Adversarially Regularized Autoencoders

Anonymous Authors[1]

Abstract

Deep latent variable models, trained using variational autoencoders or generative adversarial networks, are now a key technique for representation learning of continuous structures. However, applying similar methods to discrete structures, such as text sequences or discretized images, has proven to be more challenging. In this work, we propose a more flexible method for training deep latent variable models of discrete structures. Our approach is based on the recently proposed Wasserstein Autoencoder (WAE) which formalizes adversarial autoencoders as an optimal transport problem. We first extend this framework to model discrete sequences, and then further explore different learned priors targeting a controllable representation. Unlike many other latent variable generative models for text, this adversarially regularized autoencoder (ARAE) allows us to generate fluent textual outputs as well as perform manipulations in the latent space to induce change in the output space. Finally we show that the latent representation can be trained to perform unaligned textual style transfer, giving improvements both in automatic measures and human evaluation.

1. Introduction

Recent work on deep latent variable models, such as variational autoencoders (Kingma & Welling, 2014) and generative adversarial networks (Goodfellow et al., 2014), has shown significant progress in learning smooth representations of complex, high-dimensional continuous data such as images. These latent variable representations facilitate the ability to apply smooth transformations in latent space in order to produce complex modifications of generated outputs, while still remaining on the data manifold.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Unfortunately, learning similar latent variable models of discrete structures, such as text sequences or discretized images, remains a challenging problem. Initial work on VAEs for text has shown that optimization is difficult, as the generative model can easily degenerate into a unconditional language model (Bowman et al., 2015b). Recent work on generative adversarial networks (GANs) for text has mostly focused on dealing with the non-differentiable objective either through policy gradient methods (Che et al., 2017; Hjelm et al., 2017; Yu et al., 2017) or with the Gumbel-Softmax distribution (Kusner & Hernandez-Lobato, 2016). However, neither approach can yet produce robust representations directly.

In this work, we propose an autoencoder-based latent variable-model for discrete sequences (or structures) with explicit adversarial regularization (ARAE). The model learns an encoder from sequences to regularized latent representation. Like sequence VAEs, the model does not require using policy gradient training or continuous relaxations. Like GANs, the model provides flexibility in learning a prior through a parameterized generator.

The ARAE model can further be formalized under the recently-introduced Wasserstein Autoencoder (WAE) framework (Tolstikhin et al., 2018), which also generalizes the adversarial autoencoder (AAE) (Makhzani et al., 2015). This framework connects regularized autoencoders to an optimal transport objective for an implicit generative model. We extend this class of latent variable models to the case of discrete sequences, specifically showing that the autoencoder cross-entropy loss upper-bounds the cost objective in WAE. Under this setup, commonly-used discrete decoders such as RNNs, can be incorporated into the model.

Finally to handle non-trivial sequence examples, we consider several different (fixed and learned) prior distributions. These include a standard gaussian prior used in image models and in the AAE/WAE models, a learned parametric generator acting as a GAN in latent variable space, and a transfer-based parametric generator that is trained to ignore targeted attributes of the input. The last prior can be directly used for unaligned transfer tasks such as sentiment or style transfer.

Experiments apply ARAE to discretized images and text sequences. The latent variable model is able to generate

varied samples that can be quantitatively shown to cover the input spaces and to generate consistent image and sentence manipulations by moving around in the latent space via interpolation and offset vector arithmetic. When the ARAE model is trained with task-specific adversarial regularization, the model improves the current best results on sentiment transfer reported in Shen et al. (2017) and produces compelling outputs on a topic transfer task using only a single shared space. All code is available at *(removed for review)*.

2. Background and Notation

Discrete Autoencoder Define $\mathcal{X} = \mathcal{V}^n$ to be a set of discrete sequences where $\mathcal{V}$ is a vocabulary of symbols. Our discrete autoencoder will consist of two parameterized functions: a deterministic encoder function $\text{enc}_\phi : \mathcal{X} \mapsto \mathcal{Z}$ with parameters ϕ that maps from input to code space and a conditional decoder distribution $p_\psi(\mathbf{x} \mid \mathbf{z})$ over structures $\mathcal{X}$ with parameters ψ. The parameters are trained based on the cross-entropy reconstruction loss:

$$\mathcal{L}_{\text{rec}}(\phi, \psi) = -\log p_\psi(\mathbf{x} \mid \text{enc}_\phi(\mathbf{x}))$$

The choice of the encoder and decoder parameterization is problem-specific, for example we use RNNs for sequences. We use the notation, $\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} p_\psi(\mathbf{x} \mid \text{enc}_\phi(\mathbf{x}))$ for the (approximate) decoder mode.

Generative Adversarial Networks GANs are a class of parameterized implicit generative models (Goodfellow et al., 2014). The method approximates drawing samples from a true distribution $\mathbf{z} \sim \mathbb{P}_r$ by instead employing a noise sample $\mathbf{s}$ and a parameterized deterministic generator function $\tilde{\mathbf{z}} = g_\theta(\mathbf{s})$ to produce $\tilde{\mathbf{z}} \sim \mathbb{P}_g$. Initial work on GANs minimizes the Jensen-Shannon divergence between the distributions. Recent work on Wasserstein GAN (WGAN) (Arjovsky et al., 2017), replaces this with the *Earth-Mover* (Wasserstein-1) distance.

GAN training utilizes two separate models: a *generator* $g_\theta(\mathbf{s})$ maps a latent vector from some easy-to-sample noise distribution to a sample from a more complex distribution, and a critic/discriminator $f_w(\mathbf{z})$ aims to distinguish *real* data and *generated* samples from g_θ. Informally, the generator is trained to fool the critic, and the critic to tell real from generated. WGAN training uses the following min-max optimization over generator θ and critic w,

$$\min_\theta \max_{w \in \mathcal{W}} \mathbb{E}_{\mathbf{z} \sim \mathbb{P}_r}[f_w(\mathbf{z})] - \mathbb{E}_{\tilde{\mathbf{z}} \sim \mathbb{P}_g}[f_w(\tilde{\mathbf{z}})],$$

where $f_w : \mathcal{Z} \mapsto \mathbb{R}$ denotes the critic function, $\tilde{\mathbf{z}}$ is obtained from the generator, $\tilde{\mathbf{z}} = g_\theta(\mathbf{s})$, and $\mathbb{P}_r$ and $\mathbb{P}_g$ are real and generated distributions. If the critic parameters w are restricted to an 1-Lipschitz function set $\mathcal{W}$, this term correspond to minimizing Wasserstein-1 distance $W(\mathbb{P}_r, \mathbb{P}_g)$. We use a naive approximation to enforce this property by weight-clipping, i.e. $w = [-\epsilon, \epsilon]^d$ (Arjovsky et al., 2017).

3. Adversarially Regularized Autoencoder

ARAE combines a discrete autoencoder with a GAN-regularized latent representation. Intuitively, this method aims to provide smoother hidden encoding for discrete sequences with a flexible prior. In the next section we show how this simple network can be formally interpreted as a latent variable model under the Wasserstein Autoencoder framework. Our model consists of a discrete encoder regularized with a prior distribution,

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

Here W is the Wasserstein distance between $\mathbb{P}_Q$, the distribution from a discrete encoder model ($\text{enc}_\phi(\mathbf{x})$ where $\mathbf{x} \sim \mathbb{P}_\star$), and $\mathbb{P}_{\mathbf{z}}$ a given prior distribution. As above, the W function is computed with an embedded critic function which is optimized adversarially to the generator and encoder.[1]

The model is trained with coordinate descent across: (1) the encoder and decoder to minimize reconstruction, (2) the critic function to approximate the W term, (3) the encoder adversarially to the critic to minimize W:

$$\begin{aligned}
&1)\ \min_{\phi, \psi} && \mathcal{L}_{\text{rec}}(\phi, \psi) = \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_\star}\left[\log p_\psi(\mathbf{x} \mid \text{enc}_\phi(\mathbf{x}))\right] \\
&2)\ \max_{w \in \mathcal{W}} && \mathcal{L}_{\text{cri}}(w) = \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_\star}\left[f_w(\text{enc}_\phi(\mathbf{x}))\right] - \mathbb{E}_{\tilde{\mathbf{z}} \sim \mathbb{P}_{\mathbf{z}}}\left[f_w(\tilde{\mathbf{z}})\right] \\
&3)\ \min_{\phi} && \mathcal{L}_{\text{enc}}(\phi) = \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_\star}\left[f_w(\text{enc}_\phi(\mathbf{x}))\right] - \mathbb{E}_{\tilde{\mathbf{z}} \sim \mathbb{P}_{\mathbf{z}}}\left[f_w(\tilde{\mathbf{z}})\right]
\end{aligned}$$

The full training algorithm is shown in Algorithm 1.

