Neural Style Transfer for Non-Parallel Text

Vinitra Swamy
Computer Science Division
University of California, Berkeley
vinitra@berkeley.edu

Vasilis Oikonomou
Computer Science Division
University of California, Berkeley
v.oikonomou@berkeley.edu

Abstract

In this project, we consider and improve on existing techniques for Neural Style Transfer from Non-Parallel text. Specifically, we introduce a new setup for the problem by drawing inspiration from recent developments in the field of authorship attribution. Moreover, we attempt to reconsider the notion of style and apply non-parallel style transfer to a newly introduced dataset of Twitter Influencers. Given this new problem formulation, we also present a new style transfer content preservation metric.

1 Introduction

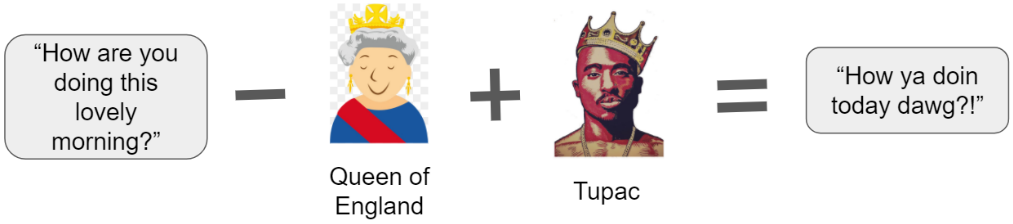
1.1 Motivation

Recent developments in deep learning have pushed the boundaries of natural language processing (NLP) in areas like machine translation, part-of-speech tagging, and sequence-to-sequence learning. In this paper, we attempt to take advantage of deep learning to tackle another problem in NLP: transferring textual style between authors. More specifically, we define our research question as follows.

Given a sentence from Person A, can we disentangle the content and its style and rewrite it in the style of Person B?

For an illustrative example, we look at Fig. 1 for a case of converting a question from the Queen of England ("How are you doing this lovely morning?") into the style of Tupac ("How ya doing today dawg?"). The content remains the same between the two sentences as they both ask their conversational partner how they are feeling currently. However, it's clear that the word "dawg" is not likely to be found in the Queen's vernacular, and that Tupac would likely prefer a less formal translation of "lovely morning".

Figure 1: Style Transfer from the Queen of England into Tupac

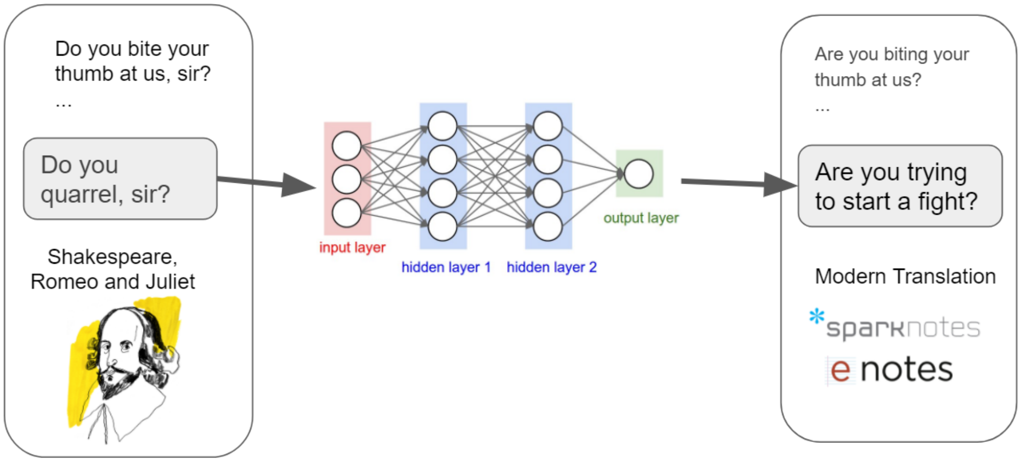

"How are you doing this lovely morning?"
−
Queen of England
+
Tupac
=
"How ya doin today dawg?!"

