Style Transfer Your Language with Encoders (STYLE): Multi-Domain Style Transfer for Natural Language Processing

Connor Baumler, Marcus Daly, Jesulayomi Kupoluyi, Sophie Salomon
Case Western Reserve University
10900 Euclid Ave, Cleveland OH 44106
connor.baumler@case.edu, marcus.r.daly@case.edu,
jesulayomi.kupoluyi@case.edu, sophie.salomon@case.edu

Abstract

This project explores multi-domain style transfer for natural language processing, based on recent research for images. The goal is to encode styled text content to an objective latent distribution which can be decoded into different styles. To accomplish this we train LSTM-based autoencoders for each domain; these are trained using a common discriminator to elicit a shared latent distribution. After a piece of text is encoded, the decoders are used to introduce differing styles back into the encoded text. We performed several experiments using newspaper articles sourced from media outlets with different political affiliations. In this scenario, the facts in the article can be the latent distribution, and political framing is the style we seek to capture. In order to evaluate the efficacy of our styled reconstructions, we used the ROUGE metric because of its tolerance for non-parallel text and focus on recall which best suits this application. Although we ran into some challenges with generating semantically meaningful sentences from the latent representations, this project left us confident regarding the potential of this uncoupled autoencoder approach for NLP style transfer applications.

1. Introduction

Natural language processing has made unprecedented breakthroughs in machine translation and content generation. Deep learning models can be trained to reduce text to and generate text from latent semantic embeddings which can be used for myriad applications. We explore the use of autoencoders for multi-domain style transfer for natural language, where content is reduced to a semantic base and then converted into one of several different styles. We based this work on existing research which accomplished multi-domain style transfer for images. There are many potential uses for this technique, such as preserving desired levels of formality when translating into languages with strict register, simplifying legal texts and news articles for laymen, converting novels into versions accessible to children, automatically writing abstracts for journal papers, or changing news and entertainment content to target different types of audiences.

We focus on news articles based on the availability of a large corpus of texts to use for training data, the breadth of styles included, and the relative consistency of syntactic formatting which made it easier to isolate style from content (e.g. compared to poetry or transcripts of spoken language). In this case, we can treat relative political alignment as style, since partisan framing can result in significantly different interpretations even when the objective semantic meaning remains the same. This approach fit our dataset well, and allowed for less ambiguous human evaluation than other more nebulous styles. Nonetheless, our approach should extend to an arbitrary set of styles which can be applied to a given latent representation to convert it to styled content.

1.1. Background

Style transfer for NLP is the culmination of several areas of research which have been enabled by deep learning techniques. Early natural language processing for style focused on detection of tone or intent, which can broadly fall under the category of style. For example, training models to detect whether a movie review is positive or negative was an early classification attempt performed using NLP. Combining this with generative processes for text, particularly with conditional GANs, has advanced the work in style introduction for NLP text generation.

This work has parallels in style translation for image generation, although the architecture of the deep neural networks differs significantly between these domains. There has also been work in one-to-one style transfer between texts, where an adversarial approach is taken to translate a work in one style to a specific different style. The use of autoencoders to generate a latent space represenation which can be stylized is an interesting area of research which could have many applications if it continues to grow. In the following sections, we go more into depth on two papers in these areas that significantly influenced our project.

1.1.1 Multi-Domain Style Transfer

However, multi-domain style text transfer is more challenging, since it depends on texts of different style domains sharing an unstyled latent representation which can be separated from the styled content, then reapplied. In the case of multi-domain text style transfer, the basic semantic meaning of the text should be reflective of a latent distribution of content which may be context specific (such as summarizing Wikipedia articles or simplifying legal texts). The use of autoencoders is a natural technique to attempt this, with an objective semantic base that supports several style domains that are removed during encoding and introduced during decoding. We based this approach for multi-domain style transfer on a paper which uses this process for image style transfer [6].

The authors demonstrate the equivalence of using uncoupled autoencoders to learning a probabilistic coupling between domains of the same underlying latent distribution. The formal problem statement structures all elements X_i of a styled domain i as being generated with function f_i from the latent distribution Z given some domain-specific noise N_i:

$$X_i = f_i(Z, N_i)$$

The joint probability distribution of a domain instance given a generative process based on Z can be factored conditioned on the domain's relationship with the latent space.