Empirically we found that the choice of the prior distribution $\mathbb{P}_{\mathbf{z}}$ strongly impacted the performance of the model. The simplest choice is to use a fixed distribution such as a Gaussian $\mathcal{N}(0, I)$, which yields a discrete version of the adversarial autoencoder (AAE). However in practice this choice is seemingly too constrained and suffers from mode-collapse.

Instead we exploit the adversarial setup and use learned prior parameterized through a generator model. This is analogous to the use of more complex priors in VAEs through, for example, applying autoregressive flows to a sample from a simple prior (Chen et al., 2017). Specifically we introduce a generator model, $g_\theta(\mathbf{s})$ over noise $\mathbf{s} \sim \mathcal{N}(0, I)$ to act as an implicit prior distribution $\mathbb{P}_{\mathbf{z}}$. We optimize its parameters θ as part of training in Step 3. The full model is shown in Figure 1.[2]

[1] Other GANs could be used for this optimization. Experimentally we found that WGANs to be more stable than other models.

[2] The downside of this approach is that the latent variable $\mathbf{z}$ is

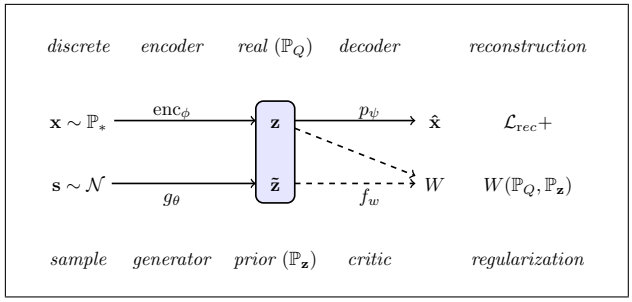

Figure 1: ARAE architecture. A discrete sequence $\mathbf{x}$ is encoded and decoded to produce $\hat{\mathbf{x}}$. A noise sample $\mathbf{s}$ is passed though a generator g_θ (possibly the identity) to produce a prior. The critic function f_w is only used at training to enforce regularization W. The model produce discrete samples $\mathbf{x}$ from noise $\mathbf{s}$. Section 5 relates these samples $\mathbf{x} \sim \mathbb{P}_\psi$ to $\mathbf{x} \sim \mathbb{P}_\star$.

Algorithm 1 ARAE Training

for each training iteration **do**

(1) Train the encoder/decoder for reconstruction (ϕ, ψ)

Sample $\{\mathbf{x}^{(i)}\}_{i=1}^m \sim \mathbb{P}_\star$ and compute $\mathbf{z}^{(i)} = \text{enc}_\phi(\mathbf{x}^{(i)})$

Backprop loss, $\mathcal{L}_{\text{rec}} = -\frac{1}{m}\sum_{i=1}^m \log p_\psi(\mathbf{x}^{(i)} | \mathbf{z}^{(i)})$

(2) Train the critic (w)

Sample $\{\mathbf{x}^{(i)}\}_{i=1}^m \sim \mathbb{P}_\star$ and $\{\mathbf{s}^{(i)}\}_{i=1}^m \sim \mathcal{N}(0, \mathbf{I})$

Compute $\mathbf{z}^{(i)} = \text{enc}_\phi(\mathbf{x}^{(i)})$ and $\tilde{\mathbf{z}}^{(i)} = g_\theta(\mathbf{z}^{(i)})$

Backprop loss $-\frac{1}{m}\sum_{i=1}^m f_w(\mathbf{z}^{(i)}) + \frac{1}{m}\sum_{i=1}^m f_w(\tilde{\mathbf{z}}^{(i)})$

Clip critic w to $[-\epsilon, \epsilon]^d$.

(3) Train the encoder/generator adversarially (ϕ, θ)

Sample $\{\mathbf{x}^{(i)}\}_{i=1}^m \sim \mathbb{P}_\star$ and $\{\mathbf{s}^{(i)}\}_{i=1}^m \sim \mathcal{N}(0, \mathbf{I})$

Compute $\mathbf{z}^{(i)} = \text{enc}_\phi(\mathbf{x}^{(i)})$ and $\tilde{\mathbf{z}}^{(i)} = g_\theta(\mathbf{s}^{(i)})$.

Backprop loss $\frac{1}{m}\sum_{i=1}^m f_w(\mathbf{z}^{(i)}) - \frac{1}{m}\sum_{i=1}^m f_w(\tilde{\mathbf{z}}^{(i)})$

end for

Extension: Unaligned Transfer Regularization of the latent space makes it more adaptable for direct continuous optimization that would be difficult over discrete sequences. For example, consider the problem of unaligned transfer, where we want to change an attribute of a discrete input without aligned examples, e.g. to change the topic or sentiment of a sentence. Define this attribute as y and redefine the decoder to be conditional $p_\psi(\mathbf{x} \mid \mathbf{z}, y)$.

To adapt ARAE to this setup, we modify the objective to learn to remove attribute distinctions from the prior (i.e. we want the prior to encode all the relevant information *except* about y). Following similar techniques from other domains, notably in images (Lample et al., 2017) and video modeling (Denton & Birodkar, 2017), we introduce a latent space attribute classifier:

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

now much less constrained. However we find experimentally that using a a simple MLP for g_θ significantly regularizes the encoder RNN.

Algorithm 2 ARAE Transfer Extension

Each loop additionally:

(2b) Train attribute classifier (u)

Sample $\{\mathbf{x}^{(i)}\}_{i=1}^m \sim \mathbb{P}_\star$, lookup $y^{(i)}$, and compute $\mathbf{z}^{(i)} = \text{enc}_\phi(\mathbf{x}^{(i)})$

Backprop loss $-\frac{1}{m}\sum_{i=1}^m \log p_u(y^{(i)} | \mathbf{z}^{(i)})$

(3b) Train the encoder adversarially (ϕ)

Sample $\{\mathbf{x}^{(i)}\}_{i=1}^m \sim \mathbb{P}_\star$, lookup $y^{(i)}$, and compute $\mathbf{z}^{(i)} = \text{enc}_\phi(\mathbf{x}^{(i)})$

Backprop loss $-\frac{1}{m}\sum_{i=1}^m \log p_u(1 - y^{(i)} \mid \mathbf{z}^{(i)})$

where $\mathcal{L}_{\text{class}}(\phi, u)$ is the loss of a classifier $p_u(y \mid \mathbf{z})$ from latent variable to labels (in our experiments we always set $\lambda^{(2)} = 1$). This requires two more update steps: (2b) training the classifier, and (3b) adversarially training the encoder to this classifier. This algorithm is shown in Algorithm 2.

4. Theoretical Properties

Standard GANs implicitly minimize a divergence measure (e.g. f-divergence or Wasserstein distance) between the true/model distributions. In our case however, we implicitly minimize the divergence between *learned* code distributions, and it is not clear if this training objective is matching the distributions in the original discrete space. Tolstikhin et al. (2018) recently showed that this style of training is minimizing the Wasserstein distance between the data distribution $\mathbb{P}_\star$ and the model distribution $\mathbb{P}_\psi$ with latent variables (with density $p_\psi(\mathbf{x}) = \int_\mathbf{z} p_\psi(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$).

In this section we apply the above result to the discrete case and show that the ARAE loss minimizes an upper bound on the *total variation distance* between $\mathbb{P}_\star$ and $\mathbb{P}_\psi$.

Definition 1 (Kantorovich's formulation of optimal transport)**.** *Let $\mathbb{P}_\star, \mathbb{P}_\psi$ be distributions over $\mathcal{X}$, and further let $c(\mathbf{x}, \mathbf{y}) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ be a cost function. Then the* optimal transport (OT) *problem is given by*

$$W_c(\mathbb{P}_\star, \mathbb{P}_\psi) = \inf_{\Gamma \in \mathcal{P}(\mathbf{x}\sim\mathbb{P}_\star, \mathbf{y}\sim\mathbb{P}_\psi)} \mathbb{E}_{\mathbf{x},\mathbf{y}\sim\Gamma}[c(\mathbf{x}, \mathbf{y})]$$

where $\mathcal{P}(\mathbf{x} \sim \mathbb{P}_\star, \mathbf{y} \sim \mathbb{P}_\psi)$ is the set of all joint distributions of $(\mathbf{x}, \mathbf{y})$ with marginals $\mathbb{P}_\star$ and $\mathbb{P}_\psi$.

