yelp reviews dataset annotated with binary gender labels (male or female)
    - Reddy and Knight, 2016: “Obfuscating gender in social media writing.”
  - Political slant: Members of Congress Facebook posts + top-level comments, labelled with political party affiliation (Democratic or Republican)
    - Voigt et al., 2018: “RtGender: A corpus for studying differential responses to gender.”
  - Sentiment: Yelp reviews dataset where ≥ 3 is labelled positive and < 3 is labelled negative
    - Shen et al., 2017: “Style transfer from non-parallel text by cross-alignment.”
      - *Also implements Cross-alignment Auto-Encoder (**CAE**) for baseline method*

Evaluation: Style Transfer Accuracy

- Pre-trained style classifier
  - Trained separately from style generator classifier
- Transfer, then test classification accuracy for opposite label
- Automatic evaluation
- Proposed method is much more accurate than (or comparable to) baseline

Experiment	CAE	BST
Gender	**60.40**	57.04
Political slant	75.82	**88.01**
Sentiment	80.43	**87.22**

Evaluation: Semantic Preservation

- Shouldn't just use BLEU score
- Style meaning often requires changing key words
- Human evaluation
- Pairwise comparison
  - Show both methods' generated sentences
  - Ask which sentence maintains same intent as the original sentence

Experiment	CAE	No Pref.	BST
Gender	15.23	41.36	**43.41**
Political slant	14.55	**45.90**	39.55
Sentiment	35.91	**40.91**	23.18

- *No Preference* wins; models do not meet human expectations
- Not on table: BST is significantly preferred for longer sentences

Evaluation: Fluency

- Human evaluation
- Select 60 generated sentences per method, Likert scale 1-4
- BST preferred overall, but sentences score low on fluency
- Example sentences vary in quality...

Experiment	CAE	BST
Gender	2.42	**2.81**
Political slant	2.79	**2.87**
Sentiment	3.09	**3.18**
Overall	2.70	**2.91**
Overall Short	3.05	**3.11**
Overall Long	2.18	**2.62**

Input Sentence	CAE	BST
male → female		
my wife ordered country fried steak and eggs.	*i got ta get the chicken breast .*	*my husband ordered the chicken salad and the fries.*
great place to visit and maybe find that one rare item you just have never seen or can not find anywhere else.	*we could n't go back and i would be able to get me to get me.*	*great place to go back and try a lot of which you ' ve never had to try or could not have been able to get some of the best.*
the place is small but cosy and very clean.	*the staff and the place is very nice.*	*the place is great but very clean and very friendly.*

Summary and Discussion

- Style transfer without parallel text
- Leverage neural machine translation to obtain latent, disentangled representation
- Three style transfer tasks, three evaluation axes
  - In general, Back-translation for Style Transfer (BST) is comparable to previous SotA (Cross-aligned Auto-Encoder)
- Example sentences demonstrate the challenging nature of the task
  - Effective separation of content and style is especially difficult

Discussion Questions

What are the potential benefits and drawbacks of using different languages for generating the latent representation?

Would you expect several “layers” of back-translation to improve disentanglement of content and style? Could there be any disadvantages with this approach?

How would you modify the training framework to accommodate multi-class styles?

Should the NMT component of the proposed pipeline consider context?

Paper #2: A Dual Reinforcement Learning Framework for Unsupervised Text Style Transfer

IJCAI 2019 (Macao, China)

