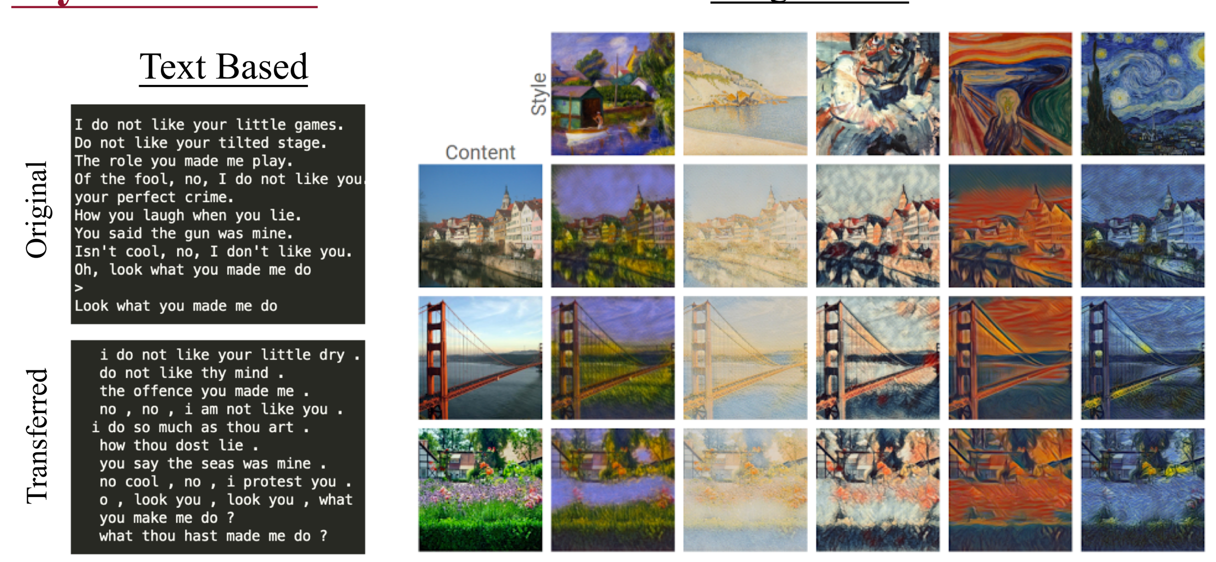
Robust Unsupervised Style Transfer Architecture for Complex Discrete Structures

Stanford University

Cairo Mo, Cherie Xu, Alice Yang

Introduction

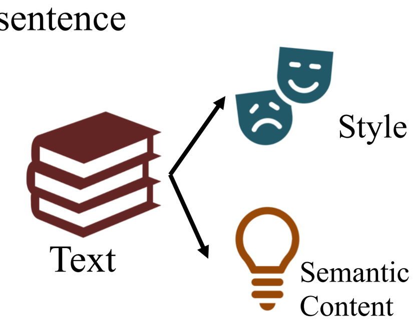
Style Transfer



Task: apply target style to the content of a source sentence

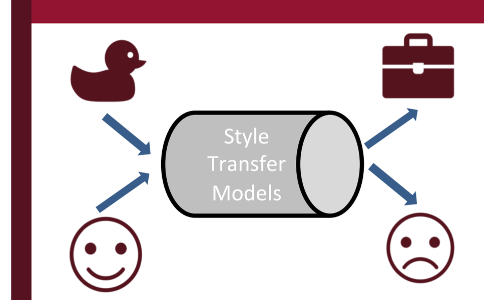
What is style?

- An abstract notion reflected in variation in word choice, sentence and paragraph structure
- Hard to identify and isolate from semantics
- Meta/pragmatic-feature of language



Style transfer requires the disentanglement of representations of attributes (e.g. negative/positive sentiment, plaintext/ ciphertext orthography) from the underlying semantic content. Breakthroughs in style transfer would indicate a certain proficiency of NLP's ability for more complex tasks.