In particular, if $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_p^p$ then $W_c(\mathbb{P}_\star, \mathbb{P}_\psi)^{\frac{1}{p}}$ is the Wasserstein-p distance between $\mathbb{P}_\star$ and $\mathbb{P}_\psi$. Now suppose we utilize a latent variable model to fit the data, i.e. $\mathbf{z} \sim \mathbb{P}_\mathbf{z}, \mathbf{x} \sim \mathbb{P}_\psi(\mathbf{x}|\mathbf{z})$. Then Tolstikhin et al. (2018) prove the following theorem:

Theorem 1. *Let $G_\psi : \mathcal{Z} \to \mathcal{X}$ be a deterministic function (parameterized by ψ) from the latent space $\mathcal{Z}$ to data space $\mathcal{X}$ that induces a dirac distribution $\mathbb{P}_\psi(\mathbf{x}|\mathbf{z})$ on $\mathcal{X}$, i.e. $p_\psi(\mathbf{x}|\mathbf{z}) = \mathbb{1}\{\mathbf{x} = G_\psi(\mathbf{z})\}$. Let $Q(\mathbf{z}|\mathbf{x})$ be any conditional distribution on $\mathcal{Z}$ with density $p_Q(\mathbf{z}|\mathbf{x})$. Define its marginal*

to be $\mathbb{P}_Q$, which has density $p_Q(\mathbf{x}) = \int_{\mathbf{x}} p_Q(\mathbf{z}|\mathbf{x})p_\star(\mathbf{x})d\mathbf{x}$). Then,

$$W_c(\mathbb{P}_\star, \mathbb{P}_\psi) = \inf_{Q(\mathbf{z}|\mathbf{x}):\mathbb{P}_Q=\mathbb{P}_\mathbf{z}} \mathbb{E}_{\mathbb{P}_\star}\mathbb{E}_{Q(\mathbf{z}|\mathbf{x})}[c(\mathbf{x}, G_\psi(\mathbf{z}))]$$

Theorem 1 essentially says that learning an autoencoder can be interpreted as learning a latent variable generative model, as long as we ensure that the marginalized encoded space is the same as the prior. This provides theoretical justification for adversarial autoencoders (Makhzani et al., 2015), and Tolstikhin et al. (2018) used the above to train deep generative models of images by minimizing the Wasserstein-2 distance (i.e. squared loss between real/generated images). We now apply Theorem 1 to discrete autoencoders trained with cross-entropy loss.

Corollary 1 (Discrete case)**.** *Suppose $\mathbf{x} \in \mathcal{X}$ where $\mathcal{X}$ is the set of all one-hot vectors of length n, and let $f_\psi : \mathcal{Z} \to \Delta^{n-1}$ be a deterministic function that goes from the latent space $\mathcal{Z}$ to the $n-1$ dimensional simplex Δ^{n-1}. Further let $G_\psi : \mathcal{Z} \to \mathcal{X}$ be a deterministic function such that $G_\psi(\mathbf{z}) = \arg\max_{\mathbf{w}\in\mathcal{X}} \mathbf{w}^\top f_\psi(\mathbf{z})$, and as above let $\mathbb{P}_\psi(\mathbf{x}|\mathbf{z})$ be the dirac distribution derived from G_ψ such that $p_\psi(\mathbf{x}|\mathbf{z}) = \mathbb{1}\{\mathbf{x} = G_\psi(\mathbf{z})\}$. Then the following is an upper bound on $\|\mathbb{P}_\psi - \mathbb{P}_\star\|_{TV}$, the total variation distance between $\mathbb{P}_\star$ and $\mathbb{P}_\psi$:*

$$\inf_{Q(\mathbf{z}|\mathbf{x}):\mathbb{P}_Q=\mathbb{P}_\mathbf{z}} \mathbb{E}_{\mathbb{P}_\star}\mathbb{E}_{Q(\mathbf{z}|\mathbf{x})}\Big[-\frac{2}{\log 2}\log \mathbf{x}^\top f_\psi(\mathbf{z})\Big]$$

The proof is in Appendix A. Since in our case $n = |\mathcal{V}|^m$ (i.e. sentences of length at most m), the total variational (TV) distance is given by

$$\|\mathbb{P}_\psi - \mathbb{P}_\star\|_{\text{TV}} = \frac{1}{2}\sum_{\mathbf{x}\in\mathcal{V}^m} |p_\psi(\mathbf{x}) - p_\star(\mathbf{x})|$$

This is an interesting alternative to the usual maximum likelihood approach which instead minimizes $\text{KL}(\mathbb{P}_\star, \mathbb{P}_\psi)$.[3] It is also clear that $-\log \mathbf{x}^\top f_\psi(\mathbf{z}) = -\log p_\psi(\mathbf{x}|\mathbf{z})$, the standard autoencoder cross-entropy loss at the sentence level with f_ψ as the decoder. As the above objective is hard to minimize directly, we follow Tolstikhin et al. (2018) and consider an easier objective by (i) restricting $Q(\mathbf{z}|\mathbf{x})$ to a family of distributions induced by a deterministic encoder parameterized by ϕ, and (ii) using a Langrangian relaxation of the constraint $\mathbb{P}_Q = \mathbb{P}_\mathbf{z}$. In particular, letting $Q(\mathbf{z}|\mathbf{x}) = \mathbb{1}\{\mathbf{z} = \text{enc}_\phi(\mathbf{x})\}$ be the dirac distribution induced by a deterministic encoder (with associated marginal $\mathbb{P}_\phi$), the objective is given by

$$\min_{\phi,\psi} \mathbb{E}_{\mathbb{P}_\star}[-\log p_\psi(\mathbf{x}|\text{enc}_\phi(\mathbf{z}))] + \lambda W(\mathbb{P}_\phi, \mathbb{P}_\mathbf{z})$$

[3] KL-divergence is itself an upper bound on twice the square of TV distance, due to Pinsker's inequality.

Note that our minimizing the Wasserstein distance in the *latent* space $W(\mathbb{P}_\phi, \mathbb{P}_\mathbf{z})$ is independent from the Wassertein distance minimization in the *output* space in WAEs. Finally, instead of using a fixed prior (which led to mode-collapse in our experiments) we parameterize $\mathbb{P}_\mathbf{z}$ implicitly by transforming a simple random variable with a generator (i.e. $\mathbf{s} \sim \mathcal{N}(0, I), \mathbf{z} = g_\theta(\mathbf{s})$). This recovers the ARAE objective from the previous section.

5. Methods and Architectures

We experiment ARAE on three general setups: (1) a small model using discretized images trained on the binarized version of MNIST, (2) a model for text sequences trained on the Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015a), and (3) a model trained for text transfer trained on the Yelp/Yahoo datasets for unaligned sentiment/topic transfer. For the experiments choosing the learned-prior option, a same generator architecture is utilized, g_θ. The generator architecture uses a low dimensional $\mathbf{s}$ with a Gaussian prior $\mathbf{s} \sim \mathcal{N}(0, \mathbf{I})$, and maps it to $\mathbf{z}$. Both the critic f_w and the generator g_θ are parameterized as MLPs.

The *image* model uses a fully-connected NN to encode/decode binarized images. Here $\mathcal{X} = \{0, 1\}^n$ where n is the image size. The encoder used is a feed-forward MLP network mapping from $\{0, 1\}^n \mapsto \mathbb{R}^m$, $\text{enc}_\phi(\mathbf{x}) = \text{MLP}(\mathbf{x}; \phi) = \mathbf{z}$. The decoder predicts each pixel in $\mathbf{x}$ as a parameterized logistic regression, $p_\psi(\mathbf{x} \mid \mathbf{z}) = \prod_{j=1}^n \sigma(\mathbf{h})^{x_j}(1 - \sigma(\mathbf{h}))^{1-x_j}$ where $\mathbf{h} = \text{MLP}(\mathbf{z}; \psi)$.