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, Zhifang Sui

Motivation and Key Insights

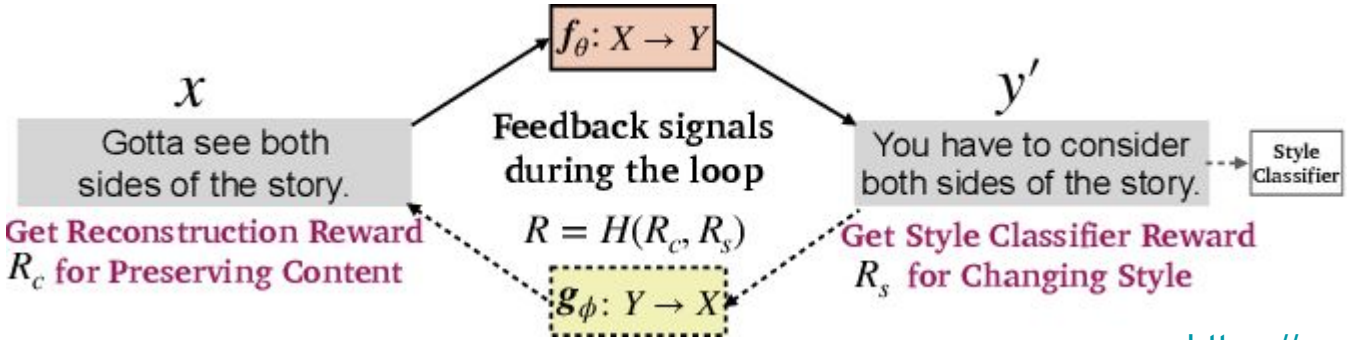
- Learning a latent disentangled representation is hard
  - Style is not completely independent from content
    - “The cat is ugly” → “The cat is pretty” NOT “The dog is ugly”
  - Style can be implicit, by discourse convention
    - “I can’t believe it’s not butter!”
- Lack of parallel corpora makes learning latent representation even harder
- Key idea: train forward and backward models in a **closed loop** with feedback signals based on transfer quality (DualRL)
- Incorporate feedback signals into a **reward function**
- Unsupervised style transfer without parallel data or latent representation

Architecture Overview

Forward model ($f\theta$) transfers sequence x to style-modified y

Backward model ($g\phi$) transfers sequence y to style-modified x

Closed loop framework: passing through forward model rewards transferring style; returning through backward model rewards preserving semantic content



https://arxiv.org/pdf/1905.10060.pdf

Reward Functions

Style Reward: maximize likelihood of classifying generated sequence with target style

$$R_s = P(s_y|\boldsymbol{y'};\varphi)$$

Content Reward: maximize likelihood of reconstructing original sequence *x*, given the backward model *g*

$$R_c = P(\boldsymbol{x}|\boldsymbol{y'};\boldsymbol{\phi})$$

Overall Reward: harmonic mean of style and content rewards

$$R = (1+\beta^2)\frac{R_c \cdot R_s}{(\beta^2 \cdot R_c) + R_s}$$

Training Details

- Reinforcement learning for text generation needs pre-training (“warm starting”)
- For supervised learning, provide parallel data + MLE
  - Ranzato et al., 2016: “Sequence level training with recurrent neural networks.”
- But DualRL is an *unsupervised* framework—so, generate simple “pseudo-parallel” data with MLE (on next slide)
- After pre-training, alternately train $f\theta$ and $g\phi$
  - Sample sentence from dataset
  - Generate sentence in opposite style
  - Compute reward functions
  - Update model
  - Repeat procedure with opposite style dataset sample, on the other model

Annealing Pseudo Teacher-forcing

- Generate pseudo-parallel data to pre-train the RL framework
- Teacher-forcing: feed parallel data into model, then use RL + MLE to update the model (either MLE loss or objective)
- Experimentally, better to exploit the latest version of the forward and backward models to generate better pseudo-parallel data
- Danger of exposure bias and distribution shift when bootstrapping with model + MLE…
- Annealing: decay the frequency of updating models with MLE, so that pseudo-parallel data is incorporated less, later

Experimental Setup

- Forward and backward models implemented as LSTM encoder-decoders
- 2 styles to transfer:
  - Sentiment: Yelp restaurant reviews dataset
    - Train/validate/test splits same as Li et al., 2018: “Delete, retrieve, generate: a simple approach to sentiment and style transfer.”
  - Formality: GYAFC dataset (discussed later in presentation)
    - parallel corpus, but alignments not used in training
    - Family and Relationships domain
- Several baselines, including BST (BackTrans) and CAE (CrossAlign)
- Automatic and human evaluation
  - Style transfer accuracy, semantic preservation, fluency (human evaluation only)

Automatic Evaluation

Style transfer accuracy measured with style classifier

Semantic preservation measured with BLEU score

DualRL has best geometric and harmonic means (scoring composite)