Problem



Past Approaches

Parallelized Methods

- Rule-based transfer
- Delete, Retrieve Generate

Unparallelized Methods

- NMT
- Seq2seq
- Encoder-Decoder

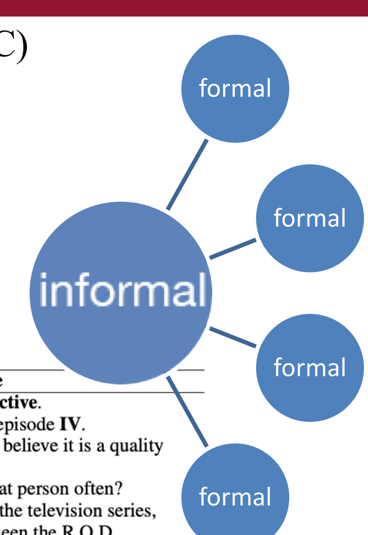
We define two data domains for this project χ_s and χ_t, for source and target data respectively. During training, we observe n samples in χ_s denoted by $X_s = \{x_s^{(1)}, x_s^{(2)}, \ldots, x_s^{(n)}\}$ and m samples in χ_t denoted by $X_t = \{x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(m)}\}$. Every $x_t^{(i)}$ is encoded into a latent representation and such representation is decoded using decoder of the target style to produce $\tilde{x}_t{}^{(i)}$.

Since the data is unparallelized, there is no semantic similarity or correspondence between any given pair of $(x_s^{(i)}, x_t^{(j)})$. We want to train a model to learn from these unparallelized data such as an unseen sample $x \in \chi_s$ can be transformed into $\tilde{x} \in \chi_t$ such as the semantic content of the sentence is persevered while having the target style.

Dataset

Grammarly Yahoo Answer Formality Corpus (GYAFC)

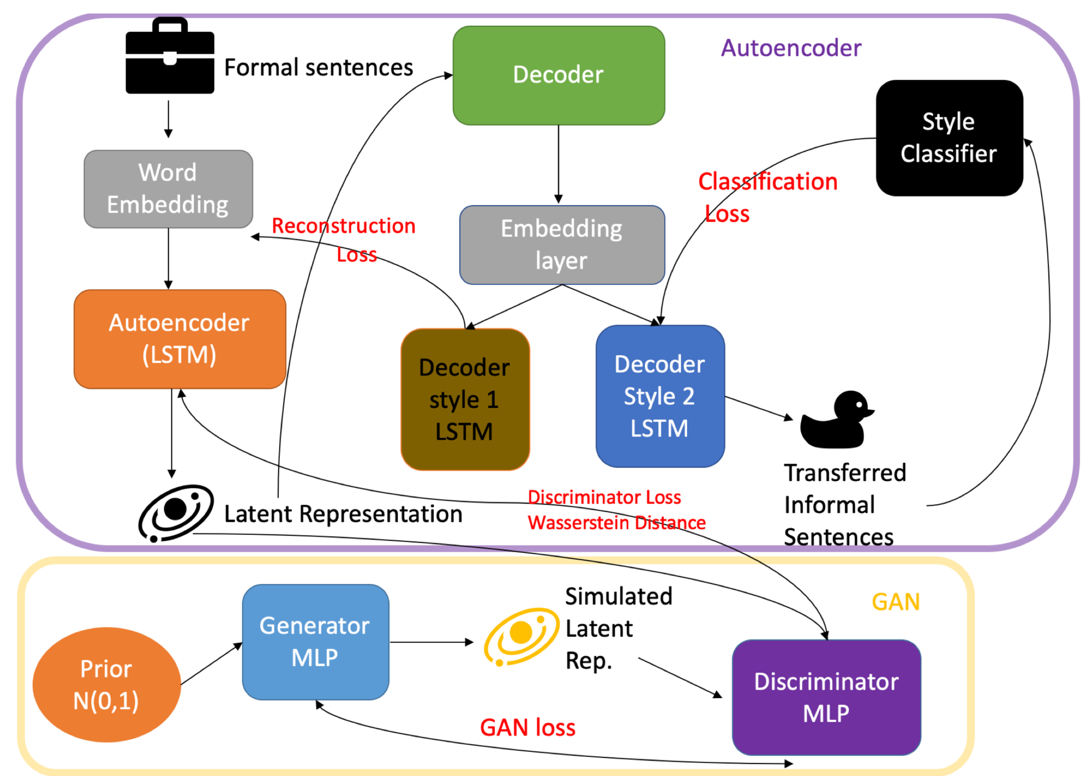
- Largest Dataset for Stylistic Transfer
- Informal/ Formal pairs
- Family-Relations + Entertainment music
- Largest aligned data for automatic eval