Previous work in this space divides the problem of neural style transfer for text into two different types of problems based on text alignment.

1.2 Style Transfer for Aligned Text

Most text style transfer literature focuses on the problem of translating between aligned pairs of input and target style sentences, usually aligning older and modern translations of text. Consider the example presented in *Paraphrasing for Style* where Wei Xu et. al. translate phrases of Shakespeare into counterparts in modern English. Xu et. al's style transfer problem is framed by comparing Shakespeare's original manuscripts to scraped data from Sparknotes / eNotes with direct sentence-by-sentence translations between the two styles. This allows the authors to frame the problem in the context of Neural Machine Translation and **sequence to sequence learning** literature, as demonstrated in Fig. 2.

Figure 2: Aligned text style transfer from Shakespeare into Modern English (Sparknotes)


Do you bite your thumb at us, sir?
...
Do you quarrel, sir?
Shakespeare, Romeo and Juliet
input layer
hidden layer 1
hidden layer 2
output layer
Are you biting your thumb at us?
...
Are you trying to start a fight?
Modern Translation
sparknotes
e notes

With aligned text, we can use traditional NLP evaluation metrics, like BLEU score (Fig. 3). BLEU score is a metric introduced by Papineni et. al at IBM Research in 2002, which is the industry standard as a metric for automatic machine translation. In the equation below, W_n represents positive weights, and p_n represents n-gram precision for n grams of length N.

Figure 3: BLEU score calculation

$$\text{BLEU}= \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

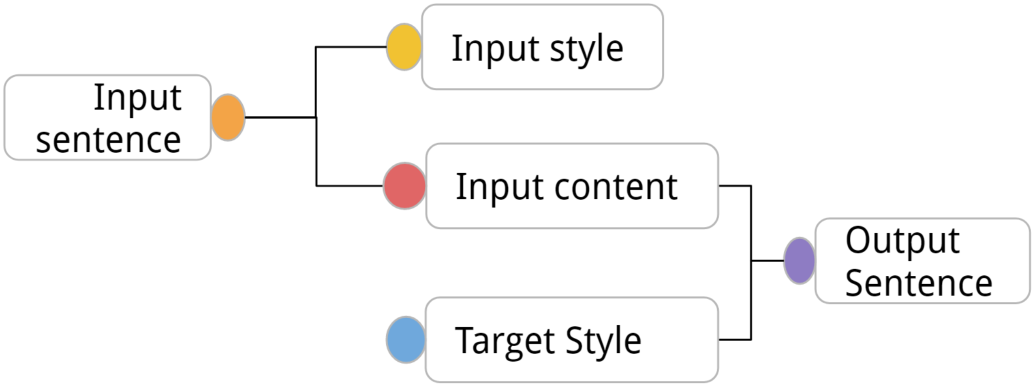
The problem with this approach is that it is very hard to obtain aligned data, and most real world style transfer problems do not have adequate data sets. Formulating the problem with aligned text narrows the scope and therefore makes it more solvable.

1.3 Style Transfer for Unaligned Text

The other (arguably more challenging) approach to style transfer uses unaligned text to develop translation models. Consider translating between Shakespeare and J. K. Rowling. Each author has a large corpus of literature (Shakespeare's plays and Rowling's Harry Potter series), but they rarely refer to the same content. We do not have a direct aligned sentence to sentence translation between sentences in *Hamlet* and *Harry Potter and the Sorcerer's Stone*.

There is not much literature in this area, but those that do exist are concerned with separately distilling the content of an input sentence from its style (Fig. 4). Both style and content are treated as latent variables that need to be learned. This approach is broadly applicable as it solves the general problem of transferring between any two styles given that examples for both appear in our dataset. The methods employed by this problem leverage mostly on literature in **Generative Modeling**.