In order to perform multi-domain style transfer in this way, some strong independence assumptions are necessary. Each domain must be effectively interchangeable from the perspective of the latent distribution, which can be thought of as a higher level structure that is foundational to all the domains. The inherent pitfall with this is that in order to achieve knowledge of this latent distribution, we must either have extensive initial knowledge of this structure, or learn it from the domains we want to model. Both approaches have shortcomings, with the first often being unrealistic and even introducing bias into the model. The second is more dependent on the specific domains being studied, and therefore may not correctly converge to the latent distribution based on unbalanced influence from the generated domains. However, there are many benefits to this approach. The paper specifically highlights the modularity, flexibility, efficiency, and generality of using autoencoders. In particular, with k autoecoders, it is possible to achieve all the pairwise style translations given a shared latent representation. From there, adding in new autoencoders allows style addition without having to train with each of the other style domains, making this a scalable solution for a large suite of style options. Using a discriminator in the latent space for adversarial training also means less customization is required to train each style domain autoencoder. These factors make it an appealing technique for style transfer given the presence of a consistent underlying structure.

Yang and Uhler used a generic adversarial model on the latent space for training, which is updated in two steps per sample iteration. With domain-specific encoder E_i and domain-specific decoder D_i, discriminator δ, weight hyperparameter λ, and sets of N samples $\{x\}, \{z\}, \{n\}$, perform gradient descent on the autoencoder components using the reconstruction loss

$$\frac{1}{N}\sum_{i=1}^{N} \|x_i - D(E(x_i))\|_2 + \lambda \text{log}(\delta(E(x_i)))$$

and gradient ascent on the discriminator using the discriminative loss

$$\frac{1}{N}\sum_{i=1}^{N} [\text{log}(\delta(E(x_i))) + \text{log}(1 - \delta(z_i, n_i))]$$

The paper goes on to mathematically demonstrate that their approach of using uncoupled autoencoders satisfies consistency and completeness properties.

1.1.2 Style Transfer for Text

There is no shortage of training data for NLP, but finding well-labeled, parallel data is much more difficult. Particularly for style transfer or sentiment modification, there is unlikely to be parallel texts availability for training. However, research by Shen, et al. suggests a strategy for using latent space representation for this and other translation applications in NLP with having any parallel texts [5]. Their use of autoencoders for text translation and generation suggested that this approach could be compatible with the approach for multi-domain style transfer for images using autoencoders described above.

The technique they used to reinforce the style transfer is called cross-alignment. This works by training a discriminator to distinguish between samples which started out in a different style, then got translated to the current style, and true samples of that style. There are many challenges associated with this, since the latent distribution is being used for style transfer but no parallel texts exists to train the autoencoder; the discriminator is supposed to just capture the stylistic differences. However, the paper describes two main strategies to overcome these challenges. The researchers found that using softmax distribution over the words as input, and Professor-Forcing to compare hidden states instead of words in the output elicited stronger results.

Although NLP tasks are often challenging, particularly when faced with non-parallel data, the notion of using shared latent space representation is very appealing, especially for style transfer. Although training autoencoders on text data is complicated, for many applications it is reasonable to assume a shared latent representation. Therefore, using this approach for multi-domain style transfer should be feasible, even if the specific nuances of achieving comprehensible sentences and text is inherently non-trivial.

2. Data

Because there is limited work in multi-domain style transfer, especially for NLP, we realized that a parallel-labeled style text dataset was unlikely to be publicly available. However, because our approach of training autoencoders on each style does not require parallel texts, we just needed an extensive enough corpus to capture a relatively consistent style. Finding data with relatively consistent format in other dimensions, including sincerity and syntax, was important so that we could isolate style differences to the greatest extent possible. For these reasons, we decided on using a collection of news articles sourced from over a dozen English-language newspapers published in the United States [2].

Since all the texts are newspaper articles, the content is broad in topic and contains writing by many different authors. However, the text is generally grouped into paragraphs of similar syntactic templates with limited slang or other confounding lexicon which may appear more frequently in other types of text. We ultimately chose six newspapers to train and test our style transfer technique, focusing on style as a reflection of political affiliation. For conservative content, we selected Fox News and Breitbart; the centrist picks are Reuters and NPR, and New York Times and CNN are the left-leaning sources.