Category	Manual	Auto	Original Informal	Formal Rewrite
Paraphrase	47%	–	he iss **wayyyy hottt**	He is **very attractive**.
Capitalization	46%	51%	**yes**, exept for episode **iv**.	**Yes**, but not for episode **IV**.
Punctuation	40%	69%	I've watched it and it is AWESOME!!!!	I viewed it and I believe it is a quality program**.**
Delete fillers	26%	–	**Well...** Do you talk to that someone much?	do you talk to that person often?
Completion	15%	–	**Haven't seen** the tv series, but R.O.D.	**I have not seen** the television series, however I have seen the R.O.D
Spelling	14%	–	that page did not give me **viroses** (i think)	I don't think that page gave me **viruses**.
Contractions	12%	8%	I **didn't** know they had an HBO in the 80's	I **did not** know HBO existed in the 1980s.
Normalization	10%	61%	my exams **r** not over yet	My exams **are** not over yet.
Lowercase	7%	8%	But you will **DEFINALTELY** know when you are in love!	You will **definitely** know when you are in love.

Architectures

ARAE



<u>Autoencoder</u>

Word Embedding with vocab size of 30004, 50 as the max sentence size, vector dimension of (128: Baseline, 256:Optimal)

<u>Encoder and Decoders</u>

LSTM with 128 hidden layers.

Embedding decoder 1 and decoder 2 both have embedding size of 30004, with 128 hidden layers

<u>Decoder 1</u> and <u>Decoder 2</u>

LSTMs with 128 hidden layers

Corresponding to each style

<u>GAN generator</u>

uses Gaussian Prior ~N(0,1) of 32 dimension and MLP

Layer sizes: (32-128-128-128)