Human references still substantially better than any method in mean scores

	YELP				GYAFC			
	ACC	BLEU	**G2**	**H2**	ACC	BLEU	**G2**	**H2**
Retri [Li *et al.*, 2018]	**96.0**	2.9	16.7	5.7	**91.3**	0.4	6.0	0.8
BackTrans [Tsvetkov *et al.*, 2018]	95.4	5.0	21.9	9.6	70.2	0.9	8.1	1.9
StyleEmbed [Fu *et al.*, 2018]	8.7	42.3	19.2	14.4	22.7	7.9	13.4	11.7
MultiDec [Fu *et al.*, 2018]	50.2	27.9	37.4	35.9	17.9	12.3	14.8	14.6
CrossAlign [Shen *et al.*, 2017]	75.3	17.9	36.7	28.9	70.5	3.6	15.9	6.8
Unpaired [Xu *et al.*, 2018]	64.9	37.0	49.0	47.1	79.5	2.0	12.6	3.9
Del [Li *et al.*, 2018]	85.3	29.0	49.7	43.3	18.8	29.2	23.4	22.9
DelRetri [Li *et al.*, 2018]	89.0	31.1	52.6	46.1	55.2	21.2	34.2	30.6
Template [Li *et al.*, 2018]	81.8	45.5	61.0	58.5	52.9	35.2	43.1	42.3
UnsuperMT [Zhang *et al.*, 2018b]	95.4	44.5	65.1	60.7	70.8	33.4	48.6	45.4
DualRL	85.6	**55.2**	**68.7**	**67.1**	71.1	**41.9**	**54.6**	**52.7**
Human	74.0	100.0	86.0	85.1	84.3	100.0	91.8	91.5

Human Evaluation

- Three human evaluators with linguistic backgrounds
- Likert scale, 1-5 on style transfer accuracy, content preservation & fluency
- Successful transfer: generated sentence is rated 4 or 5 on all three criteria
- DualRL outperforms all baseline methods
- All models are more successful on sentiment than on formality

	YELP					GYAFC				
	Sty	Con	Flu	**Avg**	**Suc**	Sty	Con	Flu	**Avg**	**Suc**
MultiDec [Fu *et al.*, 2018]	2.14	3.02	3.27	2.81	5%	2.21	1.95	2.54	2.23	4%
CrossAlign [Shen *et al.*, 2017]	2.88	2.79	3.40	3.02	14%	2.96	1.33	3.27	2.52	3%
Unpaired [Xu *et al.*, 2018]	2.93	3.38	3.44	3.25	17%	2.69	1.19	2.38	2.09	2%
Template [Li *et al.*, 2018]	3.12	3.71	3.42	3.42	23%	2.74	3.60	3.43	3.26	9%
DelRetri [Li *et al.*, 2018]	3.39	3.49	3.71	3.53	28%	2.47	2.57	2.67	2.57	5%
UnsuperMT [Zhang *et al.*, 2018b]	3.82	3.90	3.93	3.95	40%	3.27	3.54	3.76	3.52	21%
DualRL	**4.11**	**4.33**	**4.31**	**4.25**	**54%**	**3.65**	**3.62**	**3.80**	**3.69**	**28%**

Ablation and Case Studies

- Ablate RL and MLE updates to models
- Without MLE: model generates complicated, un-fluent sentences that get high rewards
- Without RL: no direct control over style modifications
- Case study: DualRL best balances transfer with semantic preservation
    - Red text: grammar errors and improperly generated tokens

	From negative to positive (YELP)	From informal to formal (GYAFC)
Source	Moving past the shape, they were dry and truly tasteless.	(That's what i called it) .. but, why?
CrossAlign	Everyone on the fish, they were fresh and filling.	And i know what this helps me.
Template	Moving past the shape, they a wonderful truly.	(That's what it is called it) .. but, why?
Del-Retri	Moving past the shape is awesome, and they will definitely be back!	(That's what i you it you but why, you?
UnsuperMT	Moving moving the shape, they were juicy and truly delicious.	(That's what i said it) but that is why you were doing.)
DualRL	Moving past the shape, they were tasty and truly delicious.	It is what i called it, but why?

Summary and Discussion

- Unsupervised text style transfer, without parallel data
- Balance semantic preservation and style modification with reward functions
- Two style transfer tasks, three evaluation axes, two evaluation modalities
  - In general, DualRL achieves the best overall success, especially w.r.t. balancing semantic preservation and style modification
  - Strong correlation between automatic and human evaluations
- Ablation study: both RL and MLE contribute to DualRL performance
- Case study: Style transfer that requires substantial semantic modification is difficult

Discussion

What are the advantages to a closed loop feedback framework compared to an adversarial training regime?

How modular is the proposed, DualRL framework?

Could the dual reinforcement learning setup be adapted for arbitrary text style transfers?