Figure 4: Conceptual overview for unaligned text style transfer



1.4 Paper Roadmap

In this paper, we've chosen to focus on the problem of Style Transfer for non-parallel (or unaligned) text. We draw heavily on Style Transfer from Non-Parallel Text by Cross-Alignment by Shen et. al from MIT CSAIL, who propose an architecture for Neural Text Style Transfer using a definition of style as sentiment in Yelp reviews. Using their work as inspiration, we provide the following contributions:

1. We consider a different formulation of the problem of Style Transfer for Text by leveraging methods from authorship attribution.
2. We apply ideas from recent research on a new dataset of Twitter influencers for distinct authorship style.
3. We propose new evaluation strategies that could be employed in this new formulation of the problem.

Our paper starts with a discussion of related work, then details our methodology, data, and experiments. We discuss our evaluation strategies and then the results of the work that we have implemented. We end with a discussion of what we have accomplished and our future work.

2 Related Work

2.1 Artistic Style Transfer

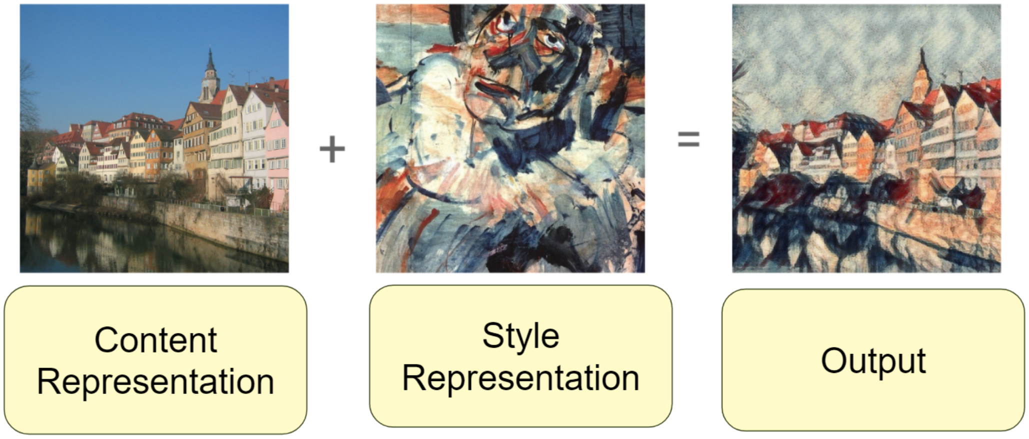
The concept of style transfer was initially introduced by Gatys et. al (2015), who show that it is possible to disentangle the style of an image from its content. Their methodology in *A Neural Algorithm of Artistic Style* involves looking at the correlations between different filter responses over the spatial extent of the feature maps.

We see their methodology of disambiguating style and content, and then applying a certain style to a content representation in Fig. 5. Our inspiration from this work was mainly in the idea of style transfer itself, and understanding how to develop an analogous content vector representation for a text input.

2.2 MIT CSAIL Paper

The paper that we found to be most closely aligned to our research goal of style transfer for non-parallel text is *Style Transfer from Non-Parallel Text by Cross-Alignment* by Shen et. al. The MIT CSAIL team formulates the technical challenge of neural style transfer as separating the content from text characteristics using a variational autoencoder, a discrimination network, and a decoding RNN. The first step in the process is identifying a content representation for a given sentence, which Shen et. al. attempt by minimizing the reconstruction error for a sentence (condensing the representation

Figure 5: Neural Algorithm of Artistic Style (Gatys et. al 2015)



and then expanding it back to its original form with the highest possible accuracy). The problem with this approach is that the autoencoder might learn style features, and we want to guarantee it only learns content.

