WiDS: NLP and Text Style Transfer

Final Report

Mugdha Bilotia

1 Introduction

Text style transfer involves rewriting sentences to embody a specified style while retaining original content. In the following report, I will be presenting my understanding of two research papers in the field of NLP and Text Style Transfer. The topics of the research papers are:

- Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation
- Controllable Unsupervised Text Attribute Transfer via Editing Entangled Latent Representation

2 Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation

2.1 Abstract

This paper introduces the Style Transformer, a novel model for text style transfer, addressing the challenge of generating natural language sentences with desired styles while preserving content. The proposed approach employs a Transformer-based architecture and a unique training algorithm integrating self-reconstruction, cycle reconstruction, and discriminator-based style control. Extensive experiments on Yelp and IMDb datasets demonstrate competitive or superior performance compared to state-of-the-art models, particularly excelling in content preservation.

2.2 Problem Formalization

The paper defines the style transfer problem, emphasizing the need for rewriting a given sentence in a new style while preserving content. It introduces datasets with distinct styles and articulates the goal as transforming an input sentence x to a new sentence xb with a desired style bs.

2.3 Model Overview:

It details the Style Transformer architecture, utilizing a Transformer encoder-decoder framework. A key challenge is addressed by incorporating an additional style embedding into the encoder, allowing the model to compute output probabilities based on both input sentence x and style s.

- The encoder is denoted as $Enc(x, s; \theta_E)$, where x is natural language sentence, s is style control variable.
- The decoder is denoted as $Dec(z; \theta_D)$, where z is a sequence of continuous representations.

2.3.1 Probability Computation:

Probability computation in the Transformer framework is expressed as p(y—x, s) = p(y—z, y, ..., y), where y is the output sentence.

2.4 Training Algorithm:

It presents a comprehensive training algorithm involving discriminator-based approaches to create supervision from non-parallel corpora. The algorithm includes discriminator learning and Style Transformer learning steps, integrating self-reconstruction, cycle reconstruction, and style-controlling losses.

2.4.1 1. Discriminator Learning:

Discriminator is trained using a combination of real and generated sentences. Positive samples are labeled as 1, and negative samples as 0.

2.4.2 2. Style Transformer Learning:

- Self-reconstruction loss (Lself): Minimizing negative log-likelihood for the same style.
- Cycle reconstruction loss (Lcycle): Minimizing negative log-likelihood for the original style.
- Style controlling loss (Lstyle): Minimizing negative log-likelihood for the desired style.

2.4.3 3. Overall Learning:

The training algorithm alternates between discriminator learning and Style Transformer learning.

2.5 Ablation Study:

2.5.1 1. Impact of Loss Functions:

Self-reconstruction loss guides the model to generate readable sentences. Cycle reconstruction loss encourages information preservation. Discriminator loss provides style supervision.

2.5.2 2. Role of Different Samples in Discriminator Training:

A mixture of real and generated samples is used for effective discriminator training.

2.6 Conclusion and Future Work:

In the end, the paper concludes the study by summarizing the contributions of the Style Transformer, emphasizing its competitive performance, particularly in content preservation. Highlights future directions, including adaptation to multiple-attribute settings.

3 Controllable Unsupervised Text Attribute Transfer via Editing Entangled Latent Representation

3.1 Introduction

The paper introduces a novel framework for text attribute transfer, emphasizing the need to edit entangled latent representations. The objective is to enable controllable unsupervised attribute transfer in text generation tasks.

3.2 Problem Formalization

3.2.1 Dataset

The dataset X comprises n sentences, each paired with an attribute vector y. This includes various attributes like sentiment, appearance, aroma, taste, etc.

3.2.2 Task Definition

Text attribute transfer involves generating a target text x' while preserving the original content and conforming to the target attribute y_0.

3.3 Model Overview

The model architecture consists of three sub-models:

- **Encoder ($\mathbf{E}\theta_e$):** Encodes the input text x into a latent representation z.
- **Decoder ($\mathbf{D}\theta_d$):** Decodes the modified latent representation z into the target text x'.
- **Attribute Classifier ($\mathbf{C}\theta_c$):** Classifies the attribute of the latent representation z.

3.4 Fast Gradient Iterative Modification Algorithm (FGIM)

The FGIM algorithm is proposed to find an optimal latent representation z_0 that conforms to the target attribute y_0. The algorithm iteratively modifies the latent representation based on the gradient of back-propagation, using a dynamic weight initialization approach.

$$z^{=}z - w_i \nabla_z L_c(C_{\theta_c}(z), y_0)$$

3.5 Transformer-based Autoencoder

3.5.1 Key Points

The model employs a Transformer-based autoencoder with low reconstruction bias.

3.5.2 Model Architecture

- Original Transformer's encoder ($\mathrm{E}_{\mathrm{transformer}}$).
- Additional positional embeddings H.
- GRU layer with self-attention.
- Sigmoid activation function applied to GRU hidden representations.

3.6 Experiments

3.6.1 Implementation Details

Detailed specifications for the Transformer-based autoencoder, classifier, and FGIM. Includes hyperparameters and optimizer details.

3.6.2 Datasets

Usage of sentiment and style transfer datasets: Yelp, Amazon, and Captions.

3.6.3 Evaluation Metrics

Performance evaluation using various metrics such as attribute transfer accuracy, BLEU score, perplexity, and human evaluation.

3.7 Results

3.7.1 Comparison with Baselines

Outperformance of the proposed model compared to existing models on sentiment and style transfer tasks.

3.7.2 Human Evaluation

Higher scores for attribute accuracy, content retainment, and fluency in human evaluation.

3.8 Multi-Aspect Sentiment Transfer

Dataset: Introduction of the Beer Advocate dataset for aspect-based sentiment transfer.
Novelty: First work addressing aspect-based sentiment transfer.
Results: High sentiment accuracy demonstrated in sentiment transfer over multiple aspects.

3.9 Transfer Degree Control

3.9.1 Modification Weight Analysis:

Impact of modification weight w on the degree of attribute transfer. Graphical representation of the relationship between w and attribute accuracy.

3.10 Latent Representation Modification Study

3.10.1 Visualization:

Use of T-SNE to visualize latent representations during the modification process. Graphical representation of the evolution of latent representations.

3.11 Conclusion and Future Work

The model is capable of transferring sentiment and style attributes while preserving content similarity. The proposed algorithm allows for control over the degree of attribute transfer, making it flexible for various applications. Experimental results on sentiment transfer tasks and a new aspect-based sentiment transfer task demonstrate the effectiveness of the proposed model. Ablation studies further confirm the contributions of the autoencoder, attribute classifier, and modification algorithm to the model's overall performance.