Additional Paper #1: Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation

ACL 2019 (Florence, Italy)
Ning Dai, Jianze Liang, Xipeng Qiu, Xuanjing Huang

Motivation and Key Insights

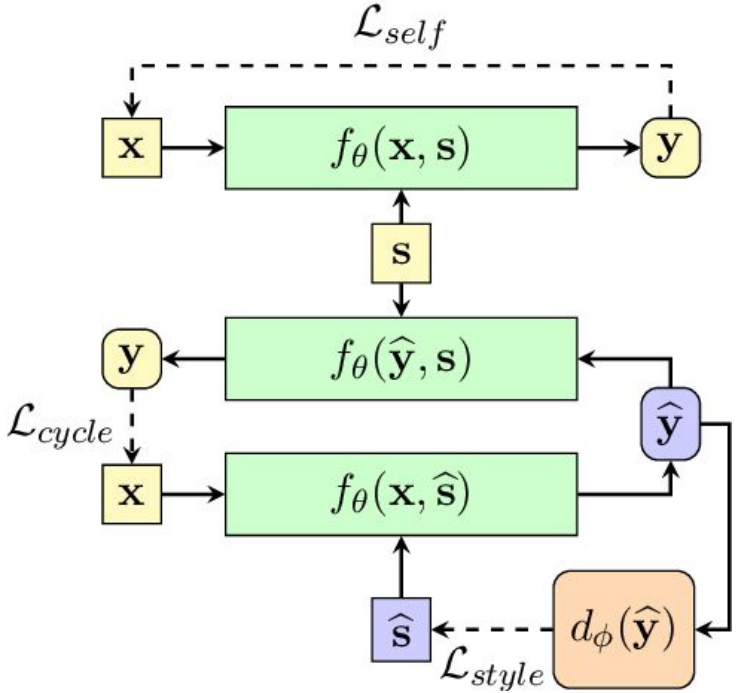
- Learning a latent disentangled representation is hard, parallel data are scarce
  - Independent content and style vectors are impossible in some cases
  - “Style” can be implicit and dependent on the content
- RNN architectures are not best for capturing long-term dependencies
  - The autoencoders and RL models in previous works were LSTMs
  - Well-known limitation of recurrent sequence-to-sequence models
- Hence, semantic preservation is hard to achieve
- Can we use Transformer architecture to improve semantic preservation?
  - Standard basis for encoder-decoder architectures for many NLP tasks
  - Stacked self-attention

Architecture Overview

- Style Transformer network, $f\theta$: Transformer encoder-decoder
  - Encoder maps input sequence to an intermediate sequence of continuous representations
  - Decoder estimates conditional probability for output sequence
  - Add a style control embedding as input to the Transformer encoder
- Train a discriminator network, $d\phi$, in adversarial setting for supervision
  - Transformer encoder
  - Given a sequence, predict if the sequence has a specified style (conditional discriminator), OR classify its style (multi-class discriminator)
- Supervision insight: Given a sequence and *its own* style embedding, an optimal solution is reconstructing the input
  - Similar idea to backward model in DualRL
  - Semantic preservation and style control are implicitly supervised

Training Procedure (1/2)

- Style Transformer network has 3 losses:
  - Self reconstruction: input + original style should reconstruct input
  - Cycle reconstruction: generated + original style should reconstruct input
  - Style control: input + target style should be classified by discriminator as target style
- Discriminator network has its own loss function (cross-entropy)



https://www.aclweb.org/anthology/P19-1601.pdf

Training Procedure (2/2)

- Adversarial training, similar to GANs
- For each training iteration:
  - First, train the discriminator network on each dataset (accumulate loss + gradient descent)
  - Then, train the Style Transformer network on each dataset, using the updated discriminator network
  - Repeat until Style Transformer network converges
- Discriminator trained on mix of real and generated sentences
- Ablation study: disable each loss function
  - No self-reconstruction loss: model does not output meaningful sequences
  - No cycle reconstruction loss: significant drop in BLEU score (content retention)
  - No style control loss: model learns to copy input sequence

Experimental Setup

- Encoder, decoder, discriminator: 4-layer Transformers, 4 attention heads in each layer
- Sentiment transfer on two datasets
  - Yelp review dataset (Li et al., 2018; same as DualRL)
  - IMDb movie review dataset (Maas et al., 2011: “Learning word vectors for sentiment analysis”
- Three axes for automatic evaluation
  - Style transfer accuracy: fastText sentiment classifiers
  - Content preservation: BLEU score (on input sentence and human-written reference)
  - Fluency: perplexity measured with a 5-gram language model
- Also conducted human evaluation, on same axes, with batch comparison against baseline methods (plus “no preference”)