Our premise for this choice was to distill political affiliation as a type of style which could be applied to the base semantics of an article. Since articles are based at some level on objective truth, the facts behind an article can be encoded to an embedded semantic meaning. However, each newspaper presents stories with a particular lens, often based on political framework which aligns with the values of the newspaper. Even when presenting the same facts, a paper's sentiment towards these facts can vary. In this way, each newspaper presents content through a political style which could be trained onto an uncoupled autoencoder. Other elements of writing style may be captured, but these should also reflect the voice and style of the newspaper as a whole, and are therefore reasonable inclusions for a style model of that paper besides just political preference.

3. Methods

3.1. Model

We based our model[1] on the two papers featured in the background which described using autoencoders for multi-domain style transfer and style-transfer for NLP. Essentially, we wanted to combine the two by training style encoders on input text which could be swapped around generically to perform multi-domain style transfer on natural language. We used bidirectional LSTMs with hyperbolic tangent activation functions to train our autoencoder. Our loss functions were based on the work done by Yang and Uhler [6].

Although we did see changes in the composition of the neural net based on the training as reflected in the loss function results, the incoherence of the sentence reconstructions suggests our model might benefit from adjustment to a Seq2Seq structure or from incorporating more of the adjustments made by Shen et al. to improve the quality of their styled text generation and decoding from latent space [5]. We might also need to adjust how we construct our latent space based on the repetitive outputs.

3.2. ROUGE

To evaluate our model, we use the metric ROUGE (Recall-Oriented Understudy for Gisting Evaluation).[3] ROUGE is designed to evaluate summaries of texts by comparing them to ideal summaries created by humans. Instead, we use it to compare our style-transferred text to the original text. We use three evaluations methods of ROUGE: ROUGE-N, ROUGE-L, and ROUGE-W.

3.2.1 ROUGE-N

ROUGE-N looks at n-gram co-occurrence. To do this, it calculates the n-gram recall between a summary and a set of reference summaries. Since we are only looking at an original and a transfered version of the text, we only have one "summary" with which to compare. This give us:

$$\text{ROUGE-N} = \frac{\sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{gram_n \in S} Count(gram_n)}$$

Here, n represents the length of $gram_n$ and $Count_{match}$ represents the maximum number of n-grams that co-occur in the original and transferred texts. For our purposes, we look at n-grams of length 1 to 4. This means that in our tables showing results, R1 through R4 represent ROUGE-N with $n = 1, 2, 3, 4$.

3.2.2 ROUGE-L

ROUGE-L looks at the longest common subsequence of words between versions of the text. The subsequence can skip entries in the sequence. For instance "the cat and dog" is a subsequence of "the cat, mouse, and dog." ROUGE-L looks at the longest common subsequence (LCS) between the original and transferred texts. Taking m to be the length of sequence X and n to be the length of sequence Y, ROUGE-L computes the recall, precision, and F1 score of this metric to be:

$$R_{lcs} = \frac{LCS(X,Y)}{m}$$
$$P_{lcs} = \frac{LCS(X,Y)}{n}$$
$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}}$$

3.2.3 ROUGE-W

Rouge-W looks at the weighted LCS. The weighting favors subsequences that don't skip as many words. For example, let's look at the sentence "I like dogs and cats. They are very nice." Let's say two style-transferred versions of the sentence are "I love dogs and cats. I think they're great.," and "Dogs are liked by me, and I like cats." Both of these have the LCS of "dogs and cats," but the first is clearly the better match as the subsequence is less spread out. So while the ROUGE-L scores may be the same, the ROUGE-W scores will favor the first version.