The *text* model uses a recurrent neural network (RNN) for both the encoder and decoder. Here $\mathcal{X} = \mathcal{V}^n$ where n is the sentence length and $\mathcal{V}$ is the vocabulary of the underlying language. We define $\text{enc}_\phi(\mathbf{x}) = \mathbf{z}$ to be the last hidden state of an encoder RNN (with parameters ϕ). For decoding we feed $\mathbf{z}$ as an additional input to the decoder RNN at each time step, and calculate the distribution over $\mathcal{V}$ at each time step via softmax, $p_\psi(\mathbf{x} \mid \mathbf{z}) = \prod_{j=1}^n \text{softmax}(\mathbf{W}h_j + \mathbf{b})_{x_j}$ where $\mathbf{W}$ and $\mathbf{b}$ are parameters (part of ψ) and h_j is the decoder RNN hidden state. Finding the most likely sequence $\tilde{\mathbf{x}}$ under this distribution is intractable, but it is possible to approximate it using greedy search or beam search. In our experiments we use the LSTM architecture (Hochreiter & Schmidhuber, 1997) for both the encoder/decoder and decode using greedy search. The *text transfer* model uses the same architecture as the text model but extends it with a classifier $p_u(y|\mathbf{z})$ which is modeled using an MLP (with parameters u) and trained to minimize cross-entropy.

We further compare our approach with a standard autoencoder (AE) and the cross-aligned autoencoder (Shen et al., 2017) for transfer. In both our ARAE and standard AE experiments, the encoder output is normalized to lie on the

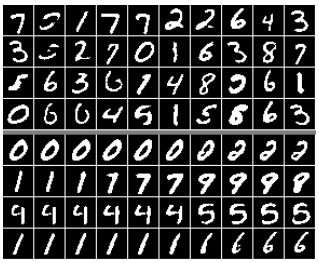

Figure 2: Image samples. The top block shows output generation of the decoder for random noise samples; the bottom block shows sample interpolation results.

Data for LM	Reverse PPL
Real data	27.4
LM samples	90.6
AE samples	97.3
ARAE samples	82.2

Table 1: Perplexity (lower is better) of language models trained on the synthetic samples from a ARAE/AE/LM, and evaluated on real data (Reverse PPL).

unit sphere, and the generator output is bounded to lie in $(-1, 1)^n$ by the $\tanh$ function at output layer.

We also tried to compare against baseline generative models, such as the VAE and AAE, but found it to be challenging. Note that learning deep latent models for text sequences has been a significantly more challenging empirical problem than for images. Standard models such as VAEs suffer from significant optimization issues that have been widely documented. We performed significant experiments with recurrent VAE, introduced by (Bowman et al., 2015b), as well as the adversarial autoencoder (AAE) (Makhzani et al., 2015). Despite extensive parameter tuning we found that neither model was able to learn meaningful latent representations—the VAE simply ignored the latent code and the AAE experienced mode-collapse and repeatedly generated the same samples. These results lead us to believe that both learning a parameterized prior through g_θ and using adversarial regularization are necessary for these models. Appendix F includes detailed descriptions of the hyperparameters, model architecture, and training regimes.

6. Experiments

6.1. Distributional Coverage

The results from Section 4 show that $\mathbb{P}_\psi$ is trained to approximate the true data distribution over discrete sequences $\mathbb{P}_\star$. While it is difficult to test for this property directly (as is the case with most GAN models), we can take samples from model to test coverage of the data space. Figure 2 (top) shows a set of samples from discretized MNIST and Appendix C shows a set of generations from the text ARAE.

A common quantitative measure of sample quality for generative models is to evaluate a strong surrogate model trained on its generated samples. While there are pitfalls of this style of evaluation methods (Theis et al., 2016), it has provided a starting point for image generation models. Here we use a similar method for text generation, which we call *reverse perplexity*. We generate 100k samples from each of the models, train an RNN language model on generated samples and evaluate perplexity on held-out data.[4] While similar metrics for images (e.g. Parzen windows) have been shown to be problematic, we argue that this is less of an issue for text as RNN language models achieve state-of-the-art perplexities on text datasets.

Positive	great indoor mall .
⇒ ARAE	no smoking mall .
⇒ Cross-AE	terrible outdoor urine .
Positive	it has a great atmosphere , with wonderful service .
⇒ ARAE	it has no taste , with a complete jerk .
⇒ Cross-AE	it has a great horrible food and run out service .
Positive	we came on the recommendation of a bell boy and the food was amazing .
⇒ ARAE	we came on the recommendation and the food was a joke .
⇒ Cross-AE	we went on the car of the time and the chicken was awful .
Negative	hell no !
⇒ ARAE	hell great !
⇒ Cross-AE	incredible pork !
Negative	small , smokey , dark and rude management .
⇒ ARAE	small , intimate , and cozy friendly staff .
⇒ Cross-AE	great , , , chips and wine .
Negative	the people who ordered off the menu did n't seem to do much better .
⇒ ARAE	the people who work there are super friendly and the menu is good .
⇒ Cross-AE	the place , one of the office is always worth you do a business .

Table 2: Sentiment transfer results, where we transfer from positive to negative sentiment (Top) and negative to positive sentiment (Bottom). Original sentence and transferred output (from ARAE and the Cross-Aligned AE (from Shen et al. (2017)) of 6 randomly-drawn examples.

Table 1 shows this measure for (i) ARAE, (ii) an autoencoder,[5] (iii) an RNN LM trained on the same data, and (iv) the real training set. We further find that with a fixed prior, the reverse PPL of an AAE-style text model (Makhzani et al., 2015) was quite high (980) due to mode-collapse. All models are of the same size to allow for fair comparison. Training directly on real data (understandably) outperforms training on generated data by a large margin. Surprisingly however, training on ARAE samples outperforms training on LM/AE samples.

6.2. Unaligned Text Style Transfer

Next we evaluate the model in the context of a learned adversarial prior, as described in Section 3. We experiment with two unaligned text transfer tasks: (i) transfer of sentiment on the Yelp corpus, and (ii) topic on the Yahoo corpus (Zhang et al., 2015). For sentiment we follow the setup of Shen et al. (2017) and split the Yelp corpus into two sets of unaligned positive and negative reviews. We train ARAE

[4] We also found this metric to be helpful for early-stopping.

[5] To "sample" from an AE we fit a multivariate Gaussian to the code space after training and generate code vectors from this Gaussian to decode back into sentence space.

Model	Automatic Evaluation			
	Transfer	BLEU	PPL	Reverse PPL
Cross-Aligned AE	77.1%	17.75	65.9	124.2
AE	59.3%	37.28	31.9	68.9
ARAE, $\lambda_a^{(1)}$	73.4%	31.15	29.7	70.1
ARAE, $\lambda_b^{(1)}$	81.8%	20.18	27.7	77.0

Model	Human Evaluation		
	Transfer	Similarity	Naturalness
Cross-Aligned AE	57%	3.8	2.7
ARAE, $\lambda_b^{(1)}$	74%	3.7	3.8

Table 3: Experiments on sentiment transfer. Left shows the automatic metrics (Transfer/BLEU/PPL/Reverse PPL) while right shows human evaluation metrics (Transfer/Similarity/Naturalness). Cross-Aligned AE is from (Shen et al., 2017)

with two separate RNNs, one for positive, $p(\mathbf{x}|\mathbf{z}, y = 1)$, and one for negative sentiment $p(\mathbf{x}|\mathbf{z}, y = 0)$, and incorporate adversarial training of the encoder to remove sentiment information from the prior. Transfer corresponds to encoding sentences of one class and decoding, greedily, with the opposite decoder.

Experiments compare against the cross-aligned AE of Shen et al. (2017) and also an AE trained without the adversarial regularization. For ARAE, we experimented with different $\lambda^{(1)}$ weighting on the adversarial loss (see section 4) with $\lambda_a^{(1)} = 1, \lambda_b^{(1)} = 10$. Both use $\lambda^{(2)} = 1$. Experimentally the adversarial regularization enhances transfer and perplexity, but tends to make the transferred text less similar to the original, compared to the AE. Randomly selected example sentences are shown in Table 2 and additional outputs are available in Appendix G.