Automatic Evaluation

- Competitive overall performance, substantially better BLEU scores
- Conditional discriminator results in higher style transfer accuracy, but multi-class is better at semantic preservation
- No mean scores computed (as with DualRL evaluation)
- Fluency as measured by perplexity is superior to RNN-based autoencoder methods

Model	Yelp				IMDb		
	ACC	*ref*-BLEU	*self*-BLEU	PPL	ACC	*self*-BLEU	PPL
Input Copy	3.8	23	100	41	5.1	100	58
RetrieveOnly (Li et al., 2018)	92.6	0.4	0.7	**7**	N/A	N/A	N/A
TemplateBased (Li et al., 2018)	84.3	13.7	44.1	117	N/A	N/A	N/A
DeleteOnly (Li et al., 2018)	85.7	9.7	28.6	72	N/A	N/A	N/A
DeleteAndRetrieve (Li et al., 2018)	87.7	10.4	29.1	60	58.8	55.4	**57**
ControlledGen (Hu et al., 2017)	88.8	14.3	45.7	219	94.1	62.1	143
CrossAlignment (Shen et al., 2017)	76.3	4.3	13.2	53	N/A	N/A	N/A
MultiDecoder (Fu et al., 2018)	49.8	9.2	37.9	90	N/A	N/A	N/A
CycleRL(Xu et al., 2018)	88.0	2.8	7.2	107	**97.8**	4.9	177
Ours (Conditional)	**93.7**	17.1	45.3	90	86.6	66.2	107
Ours (Multi-Class)	87.7	**20.3**	**54.9**	73	80.3	**70.5**	105

Human Evaluation

- Baseline selection informed by automatic evaluation
- Randomly sample 100 sentences from TEST set
- Reviewer chooses preference between 3 methods + “no preference,” against the source sentence, on 3 questions
- Over 400 human reviews
- Style Transformer preferred in majority

Model	Yelp			IMDb		
	Style	Content	Fluency	Style	Content	Fluency
CtrlGen	16.8	23.6	17.7	**30.0**	19.5	22.0
DAR	13.6	15.5	21.4	21.0	27.0	25.0
Ours	**48.6**	**36.8**	**41.4**	29.5	**35.0**	**31.5**
No Preference	20.9	24.1	19.5	19.5	18.5	21.5

Reviewer questions:

1. Which sentence has the most opposite sentiment toward the source sentence?
2. Which sentence retains the most content from the source sentence?
3. Which sentence is the most fluid one?

Case Study

- Red: good transfer
- Blue: bad transfer
- Gold: grammar error
- (with respect to human reference)

- Qualitatively, Style Transformer retains more content from the input sentence (compare with DAR)

	negative to positive
Input	the food 's ok , the service is among the worst i have encountered .
DAR	the food 's ok , the service is among great and service among .
CtrlGen	the food 's ok , the service is among the randy i have encountered .
Ours	the food 's delicious , the service is among the best i have encountered .
Human	the food is good , and the service is one of the best i 've ever encountered .
Input	this is the worst walmart neighborhood market out of any of them .
DAR	walmart market is one of my favorite places in any neighborhood out of them .
CtrlGen	fantastic is the randy go neighborhood market out of any of them .
Ours	this is the best walmart neighborhood market out of any of them .
Human	this is the best walmart out of all of them .
Input	always rude in their tone and always have shitty customer service !
DAR	i always enjoy going in always their kristen and always have shitty customer service !
CtrlGen	always good in their tone and always have shitty customer service !
Ours	always nice in their tone and always have provides customer service !
Human	such nice customer service , they listen to anyones concerns and assist them with it .
	positive to negative
Input	everything is fresh and so delicious !
DAR	small impression was ok , but lacking i have piss stuffing night .
CtrlGen	everything is disgrace and so bland !
Ours	everything is overcooked and so cold !
Human	everything was so stale .
Input	these two women are professionals .
DAR	these two scam women are professionals .
CtrlGen	shame two women are unimpressive .
Ours	these two women are amateur .
Human	these two women are not professionals .
Input	fantastic place to see a show as every seat is a great seat !
DAR	there is no reason to see a show as every seat seat !
CtrlGen	unsafe place to embarrassing lazy run as every seat is lazy disappointment seat !
Ours	disgusting place to see a show as every seat is a terrible seat !
Human	terrible place to see a show as every seat is a horrible seat !