ReLU activation between each linear layer

<u>GAN discriminator</u>

MLP layer that takes 128 input

Layer sizes (128-128-128), ReLu nonlinearity between layers

Cross Entropy Loss

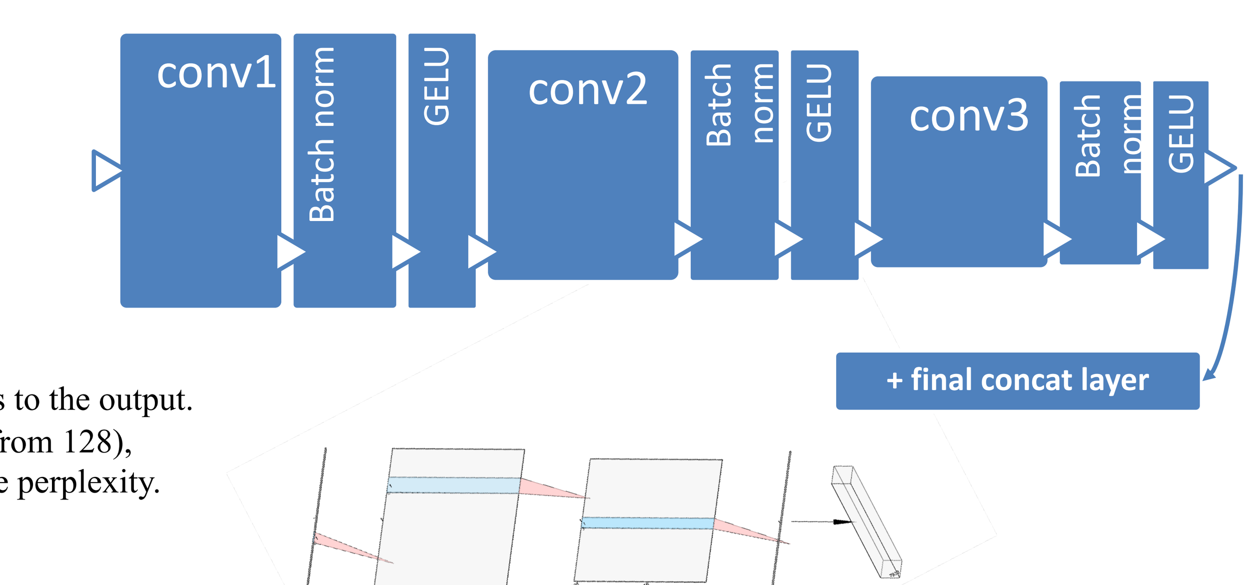
<u>Style Classifier</u>

3 linear fully connected layer

Layer sizes (128-128-128), ReLu nonlinearity between layers

Cross Entropy Softmax Loss

CNN-GAN



Full Loss Objective

$$\min_{\phi,\psi,\theta} \mathcal{L}_{\text{rec}}(\phi,\psi) + \lambda^{(1)} W(\mathbb{P}_Q, \mathbb{P}_{\mathbf{z}}) - \lambda^{(2)} \mathcal{L}_{\text{class}}(\phi, u)$$

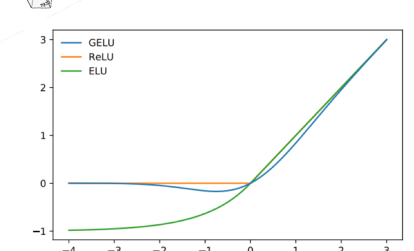
Large Latent Space

larger latent space: allows the model to perform more complex manipulations to the output.

Dimension: 128 -> 256 By increasing the latent space dimension to 256 (up from 128),

Observed: continued learning until the 50th epoch and continued reducing the perplexity.

GloVe Weight Initialization



Motivation: incorporating pre-trained data can improve the learning of semantics

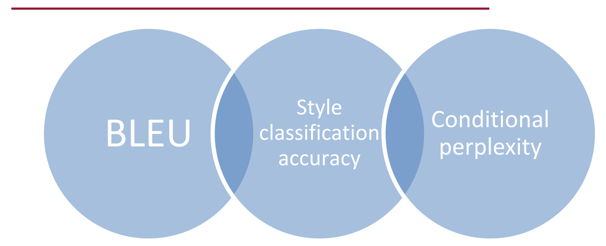
- better disentangle style and semantics at the decoding phase
- previous experiments' generated output showed semantic meaning change (undesirable)

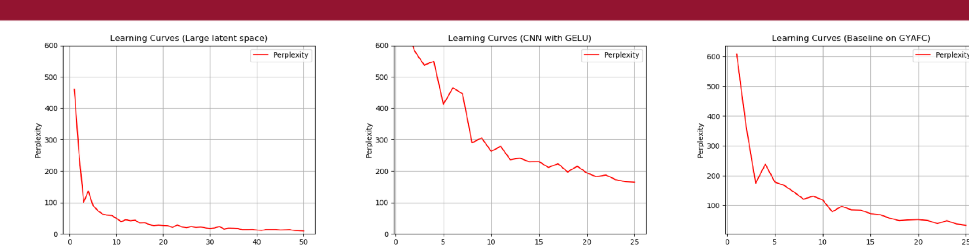
GloVe embeddings used in

1. fixed word embeddings for autoencoder
2. weight initialization for encoder layer, trainable by the data.

Results

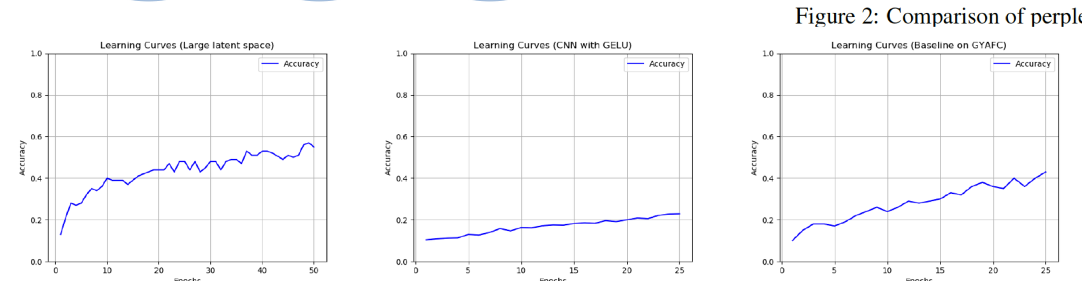
Automatic Evaluation





(a) Perplexity of large latent space (b) Perplexity of CNN + GELU model (c) Perplexity of baseline model