Table 3 shows quantitative evaluation. We use four automatic metrics: (i) Transfer: how successful the model is at altering sentiment based on an automatic classifier (we use the `fastText` library (Joulin et al., 2016)). (ii) BLEU: the consistency between the transferred text and the original; (iii) Perplexity: the fluency of the generated text; (iv) Reverse Perplexity: measuring the extent to which the generations are representative of the underlying data distribution. Both perplexity numbers are obtained by training an RNN language model. Table 3 (bottom) shows human evaluations on the cross-aligned AE and our best ARAE model. We randomly select 1000 sentences (500/500 positive/negative), obtain the corresponding transfers from both models, and ask crowdworkers to evaluate the sentiment (Positive/Neutral/Negative) and naturalness (1-5, 5 being most natural) of the transferred sentences. We create a separate task in which we show the original and the transferred sentences, and ask them to evaluate the similarity based on sentence structure (1-5, 5 being most similar). We explicitly request the reader to disregard sentiment in similarity assessment.

Science	what is an event horizon with regards to black holes ?
⇒ Music	what is your favorite sitcom with adam sandler ?
⇒ Politics	what is an event with black people ?
Science	take 1ml of hcl (concentrated) and dilute it to 50ml .
⇒ Music	take em to you and shout it to me
⇒ Politics	take bribes to islam and it will be punished .
Science	just multiply the numerator of one fraction by that of the other .
⇒ Music	just multiply the fraction of the other one that 's just like it .
⇒ Politics	just multiply the same fraction of other countries .
Music	do you know a website that you can find people who want to join bands ?
⇒ Science	do you know a website that can help me with science ?
⇒ Politics	do you think that you can find a person who is in prison ?
Music	all three are fabulous artists , with just incredible talent ! !
⇒ Science	all three are genetically bonded with water , but just as many substances , are capable of producing a special case .
⇒ Politics	all three are competing with the government , just as far as i can .
Music	but there are so many more i can 't think of !
⇒ Science	but there are so many more of the number of questions .
⇒ Politics	but there are so many more of the can i think of today .
Politics	republicans : would you vote for a cheney / satan ticket in 2008 ?
⇒ Science	guys : how would you solve this question ?
⇒ Music	guys : would you rather be a good movie ?
Politics	4 years of an idiot in office + electing the idiot again = ?
⇒ Science	4 years of an idiot in the office of science ?
⇒ Music	4) <unk> in an idiot , the idiot is the best of the two points ever !
Politics	anyone who doesnt have a billion dollars for all the publicity cant win .
⇒ Science	anyone who doesnt have a decent chance is the same for all the other .
⇒ Music	anyone who doesnt have a lot of the show for the publicity .

Table 4: Random samples from Yahoo topic transfer. Note the first row is from ARAE trained on titles while the following ones are from replies.

Model	Medium	Small	Tiny
Supervised Encoder	65.9%	62.5%	57.9%
Semi-Supervised AE	68.5%	64.6%	59.9%
Semi-Supervised ARAE	70.9%	66.8%	62.5%

Table 5: Semi-Supervised accuracy on the natural language inference (SNLI) test set, respectively using 22.2% (medium), 10.8% (small), 5.25% (tiny) of the supervised labels of the full SNLI training set (rest used for unlabeled AE training).

The same method can be applied to other style transfer tasks, for instance the more challenging Yahoo QA data (Zhang et al., 2015). For Yahoo we chose 3 relatively distinct topic classes for transfer: Science & Math, Entertainment & Music, and Politics & Government. As the dataset contains both questions and answers, we separated our experiments into titles (questions) and replies (answers). The qualitative results are showed in Table 4. See Appendix G for additional generation examples.

6.3. Semi-Supervised Training

Latent variable models can also provide an easy method for semi-supervised training. We use a natural language inference task to compare semi-supervised ARAE with other training methods. Results are shown in Table 5. The full SNLI training set contains 543k sentence pairs, and we use supervised sets of 120k (Medium), 59k (Small), and 28k (Tiny) and use the rest of the training set for unlabeled training. As a baseline we use an AE trained on the additional data, similar to the setting explored in (Dai & Le, 2015).

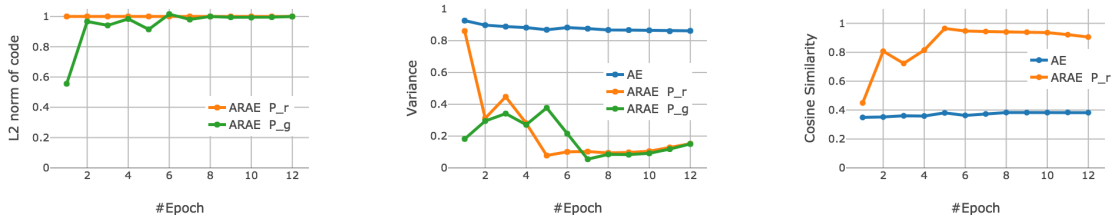

Figure 3: Left: $\ell 2$ norm of encoder output $\mathbf{z}$ and generator output $\tilde{\mathbf{z}}$ during ARAE training. ($\mathbf{z}$ is normalized, whereas the generator learns to match). Middle: Sum of the dimension-wise variances of $\mathbf{z}$ and generator codes $\tilde{\mathbf{z}}$ as well as reference AE. Right: Average cosine similarity of nearby sentences (by word edit-distance) for the ARAE and AE during training.

k	AE	ARAE
0	1.06	2.19
1	4.51	4.07
2	6.61	5.39
3	9.14	6.86
4	9.97	7.47

Model	Samples
Original	A woman wearing sunglasses
Noised	A woman sunglasses wearing
AE	A woman sunglasses wearing sunglasses
ARAE	A woman wearing sunglasses
Original	Pets galloping down the street
Noised	Pets down the galloping street
AE	Pets riding the down galloping
ARAE	Pets congregate down the street near a ravine

Figure 4: Reconstruction error (negative log-likelihood averaged over sentences) of the original sentence from a corrupted sentence. Here k is the number of swaps performed on the original sentence.

For ARAE we use the subset of unsupervised data of length < 15 (i.e. ARAE is trained on less data than AE for unsupervised training). The results are shown in Table 5. As observed by (Dai & Le, 2015), training on unlabeled data with an AE objective improves upon a model just trained on labeled data. Training with adversarial regularization provides further gains.

7. Discussion

Impact of Regularization on Discrete Encoding We further examine the impact of adversarial regularization on the encoded representation produced by the model as it is trained. Figure 3 (left), shows a sanity check that the $\ell 2$ norm of encoder output $\mathbf{z}$ and prior samples $\tilde{\mathbf{z}}$ converge quickly in ARAE training. The middle plot compares the trace of the covariance matrix between these terms as training progresses. It shows that variance of the encoder and the prior match after several epochs.

Smoothness and Reconstruction We can also assess the "smoothness" of the encoder model learned ARAE (Rifai et al., 2011). We start with a simple proxy that a smooth encoder model should map similar sentences to similar $\mathbf{z}$ values. For 250 sentences, we calculate the average cosine similarity of 100 randomly-selected sentences within an edit-distance of at most 5 to the original. The graph in Figure 3 (right) shows that the cosine similarity of nearby sentences is quite high for ARAE compared to a standard AE and increases in early rounds of training.

To further test this property, we feed noised input to the encoder and (i) calculate the score given to the original input, and (ii) compare the resulting reconstructions. Figure 4 (right) shows results for text where k words are first permuted in each sentence. We observe that the ARAE is able to map a noised sentence to a natural sentence, (though not necessarily the denoised sentence). Figure 4 (left) shows empirical results for these experiments. We obtain the reconstruction error (negative log likelihood) of the original non-noised sentence under the decoder, utilizing the noised code. We find that when $k = 0$ (i.e. no swaps), the regular AE better reconstructs the exact input. However, as the number of swaps pushes the input further away, ARAE is more likely to produce the original sentence. (Note that unlike denoising autoencoders which require a domain-specific noising function (Hill et al., 2016; Vincent et al., 2008), the ARAE is not explicitly trained to denoise an input.)

Manipulation through the Prior An interesting property of latent variable models such as VAEs and GANs is the ability to manipulate output samples through the prior. In particular, for ARAE, the Gaussian form of the noise sample $\mathbf{s}$ induces the ability to smoothly interpolate between outputs by exploiting the structure. While language models may provide a better estimate of the underlying probability space, constructing this style of interpolation would require combinatorial search, which makes this a useful feature of latent variable text models. In Appendix D we show interpolations from for the text model, while Figure 2 (bottom) shows the interpolations for discretized MNIST ARAE.