Summary and Discussion

- Style transfer using Transformer encoder-decoder architecture
- Adversarial training regime to compensate for non-parallel data
- Sentiment transfer, three evaluation axes, two evaluation modalities
  - Overall, Style Transformer is competitive in sentiment transfer, especially in content retention
- Lack of explicit latent, disentangled representation and avoidance of long-term dependency shortcoming of RNNs improves semantic preservation

Discussion

How would you expect the adversarial training regime for the Style Transformer network to perform with a more ambiguous style transfer (e.g., politeness or political slant)?

The architectures for the Style Transformer and the discriminator are very similar. Would you expect different attention mechanisms (e.g., windowed attention) to change performance on either semantic preservation or style control?

Additional Paper #2: A Recipe for Arbitrary Text Style Transfer with Large Language Models

arXiv pre-print, September 2021
Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, Jason Wei

Motivation and Key Insights

- Style transfer requires parallel text, learning a latent representation or substantial non-parallel data with a distinct style
  - Even Style Transformer requires a large amount of data and training to succeed
- Previous work results in style generators attuned to a specific style
- Other work (not presented) explores “label-free” style transfer, but requires exemplar sentences of the target style
- Can we leverage large LMs for *arbitrary* text style transfer?
  - Brown et al., 2020: “Language models are few-shot learners.”
  - Pre-pending an appropriate prompt to the input sequence widens applications of large LMs
- Augmented zero-shot learning is a prompting recipe that frames arbitrary text style transfer as a rewriting task for large LMs

Method Overview

- Zero-shot prompting
  - Give a single natural language command to the large LM
    - Input: “The day was hot.”
    - Prompt: “More negative.”
    - Output: “The day was sweltering.”
  - Flexible, can give any natural language prompt; but can give illogical responses
- Few-shot prompting
  - Give several exemplar sentences in the target style
  - Achieves higher performance, but requires pre-specified style exemplars
- Augmented zero-shot learning
  - Give exemplars in *related* styles
  - Exemplars constrain LM outputs
  - Balance flexibility with constraint satisfaction

Augmented Zero-shot Learning Example

Augmented Zero-shot Prompt: LLM

"Here is some text: {When the doctor asked Linda to take the medicine, he smiled and gave her a lollipop.}. Here is a rewrite of the text, which is more scary.
{When the doctor told Linda to take the medicine, there had been a malicious gleam in her eye that Linda didn't like at all.}
Here is some text: {they asked loudly, over the sound of the train.}. Here is a rewrite of the text, which is more intense.
{they yelled aggressively, over the clanging of the train}
Here is some text: {When Mohammed left the theatre, it was already dark out}.
Here is a rewrite of the text, which is about the movie itself. {The movie was longer than Mohammed had expected, and despite the excellent ratings he was a bit disappointed when he left the theatre.}
Here is some text: {next to the path}. Here is a rewrite of the text, which is about France.
{next to la Seine}
Here is some text: {The man stood outside the grocery store, ringing the bell.}. Here is a rewrite of the text, which is about clowns.
{The man stood outside the circus, holding a bunch of balloons.}
Here is some text: {the bell ringing}. Here is a rewrite of the text, which is more flowery.
{the peales of the jangling bell}
Here is some text: {against the tree}. Here is a rewrite of the text, which is includes the word 'snow'.
{against the snow-covered bark of the tree}
Here is some text: {That is an ugly dress}. Here is a rewrite of the text, which is more positive."

Experiment Setup

- Two large language models trained:
  - LLM: dense, left-to-right, decoder-only Transformer trained on public web documents corpus
  - LLM-Dialog: extension of LLM finetuned on conversational format subset of the corpus
- Two varieties of transfer tasks:
  - Standard style transfer: sentiment (Yelp polarity dataset) and formality (GYAFC)
  - Non-standard style transfer: 6 descriptions of adjustments made by an interactive editor
- Automatic evaluation metrics on standard style transfers:
  - Style transfer accuracy
  - Semantic preservation via BLEU score
  - Fluency via perplexity
- Human evaluation for standard and non-standard style transfers