The solution proposed by the CSAIL team is a discriminator neural network that takes the latent variable z from the autoencoder and attempts to predict the initial sentence style it was extracted from. If our discriminator network is good at identifying the input style, then we know that z encodes some sort of stylistic representation, which was not our intention. Therefore, we jointly optimize over the variational autoencoder and the discriminator network such that the the latent variable z is effective in minimizing reconstruction error as well as maximizing loss for the discriminator network, and the discriminator network has the highest accuracy it can obtain. How do we jump from the latent z content variable to stylized output? CSAIL uses a third neural network (decoder RNN) that outputs content in the target style given z and the target style as input parameters.

There is quite a bit of related work that is related to the aforementioned approach. One of their main contributions is the fact that imposing a Gaussian constraint on the distribution of the latent code is not enough and that we need to enable our architecture to impose, implicitly, more complex distributions. For that reason, we looked at the related literature on Adversarial Autoencoders.

2.3 Authorship Attribution Literature

An important part of our efforts was to develop a reliable model-based evaluation metric that would attribute the style of the generated sentence to the correct author. In working towards this model, we explored many different approaches. Typically authorship attribution is done on large corpuses; there has been extensive research that has looked at ways of identifying the authors of books or articles.

However, since we were operating on a Tweet level, those methods proved to be inadequate. However, recent research on Character-Level Convolutional Neural Networks has delivered promising results that we found to be particularly useful.

In particular, Ruder et. al, use a character-level CNN for large scale authorship attribution. In particular, they emphasize that their approach does a really good job at handling multiple classes and, combined with a very short inference time improves on current state-of-the-art by quite a bit. What makes this approach particularly relevant to our problem is that character level CNN capture information about morphemes that other techniques that operate on a word level cannot identify. This can help us extract more signal from our sequences. An interesting survey of modern techniques for authorship attribution that we used extensively can be found as part of Efstathios Stamatatos's work at the University of the Aegean who used relative frequencies of function words to compare different authors.

3 Methodology

3.1 Modifying Shen et. al

While the process and the methodology is sound, we are uncomfortable with the data and evaluation strategy of the Shen et. al. paper. The input data for the CSAIL formulation is a Yelp dataset that has reviews with classified sentiment – this approach makes a leap in the problem space by equating sentiment with style. Given this assumption, the evaluation strategy used in the paper is a pretrained sentiment classifier that takes the newly generated reviews and classifies the sentiment for that sentence in comparison to the "transfer style" passed into the decoder RNN.

3.2 Challenges

The discussion of what is style is a pervasive challenge throughout style transfer work. In the Gatys et. al work, style is decipherable from content, because they can visualize their content representations after half of the style transfer process. Shen et. al defined style as sentiment, which helps them formulate the problem cohesively. However, we define style as something distinct to author, and therefore we would like to approach the style transfer problem with a dataset that is distinctive on the author domain.

4 Data

4.1 Dataset Considerations

In looking for a dataset with distinct authors, we considered a number of options in both the supervised and unsupervised realms. We entertained the thought of introducing a new technique for Xu et. al's dataset of aligned Shakespeare and modern english translations. We also found an IMDb2 dataset of 62,000 movie reviews from 62 unique authors written by Serrousi et. al that was available by request.

For unsupervised methods, we were considering scraping Project Gutenburg for a new corpus of text from 10 unique authors. We experimented with developing scripts and managed to scrape the data. Most of the authorship attribution works we found had used some variation of this data. However, we finally decided on a different unsupervised learning dataset because of our desire to try out recent advances in authorship attribution for smaller quantities of text.

5 Twitter Data Description

For this project, we decided on using the Twitter influencer dataset introduced by Ruder et. al in 2016. The dataset is comprised of 4,391 celebrity and power-user influencers in 68 domains ranging from politics and tech to arts and writing and collected just over 1M tweets for these users in October and November 2015. Our main concern when selecting a dataset was that there needed to be enough signal to differentiate between different author styles. Twitter proved to be a good place to look for that since most of the celebrities and power users in the dataset use language in their own, distinctive way to cultivate their own brand image. This is probably also the reason why we observed authorship attribution models performing significantly better compared to other domains. It is worth noting that the tweets are not evenly distributed among users. Specifically, we found that around 100,000 of all the tweets in the dataset came from the 10 most prolific authors which largely consisted of various corporate accounts. In the end, we worked with a subsample of that dataset that consisted of the 10 accounts from which we had the most tweets.