Figure 2: Comparison of perplexity of three models during ARAE training

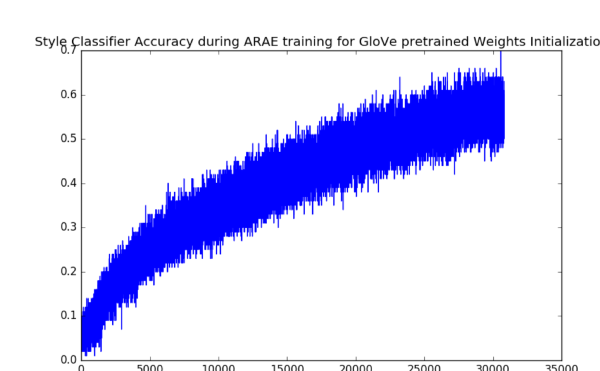


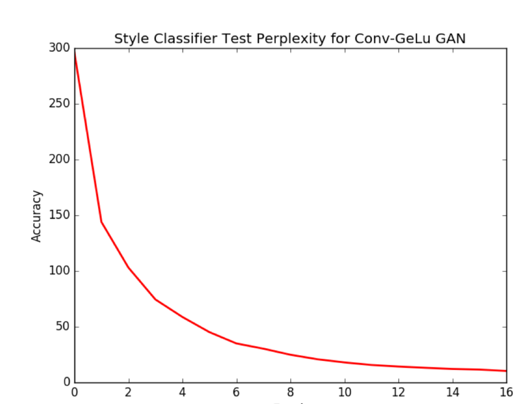
(a) Accuracy of large latent space (b) Accuracy of CNN + GELU model (c) Accuracy of baseline model

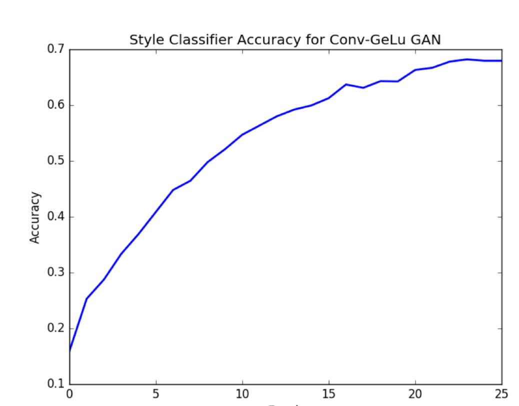
Figure 3: Comparison of accuracy of three models during ARAE training

Automatic Evaluation*			
	Baseline on GYAFC	Large Latent Space	CNN and GELU (10k)
Corpus BLEU Decoder 1 Source	84.9	72.4	68.5
Corpus BLEU Decoder 1 Target	1.02	2.32	1.66
Corpus BLEU Decoder 2 Source	67.3	64.9	60.8
Corpus BLEU Decoder 2 Target	3.04	2.08	2.05
Accuracy	0.703	0.773	0.229
Perplexity	18.53	9.65	164.3

Table 1: **Style transfer automatic evaluation**: Evaluation of the experimentalal models on the 10k subsampled dataset



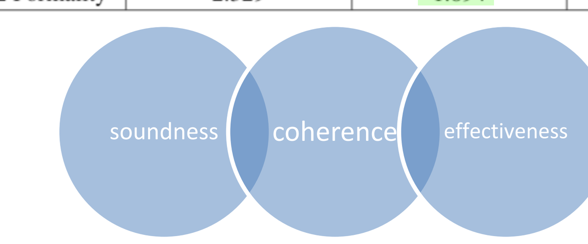




Results

Human Evaluation

	Baseline on GYAFC	Large Latent Space	CNN and GELU (10k)
Soundness	1.771	2.036	2.010
Decoder 1 Coherence	3.365	2.900	4.670
Decoder 2 Coherence	3.482	2.612	2.670
Decoder 1 Formality	2.841	2.771	3.450
Decoder 2 Formality	2.329	1.894	2.560