ROUGE-W's weighting function f will have the property that $f(x + y) > f(x) + f(y)$ which makes sure non-consecutive matches are scored worse than consecutive ones. The metrics can then be expressed as:

$$R_{wlcs} = f^{-1}\Big(\frac{WLCS(X,Y)}{f(m)}\Big)$$
$$P_{wlcs} = f^{-1}\Big(\frac{WLCS(X,Y)}{f(n)}\Big)$$
$$F_{wlcs} = \frac{(1+\beta^2)R_{wlcs}P_{wlcs}}{R_{wlcs}+\beta^2 P_{wlcs}}$$

4. Experiments

Our experiments focused on transferring styles from four different pairings of media sources that differed based on political affiliation. In order to train our each model, we first needed to select a pair of media sources whose shared latent distribution would represent the latent distribution of all 6 media sources. In order to determine the best starting pair, we varied the sources with which we started training. As the original pair will have to together represent the latent distribution for all six sources, we hypothesize that sources that are centrist will best capture the truth behind the content in the latent representation, as little political bias will be present in the latent representation so will not easily be encodable.

The sources we chose were Fox News and Breitbart for conservative content, Reuters and NPR for centrist content and the New York Times and CNN for left-leaning sources. We then tried to transfer text styles across all domains. The domains we experimented on were divided into two conservative, two centrist, two left-leaning and one left and one conservative pair. So we tried approximating the latent distribution using Breitbart and Fox News, Reuters and NPR, CNN and the New York Times and CNN and Fox news respectively in four different trials. Our intention with this was to find an initial domain pairing that produced the best result.

Our goal with the different experiments was to find the ideal initial domains that could lead to an accurate value of P_Z,(the shared latent space amongst all of the domains). So, when training, we select two initial domains, and assume that their shared latent space is the same as the shared latent space amongst all domains. The shared latent space between the two initial domains thus serves as a proxy for P_Z. Varying the selected domains essentially calculates the shared space between the selected domains, so finding the best initial pairing helps us to find the most accurate P_Z for all domains. The following sections go over hypotheses on starting from each of the initial four pairs as latent proxies.

4.1. Fox News and Breitbart Latent Distribution Proxy

As Fox News and Breitbart are both conservative media sources, we would expect the sentences here to be conservative-biased, as we learned the latent distribution from only conservative media sources.

4.2. Reuters and NPR Latent Distribution Proxy

As Reuters and NPR are both centrist media sources, we would expect the sentences here to be fairly unpolitical, as we learned the latent distribution from only centrist media washington — congressional republicans have the incoming trump administration could but a sudden loss of that could lead to chaos to stave off that outcome in another twist, donald eager to avoid an ugly they are not yet ready to divulge their strategy . ?@_ ?@_ ?@_ ?@_ ?@_

(a) Original text

although although although although although although although although although although although although message message phrase phrase solution solution delivering delivering delivering delivering orientation ?@_

(b) Transferred Text

Figure 1: Sample translation from New York Times to Breitbart starting with the Fox and Breitbart domains. Note that ?@_ is the padding token, which was placed at the end of each piece of text in preprocessing to pad each sequence to the same token length of 50 tokens.

sources, leaving any political bias to be a part of the random domain-specific latent distributions, not in the shared latent distribution. We hypothesize this would be the best latent distribution to use in order to capture only the actual content in the latent distribution.

4.3. CNN and New York Times Latent Distribution Proxy

As CNN and New York Times are both liberal media sources, we would expect the sentences here to be liberal-biased, as we learned the latent distribution from only liberal media sources.

4.4. CNN and Fox News Latent Distribution Proxy

As CNN is a liberal media source while Fox News is a conservative media source, we would expect sentences here to capture political bias, as we learned the latent distribution from two oppositely politically biased media sources.

5. Results

We evaluate all of our translators using ROUGE (as explained in section 3.2) on 300 articles held out separate from the training articles with each source style being tested with each target style. For an example, let's look at our results converting from New York Times to Breitbart starting with the Fox and Breitbart domains. As this translation goes from one politically biased source on the left side to another politically biased source but on the right side, this will determine whether we can retain underlying information while translating from one political bias to another. One of our evaluation sentences and its translation can be seen in Figure 1. The ROUGE results for this experiment can be seen in Table 1. As expected, the results were not very good. We find a very small number of matching 1-grams and no matching 2-, 3-, or 4- grams. This means that the longest common subsequences will always be of length 1 with similar results for the weighted common subsequences.