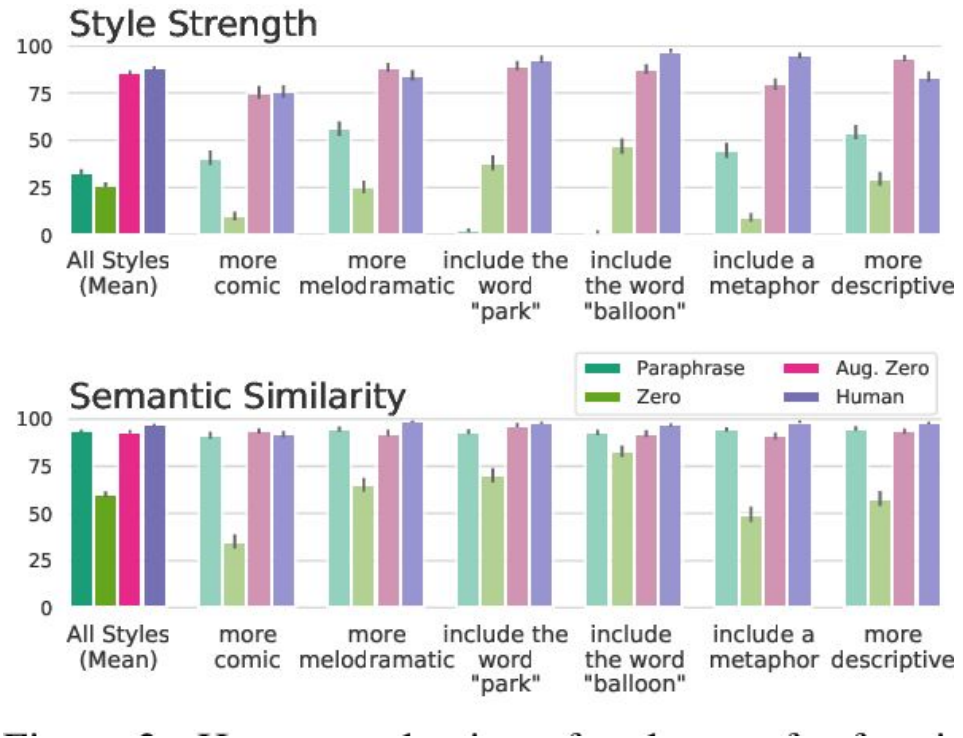
Standard Style Transfer Results

- Compared to baselines, aug. zero achieves comparable accuracy and low perplexity, but low BLEU score
- Tendency to add information to generated sentences
- Evaluated off-the-shelf GPT-3 model, recipe generalizes to other LMs
- *Larger* LM improves performance

	Acc	BLEU	PPL
SUPERVISED METHODS			
Cross-alignment (Shen et al., 2017)	73.4	17.6	812
Backtrans (Prabhumoye et al., 2018)	90.5	5.1	424
Multidecoder (Fu et al., 2018)	50.3	27.7	1,703
Delete-only (Li et al., 2018)	81.4	28.6	606
Delete-retrieve (Li et al., 2018)	86.2	31.1	948
Unpaired RL (Xu et al., 2018)	52.2	37.2	2,750
Dual RL (Luo et al., 2019b)	85.9	55.1	982
Style transformer (Dai et al., 2019)	82.1	55.2	935
INFERENCE-ONLY METHODS			
GPT-3 ada, aug zero-shot	31.5	39.0	283
GPT-3 curie, aug zero-shot	53.0	48.3	207
GPT-3 da vinci, aug zero-shot	74.1	43.8	231
LLM: zero-shot	69.7	28.6	397
five-shot	83.2	19.8	240
aug zero-shot	79.6	16.1	173
LLM-dialog: zero-shot	59.1	17.6	138
five-shot	94.3	13.6	126
aug zero-shot	90.6	10.4	79

Non-standard Style Transfer Results

- Human scoring, 1-100
- LLM-Dialog used for these evaluations
- “Paraphrase” is a recipe with the target style of “paraphrase” as a control condition
- Proper aug. zero is closest to human-written reference overall
- Vanilla zero-shot prompting performed worst



https://arxiv.org/pdf/2109.03910.pdf

Limitations

- Mostly related to known limitations of large LMs
- Privacy, social biases, resource barriers to entry
- Hallucinations: inject superfluous text content
- Unparsable answers: Augmented zero-shot learning can still produce outputs that are illogical, since training data for the LM might not reflect input/output format
    - “That is an ugly dress” → “Sounds like you are a great writer!” (???)
- The quality of the prompt is important, but still being explored
- Overall, prompting is less reliable and robust, but more flexible, than trained text style transfer models