A related property of GANs is the ability to move in the latent space via offset vectors.[6] To experiment with this property we generate sentences from the ARAE and compute vector transforms in this space to attempt to change main verbs, subjects and modifier (details in Appendix E). Some examples of successful transformations are shown in Figure 5 (bottom). Quantitative evaluation of the success of the vector transformations is given in Figure 5 (top).

[6]Similar to the case with word vectors (Mikolov et al., 2013), Radford et al. (2016) observe that when the mean latent vector for "men with glasses" is subtracted from the mean latent vector for "men without glasses" and applied to an image of a "woman without glasses", the resulting image is that of a "woman with glasses".

Transform	Match %	Prec
walking	85	79.5
man	92	80.2
two	86	74.1
dog	88	77.0
standing	89	79.3
several	70	67.0

A man in a tie is sleeping and clapping on balloons . A man in a tie is clapping and walking dogs .	$\Rightarrow_{\text{walking}}$
The jewish boy is trying to stay out of his skateboard . The jewish man is trying to stay out of his horse .	$\Rightarrow_{\text{man}}$
Some child head a playing plastic with drink . Two children playing a head with plastic drink .	$\Rightarrow_{\text{Two}}$
The people shine or looks into an area . The dog arrives or looks into an area .	$\Rightarrow_{\text{dog}}$
A women are walking outside near a man . Three women are standing near a man walking .	$\Rightarrow_{\text{standing}}$
A side child listening to a piece with steps playing on a table . Several child playing a guitar on side with a table .	$\Rightarrow_{\text{Several}}$

Figure 5: Top: Quantitative evaluation of transformations. Match % refers to the % of samples where at least one decoder samples (per 100) had the desired transformation in the output, while Prec. measures the average precision of the output against the original sentence. Bottom: Examples where the offset vectors produced successful transformations of the original sentence. See Appendix E for the full methodology.

8. Related Work

Ideally, autoencoders would learn latent spaces which compactly capture useful features that explain the observed data. However in practice unregularized autoencoders often learn a degenerate *identity* mapping where the latent code space is free of any structure, necessitating the need for some regularization on the latent space. A popular approach is to regularize through an explicit prior on the code space and use a variational approximation to the posterior, leading to a family of models called variational autoencoders (VAE) (Kingma & Welling, 2014; Rezende et al., 2014). Unfortunately VAEs for discrete text sequences can be challenging to train—for example, if the training procedure is not carefully tuned with techniques like word dropout and KL annealing (Bowman et al., 2015b), the decoder simply becomes a language model and ignores the latent code. One possible reason for the difficulty in training VAEs is due to the strictness of the prior (usually a spherical Gaussian) and/or the parameterization of the posterior. There has been some work on making the prior/posterior more flexible through explicit parameterization (Rezende & Mohamed, 2015; Kingma et al., 2016; Chen et al., 2017). A notable technique is adversarial autoencoders (AAE) (Makhzani et al., 2015) which attempt to imbue the model with a more flexible prior implicitly through adversarial training. Recent work on Wasserstein Autoencoders (Tolstikhin et al., 2018) provides a theoretical foundation for the AAE and shows that AAE minimizes the Wasserstein distance between the data/model distributions. Along this line of work, we explore the use of a parametric prior (i.e. we do not sample from a fixed prior distribution) where the 'prior' is instead parameterized through a flexible generator.

The success of GANs on images have led many researchers to consider applying GANs to discrete data such as text. Policy gradient methods are a natural way to deal with the resulting non-differentiable generator objective when training directly in discrete space (Glynn, 1987; Williams, 1992). When trained on text data however, such methods often require pre-training/co-training with a maximum likelihood (i.e. language modeling) objective (Che et al., 2017; Yu et al., 2017; Li et al., 2017). Another direction of work has been through reparameterizing the categorical distribution with the Gumbel-Softmax trick (Jang et al., 2017; Maddison et al., 2017)—while initial experiments were encouraging on a synthetic task (Kusner & Hernandez-Lobato, 2016), scaling them to work on natural language is a challenging open problem. There has also been many recent related approaches that work directly with the soft outputs from a generator (Gulrajani et al., 2017; Sai Rajeswar, 2017; Shen et al., 2017; Press et al., 2017). For example, Shen et al. (2017) exploits adversarial loss for unaligned style transfer between text by having the discriminator act on the RNN hidden states and using the soft outputs at each step as input to an RNN generator, utilizing the Professor-forcing framework (Lamb et al., 2016). Our approach instead works entirely in fixed-dimensional continuous space and does not require utilizing RNN hidden states directly. It is therefore also different from methods that discriminate in the joint latent/data space, such as adversarially learned inference (Vincent Dumoulin, 2017) and BiGAN (Donahue et al., 2017). Finally, our work also adds to the recent line of work on connecting VAEs to GANs (Tran et al., 2017; Mescheder et al., 2017; Makhzani & Frey, 2017; Zhao et al., 2017; Hu et al., 2018).

9. Conclusion

We present adversarially regularized autoencoders (ARAE) as a simple approach for training a discrete structure autoencoder jointly with a code-space generative adversarial network. Utilizing the Wasserstein Autoencoder framework (Tolstikhin et al., 2018), we also interpret ARAE as learning a latent variable model that minimizes an upper bound on the total variation distance between the data/model distributions. We find that the model learns an improved autoencoder and exhibits a smooth latent space, as demonstrated by semi-supervised experiments, improvements on text style transfer, and manipulations in the latent space. Finally, we note that (as has been frequently observed when training GANs) the proposed model seemed to be quite sensitive to hyperparameters, and that we only tested our model on simple structures such as binarized digits and short sentences. Training deep latent variable models that can robustly model complex discrete structures (e.g. paragraphs/documents) remains an important open issue in the field.

References

Arjovsky, Martin, Chintala, Soumith, and Bottou, Léon. Wasserstein gan. *arXiv:1701.07875*, 2017.

Bowman, Samuel R., Angeli, Gabor, Potts, Christopher, and Manning., Christopher D. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*, 2015a.

Bowman, Samuel R, Vilnis, Luke, Vinyals, Oriol, Dai, Andrew M, Jozefowicz, Rafal, and Bengio, Samy. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015b.

Che, Tong, Li, Yanran, Zhang, Ruixiang, Hjelm, R Devon, Li, Wenjie, Song, Yangqui, and Bengio, Yoshua. Maximum-Likelihood Augment Discrete Generative Adversarial Networks. *arXiv:1702.07983*, 2017.

Chen, Xi, Kingma, Diederik P., Salimans, Tim, Duan, Yan, Dhariwal, Prafulla, Schulman, John, Sutskever, Ilya, and Abbeel, Pieter. Variational Lossy Autoencoder. In *Proceedings of ICLR*, 2017.

Dai, Andrew M and Le, Quoc V. Semi-supervised sequence learning. In *Proceedings of NIPS*, 2015.

Denton, Emily and Birodkar, Vighnesh. Unsupervised learning of disentangled representations from video. *arXiv preprint arXiv:1705.10915*, 2017.

Donahue, Jeff, Krahenbühl, Philipp, and Darrell, Trevor. Adversarial Feature Learning. In *Proceedings of ICLR*, 2017.

Glynn, Peter. Likelihood Ratio Gradient Estimation: An Overview. In *Proceedings of Winter Simulation Conference*, 1987.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial nets. In *Proceedings of NIPS*, 2014.

Gulrajani, Ishaan, Ahmed, Faruk, Arjovsky, Martin, and Vincent Dumoulin, Aaron Courville. Improved Training of Wasserstein GANs. *arXiv:1704.00028*, 2017.

Hill, Felix, Cho, Kyunghyun, and Korhonen, Anna. Learning distributed representations of sentences from unlabelled data. In *Proceedings of NAACL*, 2016.

Hjelm, R Devon, Jacob, Athul Paul, Che, Tong, Cho, Kyunghyun, and Bengio, Yoshua. Boundary-Seeking Generative Adversarial Networks. *arXiv:1702.08431*, 2017.

Hochreiter, Sepp and Schmidhuber, Jürgen. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Hu, Zhiting, Yang, Zichao, Salakhutdinov, Ruslan, and Xing, Eric P. On Unifying Deep Generative Models. In *Proceedings of ICLR*, 2018.