Method	Precision	Recall	F1
R1	0.04	0.03	0.03
R2	0.00	0.00	0.00
R3	0.00	0.00	0.00
R4	0.00	0.00	0.00
RL	0.07	0.05	0.06
RW	0.04	0.01	0.02

Table 1: Rouge results converting from New York Times to Breitbart starting with the Fox News and Breitbart domains

Method	Precision	Recall	F1
R1	0.11	0.01	0.02
R2	0.00	0.00	0.00
R3	0.00	0.00	0.00
R4	0.00	0.00	0.00
RL	0.13	0.02	0.03
RW	0.11	0.00	0.01

Table 2: Rouge results converting from New York Times to Breitbart starting with the Reuters and NPR domains

We also show results for latent spaces produced using the three other pairs of Reuters and NPR, CNN and New York Times, and CNN and Fox News in Tables 2, 3, and 4 respectively. Most notably from these results, we see that Table 2 shows the highest precision in R1, RL, and RW among all starting domain pairs. Even with generation that is not overly recognizable as news text, our results limitedly confirm our hypothesis that building a latent space from the pair of centrist media sources results in the model that retains the most information when translating between sources of varying political biases. It is important to note that, although

Method	Precision	Recall	F1
R1	0.08	0.04	0.05
R2	0.00	0.00	0.00
R3	0.00	0.00	0.00
R4	0.00	0.00	0.00
RL	0.12	0.07	0.09
RW	0.08	0.02	0.03

Table 3: Rouge results converting from New York Times to Breitbart starting with the CNN and New York Times domains

Method	Precision	Recall	F1
R1	0.02	0.02	0.02
R2	0.00	0.00	0.00
R3	0.00	0.00	0.00
R4	0.00	0.00	0.00
RL	0.03	0.03	0.03
RW	0.02	0.01	0.01

Table 4: Rouge results converting from New York Times to Breitbart starting with the CNN and Fox News domains

6. Conclusions and Future Work

To work towards being able to do multi-domain style transfer for text, we applied a computer vision technique using multiple autoencoders with a shared latent distribution. LSTMs take in text from various news publications and encode them into the latent space. From there, the decoder matching the desired output style is used to produce the final result.

We found limited success with this approach. Our results barely approached human readability let alone converting style or even conserving any real semantic meaning. Also, in some experiments, our model mysteriously collapsed (which can be seen in the charts of loss values at the end of this report). We believe that some of this challenge was due to how our underlying latent representation was distributed, as it appears that even for the autoencoders, they struggled to restore to anything meaningful. With the decoupled results for style transfer, it is natural that this pattern would persist. This is one of the biggest challenges in NLP.

Perhaps modifying our model to incorporate some of the state-of-the-art techniques used in NLP would improve our output. We also explored converting to a Seq2Seq model, so exploring that avenue in the future could make this approach work better. The nuances of the strategies used for text style transfer by Shen et al. make it clear that this is not a fundamental problem with our efforts, but rather an ongoing struggle with extracting a semantically meaningful sequence of words from a latent distribution. The sequential nature of NLP introduces many challenges that are not present in use of this approach for, e.g. computer vision applications.

Although our initial attempts were not very successful, this strategy has a lot of potential. With more refining, multi-domain style transfer could be used on a variety of applications where more than two style domains are required. We ran into many problems where NLP challenges overlapped with style-transfer challenges, which is consistent with the paper on one-to-one style transfer. Therefore, we are optimistic that, despite the shortcomings of our results, this technique can be extended to many of the example applications we posited throughout this report.

References

[1] https://github.com/marcusdaly/style/.