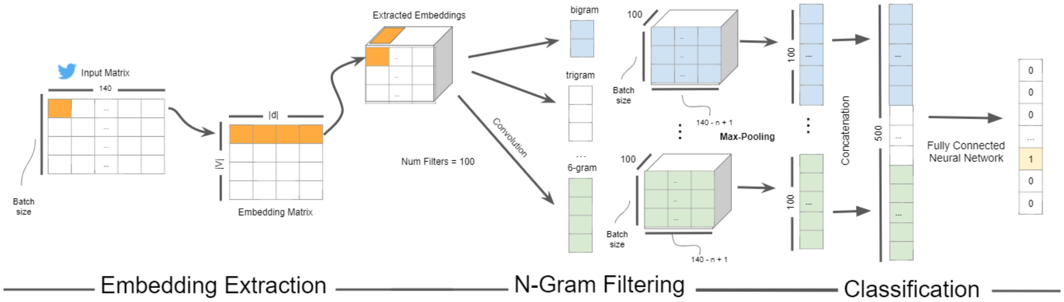
6 Experiment

6.1 Authorship Attribution Architecture

We put quite a bit of thought in designing our architecture as well as the different trade offs that were involved in the process. First off, a critical component of any model-based evaluation metric is reliability. For that reason, we wanted an architecture that performed well in the subset of the original dataset that we chose to work with. Moreover, however, we wanted to make sure that our model would use as little content-related information. One can imagine for example that content should be a discriminative enough feature for the purposes of authorship attribution.

Given these constraints, we experimented with various architectures that revolved in the sphere of Character-Level Models. We consciously avoided using a word channel as has been tried in the past by others (Ruder et. al) and we instead focused on shallow character-level models. A typical architecture that we experimented with consists of indexing an embedding matrix using the unique word id for every word in the tweet. This generates a tensor over which we apply our one dimensional convolutions. For those, we used n-grams ranging from 3 to 7, with 100 filters for each. After that we perform max pooling over each map, concatenate the results and pass it through a fully connected layer with a softmax activation as shown in Fig. 6.

Figure 6: Authorship Attribution Architecture)



Moreover, there were two families of models that we experimented with. In one, we allowed the model to dynamically learn word embeddings for the task while in the other, the embeddings were stationary. that included both dynamically learning character-level embeddings as well as simply having static embeddings.

6.2 Neural Style Transfer Architecture

Our main architecture consists of 4 components:

1. Use one layer Recurrent Neural Network GRU cells to find the latent representation of the content variables, z.
2. A simple one layer feedforward Neural Net that acts as our discriminator. This allowed us to enforce an implicit distribution on z.
3. After that, the latent vector z is fed into a decoder RNN that has the same architecture as the encoder and attempts to reconstruct the original sentence. The whole system is jointly optimized to learn a good latent representation. (Fig. 7)
4. After a good latent representation has been learned, we optimize the decoder network. The training for that is done in a supervised learning fashion by using the latent code z and the one-hot-encoded representation of the style, we train the decoder RNN to reconstruct the original sentence. (Fig. 8)

7 Evaluation Metrics

7.1 Evaluation Literature

There were a lot of ideas that we considered when coming up with the proposed evaluation for this task, especially in our attempt to evaluate style transfer in our generated sentences. Initially, we considered approaching the task using statistical similarity measures that have been proven to be successful in the field of authorship attribution. However, non were actually applicable to our task, since we wanted to focused more on generating sentences rather than longer texts. We also read up on a variety of evaluation metrics from related fields such as text summarization and neural machine translation, however, none proved to be particularly helpful in the absence of target sentences.