Jang, Eric, Gu, Shixiang, and Poole, Ben. Categorical Reparameterization with Gumbel-Softmax. In *Proceedings of ICLR*, 2017.

Joulin, Armand, Grave, Edouard, Bojanowski, Piotr, and Mikolov, Tomas. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

Kingma, Diederik P. and Welling, Max. Auto-Encoding Variational Bayes. In *Proceedings of ICLR*, 2014.

Kingma, Diederik P., Salimans, Tim, and Welling, Max. Improving Variational Inference with Autoregressive Flow. *arXiv:1606.04934*, 2016.

Kusner, Matt and Hernandez-Lobato, Jose Miguel. GANs for Sequences of Discrete Elements with the Gumbel-Softmax Distribution. *arXiv:1611.04051*, 2016.

Lamb, Alex M, Goyal, Anirudh, Zhang, Ying, Zhang, Saizheng, Courville, Aaron C, and Bengio, Yoshua. Professor forcing: A new algorithm for training recurrent networks. In *Advances In Neural Information Processing Systems*, pp. 4601–4609, 2016.

Lample, Guillaume, Zeghidour, Neil, Usuniera, Nicolas, Bordes, Antoine, Denoyer, Ludovic, and Ranzato, Marc'Aurelio. Fader networks: Manipulating images by sliding attributes. In *Proceedings of NIPS*, 2017.

Li, Jiwei, Monroe, Will, Shi, Tianlin, Jean, Sébastien, Ritter, Alan, and Jurafsky, Dan. Adversarial Learning for Neural Dialogue Generation. *arXiv:1701.06547*, 2017.

Maddison, Chris J., Mnih, Andriy, and Teh, Yee Whye. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In *Proceedings of ICLR*, 2017.

Makhzani, Alireza and Frey, Brendan. PixelGAN Autoencoders. *arXiv:1706.00531*, 2017.

Makhzani, Alireza, Shlens, Jonathon, Jaitly, Navdeep, Goodfellow, Ian, and Frey, Brendan. Adversarial Autoencoders. *arXiv:1511.05644*, 2015.

Mescheder, Lars, Nowozin, Sebastian, and Geiger, Andreas. Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks. *arXiv:1701.04722*, 2017.

Mikolov, Tomas, tau Yih, Scott Wen, and Zweig, Geoffrey. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of NAACL*, 2013.

Press, Ofir, Bar, Amir, Bogin, Ben, Berant, Jonathan, and Wolf, Lior. Language Generation with Recurrent Generative Adversarial Networks without Pre-training. *arXiv:1706.01399*, 2017.

Radford, Alec, Metz, Luke, and Chintala, Soumith. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *Proceedings of ICLR*, 2016.

Rezende, Danilo J. and Mohamed, Shakir. Variational Inference with Normalizing Flows. In *Proceedings of ICML*, 2015.

Rezende, Danilo Jimenez, Mohamed, Shakir, and Wierstra, Daan. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proceedings of ICML*, 2014.

Rifai, Salah, Vincent, Pascal, Muller, Xavier, Glorot, Xavier, and Bengio, Yoshua. Contractive Auto-Encoders: Explicit Invariance During Feature Extraction. In *Proceedings of ICML*, 2011.

Sai Rajeswar, Sandeep Subramanian, Francis Dutil Christopher Pal Aaron Courville. Adversarial Generation of Natural Language. *arXiv:1705.10929*, 2017.

Shen, Tianxiao, Lei, Tao, Barzilay, Regina, and Jaakkola, Tommi. Style Transfer from Non-Parallel Text by Cross-Alignment. *arXiv:1705.09655*, 2017.

Theis, Lucas, van den Oord, Aaron, and Bethge, Matthias. A note on the evaluation of generative models. In *Proceedings of ICLR*, 2016.

Tolstikhin, Ilya, Bousquet, Olivier, Gelly, Sylvain, and Schoelkopf, Bernhard. Wasserstein Auto-Encoders. 2018.

Tran, Dustin, Ranganath, Rajesh, and Blei, David M. Deep and Hierarchical Implicit Models. *arXiv:1702.08896*, 2017.

Villani, Cédric. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Vincent, Pascal, Larochelle, Hugo, Bengio, Yoshua, and Manzagol, Pierre-Antoine. Extracting and Composing Robust Features with Denoising Autoencoders. In *Proceedings of ICML*, 2008.

Vincent Dumoulin, Ishmael Belghazi, Ben Poole Olivier Mastropietro Alex Lamb Martin Arjovsky Aaron Courville. Adversarially Learned Inference. In *Proceedings of ICLR*, 2017.

Williams, Ronald J. Simple Statistical Gradient-following Algorithms for Connectionist Reinforcement Learning. *Machine Learning*, 8, 1992.

Yu, Lantao, Zhang, Weinan, Wang, Jun, and Yu, Yong. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. In *Proceedings of AAAI*, 2017.

Zhang, Xiang, Zhao, Junbo, and LeCun, Yann. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pp. 649–657, 2015.

Zhao, Shengjia, Song, Jiaming, and Ermon, Stefano. The Information-Autoencoding Family: A Lagrangian Perspective on Latent Variable Generative Modeling. *OpenReview:* https://openreview.net/pdf?id=ryZERzWCZ, 2017.

A. Proof of Corollary 1

Corollary (Discrete case)**.** *Suppose $\mathbf{x} \in \mathcal{X}$ where $\mathcal{X}$ is the set of all one-hot vectors of length n, and let $f_\psi : \mathcal{Z} \to \Delta^{n-1}$ be a deterministic function that goes from the latent space $\mathcal{Z}$ to the $n-1$ dimensional simplex Δ^{n-1}. Further let $G_\psi : \mathcal{Z} \to \mathcal{X}$ be a deterministic function such that $G_\psi(\mathbf{z}) = \arg\max_{\mathbf{w} \in \mathcal{X}} \mathbf{w}^\top f_\psi(\mathbf{z})$, and as above let $\mathbb{P}_\psi(\mathbf{x}|\mathbf{z})$ be the dirac distribution derived from G_ψ such that $p_\psi(\mathbf{x}|\mathbf{z}) = \mathbb{1}\{\mathbf{x} = G_\psi(\mathbf{z})\}$. Then the following is an upper bound on $\|\mathbb{P}_\psi - \mathbb{P}_\star\|_{TV}$, the total variation distance between $\mathbb{P}_\star$ and $\mathbb{P}_\psi$.*

$$\inf_{Q(\mathbf{z}|\mathbf{x}):\mathbb{P}_Q = \mathbb{P}_\mathbf{z}} \mathbb{E}_{\mathbb{P}_\star} \mathbb{E}_{Q(\mathbf{z}|\mathbf{x})} \Big[-\frac{2}{\log 2} \log \mathbf{x}^\top f_\psi(\mathbf{z}) \Big]$$

Proof. Let our cost function be $c(\mathbf{x}, \mathbf{y}) = \mathbb{1}\{\mathbf{x} \neq \mathbf{y}\}$. We first note that for all $\mathbf{x}, \mathbf{z}$

$$\log 2\mathbb{1}\{\mathbf{x} \neq \arg\max_{\mathbf{w} \in \mathcal{X}} \mathbf{w}^\top f_\psi(\mathbf{z})\} < -\log \mathbf{x}^\top f_\psi(\mathbf{z})$$