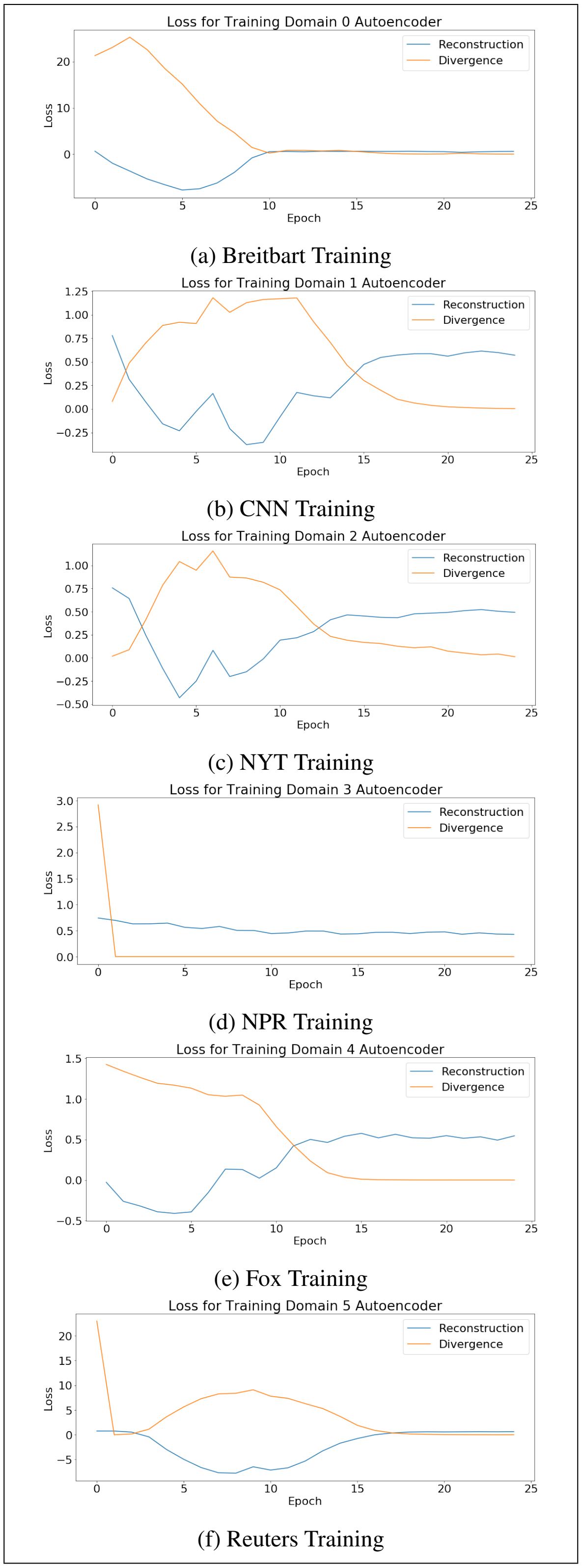
[2] Andrew Thompson. All the news: 143,000 articles from 15 american publications, 2017. Data retrieved from Kaggle, https://www.kaggle.com/snapcrack/all-the-news/.

[3] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[4] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA, 2013. Curran Associates Inc.

[5] T. Shen, T. Lei, R. Barzilay, and T. Jaakkola. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems*, pages 6830–6841, 2017.

[6] K. D. Yang and C. Uhler. Multi-domain translation by learning uncoupled autoencoders. *CoRR*, abs/1902.03515, 2019.