Figure 7: Autoencoder content variable learning architecture (Components 1-3))

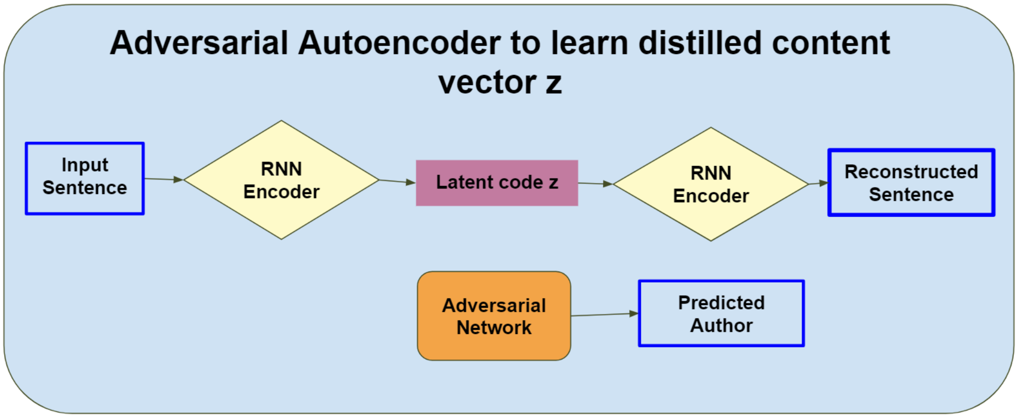
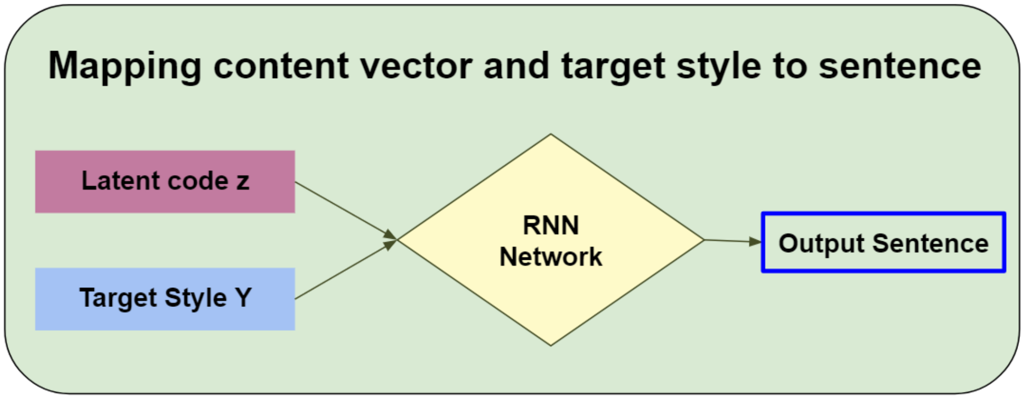


Figure 8: Generating stylized sentence from content variable z (Component 4)



7.2 Style Transformation Metric

Our system seeks to successfully transform the input sentence into the target style. We need a metric to measure the efficacy of style transformation to understand how well our model performs. In the Shen et. al paper, this is done by a neural network trained on classifying the sentiment of their Yelp reviews. However, given that our problem formulation is different, we propose using the accuracy from our character-level CNN for Authorship Attribution.

This entails taking the output of our style transfer architecture, and passing it through our network to classify authorship, and seeing if the author it outputs is our desired author.