Large Latent Space Model-generated	
Noise to Formal	Noise to Informal
you can do not have it only anyone.	you can do it only for it as long time.
you should tell him the truth of a boyfriend.	so get him back on a man!
even a friend may not go out with him, and it's not working out myself.	sounds like trying to never give him some one and talk about it

CNN + GELU + Concat generated

Informal

- does this mean that he is attracted to me?
- i have never wanted to it to be a married woman. it it has to want to have been it to be quite a woman.

Formal

- have fun in order to make that type girl you are making to your husband.
- be yourself, when you want her back. tell her. and tell her. tell her. you tell her.

Discussion

Larger latent space improved slightly, 256 is the optimal hidden size on a unmodified model. >300 overfits the model.

GloVe pretrained weight initialization overfits the data and does not handle <unk> well

Naïve Convnet without GELU diminishing gradient problem

Conv + GELU improved style transfer accuracy and reduced perplexity

- For longer sentences, generated output became less grammatical

Conv + GELU + Concat can create grammatical sentences of longer length

- Achieved 0.7 accuracy

Main challenges:

- symbolic info lost during the encoding, decoding, and the generation of the sentences.
- Similar challenges found in previous architectures using LSTM and CNN for text generation
- GAN for text problem -- gradients from the discriminator cannot effectively back-propagate through discrete variables.

Future Work

- Exploring the combination of rule-based and statistical approaches
- Handle <unk> with a spell checker or character-based embedding
- Our model in its current state does not learn explicitly a latent representation of language style
- Explicitly extract a style embedding of the sentence (our model instead uses a style decoder)
- The interpretability of the latent representation is poor.
- Multitask learning for disentangling semantic content.

References

[1] Tetreault Rao. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. 2018. https://arxiv.org/pdf/1803.06535.pdf.

[2] Zhao, Kim, Zhang, Rush, LeCun. Adversarially regularized autoencoders. 2018. https://arxiv.org/pdf/1706.04223.pdf.

[3] Singh, Palod. Sentiment transfer using seq2seq adversarial autoencoders. 2018.https://arxiv.org/pdf/1804.04003.pdf.

[4] Alexey Tikhonov and Ivan P. Yamshchikov. What is wrong with style transfer for texts? https://arxiv.org/pdf/1808.04365.pdf, 2018.

[5] Jhamtani, Gangal, Hovy, Nyberg. Shakespearizing modern language using copy-enriched sequence-to-sequence models. 2017. https://arxiv.org/pdf/1707.01161.pdf.

[12] Soujanya Poria Erik Cambria Tom Young, Devamanyu Hazarika. Recent trends in deep learning based natural language processing. https://arxiv.org/pdf/1708.02709.pdf?, 2018.

[6] Zhang, Ren, Liu, Wang, Chen, Li, Zhou, Chen. Style transfer as unsupervised machine translation. 2018. https://arxiv.org/pdf/1808.07894.pdf.

[7] David Lopez-Paz Arthur Szlam Piotr Bojanowski, Armand Joulin. Optimizing the latent space of generative networks. https://arxiv.org/pdf/1707.05776.pdf, 2017.

[8] Yoon Kim. Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014. https://www.aclweb.org/anthology/D14-1181.

[9] Kevin Gimpel Dan Hendrycks. Gaussian error linear unit. https://arxiv.org/pdf/1606.08415.pdf, 2018.9 A PREPRINT - MARCH 19, 2019

[10] Nanyun Peng-Dongyan Zhao Rui Yan Zhenxin Fu, Xiaoye Tan. Style transfer in text: Exploration and evaluation.https://arxiv.org/pdf/1711.06861.pdf, 2017.

[11] Christopher D. Manning Jeffrey Pennington, Richard Socher. Glove: Global vectors for word representation. https://nlp.stanford.edu/pubs/glove.pdf, 2014.