(a) Breitbart Training

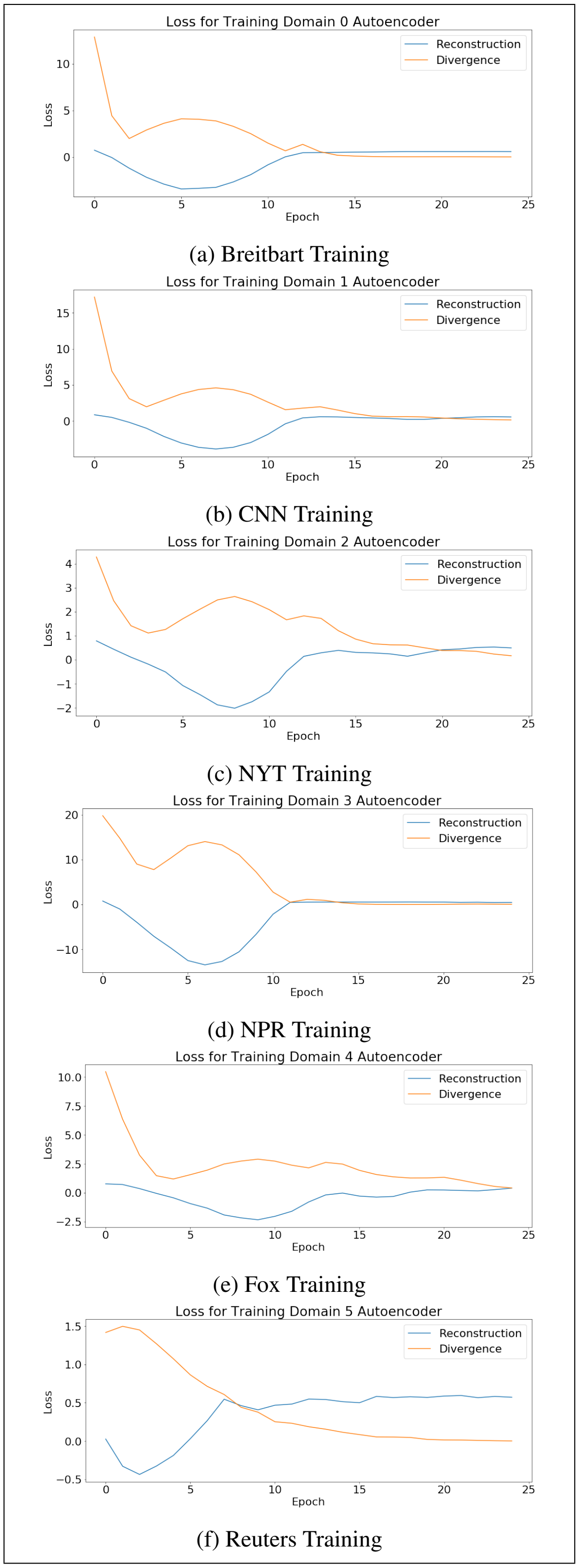
(b) CNN Training

(c) NYT Training

(d) NPR Training

(e) Fox Training

(f) Reuters Training



(a) Breitbart Training

(b) CNN Training

(c) NYT Training

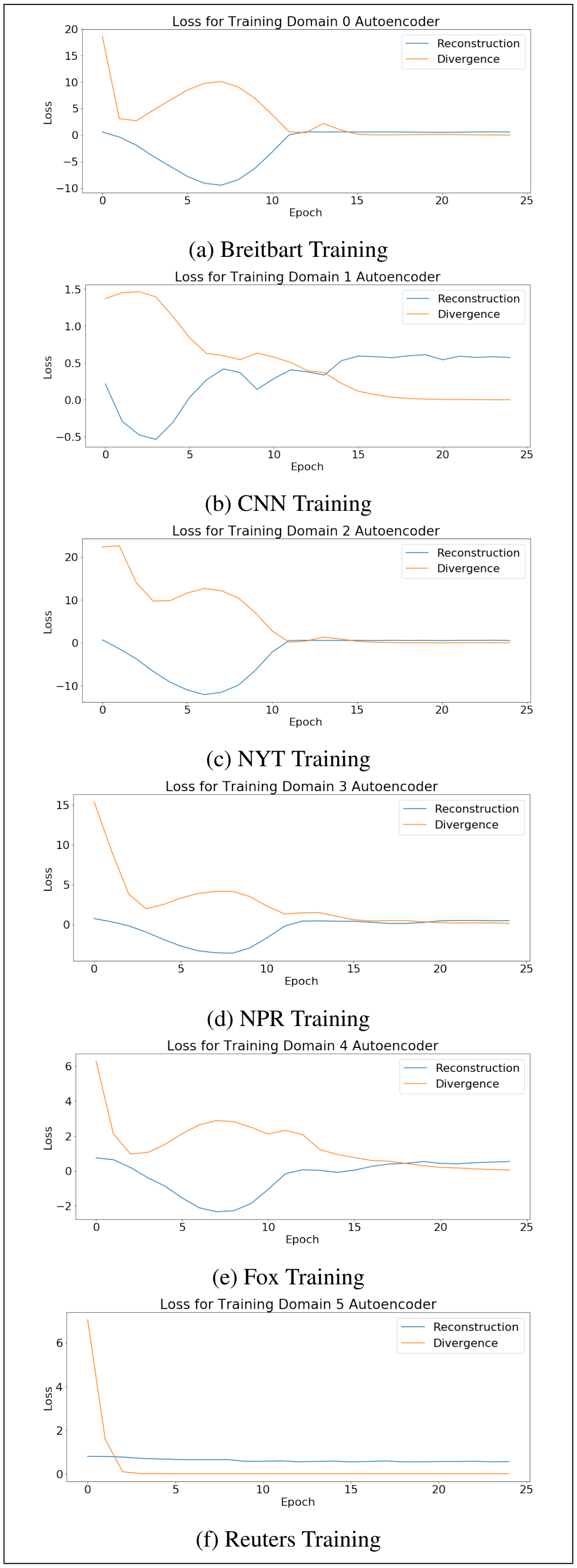
(d) NPR Training

(e) Fox Training

(f) Reuters Training

Figure 2: Learning Rates for the experiment with initial domains Fox News and Breitbart.

Figure 3: Learning Rates for the experiment with initial domains Reuters and NPR.