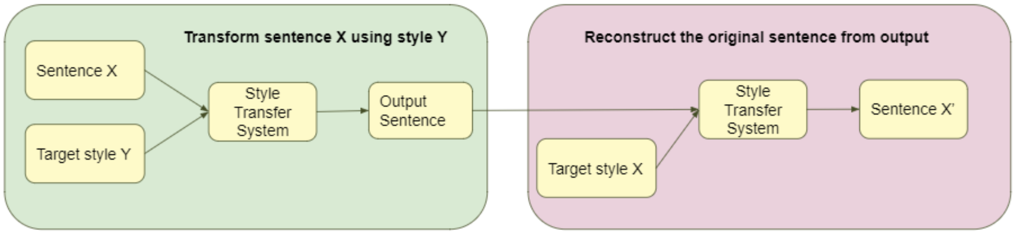
7.3 Content Preservation Metric

Another goal of our system is to preserve the content of the input sentence through the transformation process. For this we propose a metric for neural text style transfer – a roundtrip style transformation as depicted in Fig. 9.

We define our metric using sentence X of style x. We initially pass in our inputs sentence X and target style y through our model, transforming sentence X using style y to create output sentence Y. We then take output sentence Y and original style x, and reconstruct the original sentence by passing it through the architecture once more. The end of this reconstruction process will create sentence X', our attempt at recreating sentence X after converting it into style y and back using style x.

Once we have X and X', we can regress to BLEU score (Fig. 3) to compare the two sentences – this will allow us to see how much content has been preserved.

Figure 9: Evaluation metric for content preservation in neural style transfer



8 Results and Analysis

8.1 Authorship Attribution Results

We performed a comparison between the models with the static and dynamic embeddings. In general, the results between the two are comparable and one model does not seem to outperform the other. Although we observe that the dynamic embedding model seems to develop a slight edge over the static embedding one as we increase the number of authors, this did not particularly affect our results since we almost exclusively experimented with the 10 person subset from our dataset.

Table 1: Authorship Attribution Results

Model	Number of Authors		
	10	20	50
$Char - CNNDynamic$	96.80%	90.60%	81.10%
$Char - CNNStatic$	96.70%	89.40%	78.70%

Unfortunately, we ran out of time and compute power to present results for the neural style transfer architecture.

9 Conclusion

9.1 Summary

In this paper, we present a reformulation and modified approach to the problem of Neural style transfer for non-parallel text. Our novelty comes from tying in the field of authorship attribution to recent Neural Style Transfer techniques. We conduct an extensive literature review of style transfer, authorship attribution, and evaluation metrics. We present a Twitter Influencer dataset, discuss methodology inspirations, and outline our architecture. We also present a new 2-part evaluation metric that takes into account transferring style and content preservation separately. Although we do not yet have results for the Neural Style Transfer architecture, we have accomplished a 96.80% accuracy using dynamic character-CNN embeddings for an authorship attribution model.

9.2 Future Work

If we had more time and compute resources, our first priority would be to get the Adverserial auto-encoder network fully trained, and adjust the hyperparameters for optimized results.

We would then test our proposed content preservation evaluation metric, as well as use Mechanical Turk (or a small army of undergrads) to have a gold-standard evaluation of our model.

We would like to explore more datasets (i.e. Project Gutenburg) for a famous author style transfer project.

10 Acknowledgements

We would like to thank Professor Bamman and Nick Altieri for their guidance in this project. Jacob Andreas's guest lecture was inspirational for this work.

11 References

1. Shen, Tianxiao, et al. "Style Transfer from Non-Parallel Text by Cross-Alignment." 2017.
2. Leon A Gatys, et al. Image style transfer using convolutional neural networks, 2015.
3. Sebastian Ruder, et al., "Character-level and Multi-channel Convolutional Neural Networks for Large-scale Authorship Attribution", 2016.
4. Alireza Makhzani, et al. "Adversarial Autoencoders", 2016.
5. Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. Controllable text generation.
6. Wei Xu, et al. "Paraphrasing for Style", 2012.
7. Marius Popescu, et al. "Comparing Statistical Similarity Measures for Stylistic Multivariate Analysis", 2009.
8. Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. "Authorship Attribution with Topic Models", 2014.
9. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "BLEU: a Method for Automatic Evaluation of Machine Translation", 2002.
10. Efstathios Stamatatos, "A survey of modern authorship attribution methods", 2009.