Summary and Conclusions

- Style transfer can be achieved by constructing prompts in an augmented zero-shot learning paradigm for a large LM
- Style can be “arbitrary” provided with sufficient exemplars
- Does not require style-specific corpora, learning latent representation, or specialized training regime
- Two style transfer “flavors” evaluated by human scoring or automatic metrics, as appropriate
    - Augmented zero-shot learning is substantially better than zero-shot prompting, and mostly competitive against existing SotA
    - Worth noting, SotA for formality and sentiment transfers are specialized models
- Potential application as interactive creative writing editor

Discussion

What are some other applications (besides an interactive editor) of an arbitrary text style transfer “engine”?

Could back-translation be incorporated with augmented zero-shot learning?

Resources and Conclusion

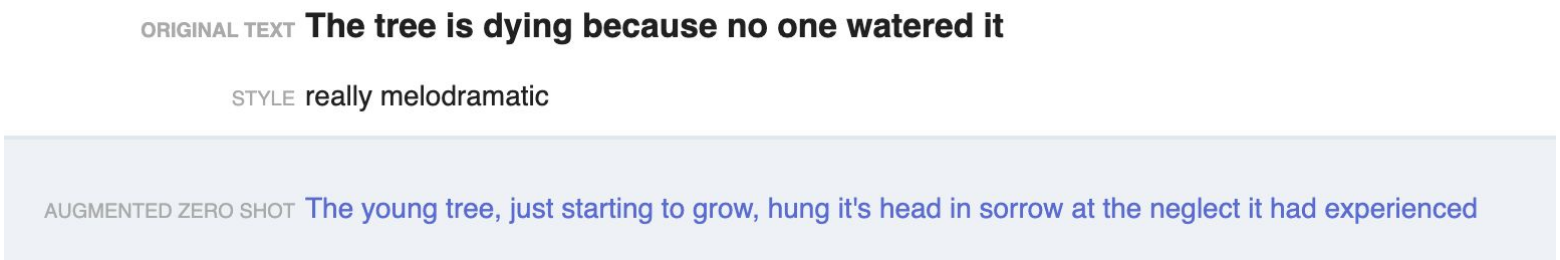
GYAFC Dataset

- Grammarly's Yahoo Answers Formality Corpus [https://github.com/raosudha89/GYAFC-corpus]
- 110K informal/formal sentence pairs (parallel data!)
  - Sentences collected from Yahoo Answers forums (Entertainment & Music; Family & Relationships) [https://webscope.sandbox.yahoo.com/catalog.php?datatype=l]
  - Human-written reference pairing for each collected sentence
- Released in 2018 with the intention of catalyzing formality transfer research
- Also published benchmarks on both human and automatic metrics
  - Rule-based approach
  - Phrase-based approach (PBMT)
  - Neural machine translation (NMT)
  - Formality classifier
  - Statistical model for sentence grammaticality (fluency)
  - CNN-based sentence similarity for meaning preservation

Links to Papers and Implementations

- GYAFC: https://github.com/raosudha89/GYAFC-corpus
- BTS: https://github.com/shrimai/Style-Transfer-Through-Back-Translation
- DualRL: https://github.com/luofuli/DualRL
- Style Transformer: https://github.com/fastnlp/style-transformer

- Examples of augmented zero-shot learning: https://storage.googleapis.com/style-transfer-paper-123/index.html



Slide Reserves

Extra Resources

Visual style transfer web demo:
https://reiinakano.com/arbitrary-image-stylization-tfjs/

Recent survey paper: Jin et al., 2020: “Deep Learning for Text Style Transfer.” [https://arxiv.org/pdf/2011.00416v3.pdf]

Text style transfer web demo: Krishna et al., 2020: “Reformulating unsupervised style transfer as Paraphrase Generation. ” [https://arxiv.org/pdf/2010.05700.pdf] [http://arkham.cs.umass.edu:8553/]

Discussion Questions

Are the automatic metrics commonly assessed really suitable for text style transfer?

Is “style” too broadly defined for an arbitrary style transfer task?

How do the various kinds of machine text style transfer compare to a human approach to the task (introspective design)?