(a) Breitbart Training

(b) CNN Training

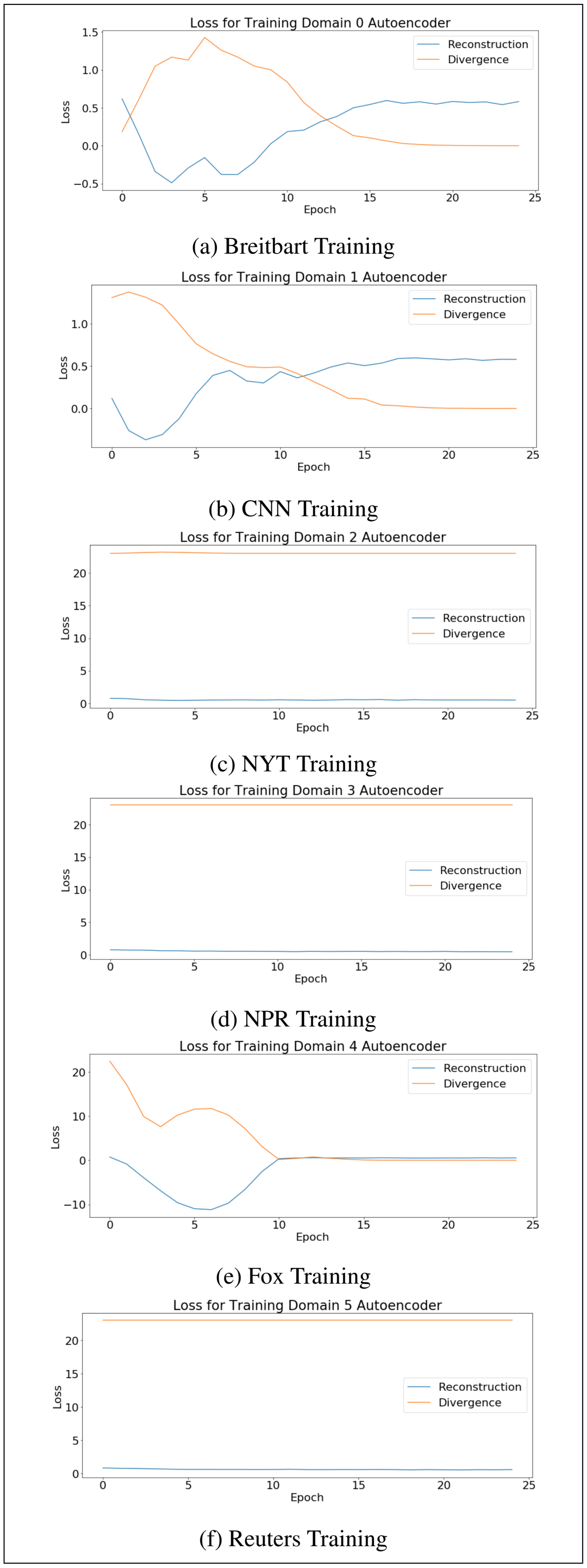
(c) NYT Training

(d) NPR Training

(e) Fox Training

(f) Reuters Training

Figure 4: Learning Rates for the experiment with initial domains CNN and NYT.



(a) Breitbart Training

(b) CNN Training

(c) NYT Training

(d) NPR Training

(e) Fox Training

(f) Reuters Training

Figure 5: Learning Rates for the experiment with initial domains CNN and Fox News.