This holds since if $\mathbb{1}\{\mathbf{x} \neq \arg\max_{\mathbf{w} \in \mathcal{X}} \mathbf{w}^\top f_\psi(\mathbf{z})\} = 1$, we have $\mathbf{x}^\top f_\psi(\mathbf{z}) < 0.5$, and $-\log \mathbf{x}^\top f_\psi(\mathbf{z}) > -\log 0.5 = \log 2$. Then,

$$\begin{aligned}
&\inf_{Q:\mathbb{P}_Q = \mathbb{P}_\mathbf{z}} \mathbb{E}_{\mathbb{P}_\star} \mathbb{E}_{Q(\mathbf{z}|\mathbf{x})} [-\frac{2}{\log 2} \log \mathbf{x}^\top f_\psi(\mathbf{z})] \\
>& \inf_{Q:\mathbb{P}_Q = \mathbb{P}_\mathbf{z}} \mathbb{E}_{\mathbb{P}_\star} \mathbb{E}_{Q(\mathbf{z}|\mathbf{x})} [2\mathbb{1}\{\mathbf{x} \neq \arg\max_{\mathbf{w} \in \mathcal{X}} \mathbf{w}^\top f_\psi(\mathbf{z})\}] \\
=& 2 \inf_{Q:\mathbb{P}_Q = \mathbb{P}_\mathbf{z}} \mathbb{E}_{\mathbb{P}_\star} \mathbb{E}_{Q(\mathbf{z}|\mathbf{x})} [\mathbb{1}\{\mathbf{x} \neq G_\psi(\mathbf{z})\}] \\
=& 2 \inf_{Q:\mathbb{P}_Q = \mathbb{P}_\mathbf{z}} \mathbb{E}_{\mathbb{P}_\star} \mathbb{E}_{Q(\mathbf{z}|\mathbf{x})} [c(\mathbf{x}, G_\psi(\mathbf{z}))] \\
=& 2W_c(\mathbb{P}_\star, \mathbb{P}_\psi) \\
=& \|\mathbb{P}_\star - \mathbb{P}_\psi\|_{\text{TV}}
\end{aligned}$$

The fifth line follows from Theorem 1, and the last equality uses the well-known correspondence between total variation distance and optimal transport with the indicator cost function. □

B. Optimality Property

One can interpret the ARAE framework as a dual pathway network mapping two distinct distributions into a similar one; enc_ϕ and g_θ both output code vectors that are kept similar in terms of Wasserstein distance as measured by the critic. We provide the following proposition showing that under our parameterization of the encoder and the generator, as the Wasserstein distance converges, the encoder distribution ($\mathbb{P}_Q$) converges to the generator distribution ($\mathbb{P}_z$), and further, their moments converge.

This is ideal since under our setting the generated distribution is simpler than the encoded distribution, because the input to the generator is from a simple distribution (e.g. spherical Gaussian) and the generator possesses less capacity than the encoder. However, it is not so simple that it is overly restrictive (e.g. as in VAEs). Empirically we observe that the first and second moments do indeed converge as training progresses (Section 7).

Proposition 1. *Let $\mathbb{P}$ be a distribution on a compact set $\mathcal{X}$, and $(\mathbb{P}_n)_{n \in N}$ be a sequence of distributions on $\mathcal{X}$. Further suppose that $W(\mathbb{P}_n, \mathbb{P}) \to 0$. Then the following statements hold:*

(i) $\mathbb{P}_n \rightsquigarrow \mathbb{P}$ *(i.e. convergence in distribution).*

(ii) All moments converge, i.e. for all $k > 1, k \in \mathbb{N}$,

$$\mathbb{E}_{X \sim \mathbb{P}_n} \Big[\prod_{i=1}^d X_i^{p_i} \Big] \to \mathbb{E}_{X \sim \mathbb{P}} \Big[\prod_{i=1}^d X_i^{p_i} \Big]$$

for all $p_1, \ldots, p_d$ such that $\sum_{i=1}^d p_i = k$

Proof. (i) has been proved in (Villani, 2008) Theorem 6.9.

For (ii), using *The Portmanteau Theorem*, (i) is equivalent to the following statement:

$\mathbb{E}_{X \sim \mathbb{P}_n}[f(X)] \to \mathbb{E}_{X \sim \mathbb{P}}[f(X)]$ for all bounded and continuous function f: $\mathbb{R}^d \to \mathbb{R}$, where d is the dimension of the random variable.

The k-th moment of a distribution is given by

$$\mathbb{E}\Big[\prod_{i=1}^d X_i^{p_i} \Big] \text{ such that } \sum_{i=1}^d p_i = k$$

Our encoded code is bounded as we normalize the encoder output to lie on the unit sphere, and our generated code is also bounded to lie in $(-1, 1)^n$ by the $\tanh$ function. Hence $f(X) = \prod_{i=1}^d X_i^{q_i}$ is a bounded continuous function for all $q_i \geq 0$. Therefore,

$$\mathbb{E}_{X \sim \mathbb{P}_n} \Big[\prod_{i=1}^d X_i^{p_i} \Big] \to \mathbb{E}_{X \sim \mathbb{P}} \Big[\prod_{i=1}^d X_i^{p_i} \Big]$$

where $\sum_{i=1}^d p_i = k$ □

C. Sample Generations

In Figure 6 we show some generated samples from the ARAE, AE, and a LM.

D. Sentence Interpolations

In Figure 7 we show generations from interpolated latent vectors. Specifically, we sample two points $\mathbf{z}_0$ and $\mathbf{z}_1$ from $p(\mathbf{z})$ and construct intermediary points $\mathbf{z}_\lambda = \lambda \mathbf{z}_1 + (1 - \lambda)\mathbf{z}_0$. For each we generate the argmax output $\tilde{\mathbf{x}}_\lambda$.

ARAE Samples	AE Samples	LM Samples
A woman preparing three fish .	Two Three woman in a cart tearing over of a tree .	a man walking outside on a dirt road , sitting on the dock .
A woman is seeing a man in the river .	A man is hugging and art .	A large group of people is taking a photo for Christmas and at night .
There passes a woman near birds in the air .	The fancy skier is starting under the drag cup in .	Someone is avoiding a soccer game .
Some ten people is sitting through their office .	A dog are <unk> a	The man and woman are dressed for a movie .
The man got stolen with young dinner bag .	A man is not standing .	Person in an empty stadium pointing at a mountain .
Monks are running in court .	The Boys in their swimming .	Two children and a little boy are <unk> a man in a blue shirt .
The Two boys in glasses are all girl .	A surfer and a couple waiting for a show .	A boy rides a bicycle .
The man is small sitting in two men that tell a children .	A couple is a kids at a barbecue .	A girl is running another in the forest .
The two children are eating the balloon animal .	The motorcycles is in the ocean loading	the man is an indian women .
A woman is trying on a microscope .	I 's bike is on empty	
The dogs are sleeping in bed .	The actor was walking in a a small dog area .	
	no dog is young their mother	

Figure 6: Text samples generated from ARAE, a simple AE, and from a baseline LM trained on the same data. To generate from an AE we fit a multivariate Gaussian to the learned code space and generate code vectors from this Gaussian.

A man is on the corner in a sport area .	A man is on a ship path with the woman .	A man in a cave is used an escalator .
A man is on corner in a road all .	A man is on a ship path with the woman .	
A lady is on outside a race-track .	A man is passing on a bridge with the girl .	A man in a cave is used an escalator
A lady is outside on a race-track .	A man is passing on a bridge with the girl .	A man in a cave is used chairs .
A lot of people is outdoors in an urban setting .	A man is passing on a bridge with the girl .	A man in a number is used many equipment
A lot of people is outdoors in an urban setting .	A man is passing on a bridge with the dogs .	A man in a number is posing so on a big rock .
A lot of people is outdoors in an urban setting .	A man is passing on a bridge with the dogs .	People are posing in a rural area .
		People are posing in a rural area.

Figure 7: Sample interpolations from the ARAE. Constructed by linearly interpolating in the latent space and decoding to the output space. Word changes are highlighted in black.

E. Vector Arithmetic

We generate 1 million sentences from the ARAE and parse the sentences to obtain the main verb, subject, and modifier. Then for a given sentence, to change the main verb we subtract the mean latent vector ($\mathbf{t}$) for all other sentences with the same main verb (in the first example in Figure 5 this would correspond to all sentences that had "sleeping" as the main verb) and add the mean latent vector for all sentences that have the desired transformation (with the running example this would be all sentences whose main verb was "walking"). We do the same to transform the subject and the modifier. We decode back into sentence space with the transformed latent vector via sampling from $p_\psi(g(\mathbf{z}+\mathbf{t}))$. Some examples of successful transformations are shown in Figure 5 (right). Quantitative evaluation of the success of the vector transformations is given in Figure 5 (left). For each original vector $\mathbf{z}$ we sample 100 sentences from $p_\psi(g(\mathbf{z}+\mathbf{t}))$ over the transformed new latent vector and consider it a match if *any* of the sentences demonstrate the desired transformation. Match % is proportion of original vectors that yield a match post transformation. As we ideally want the generated samples to only differ in the specified transformation, we also calculate the average word precision against the original sentence (Prec) for any match.

F. Experimental Details

MNIST experiments

- The encoder is a three-layer MLP, `784-800-400-100`.
- Additive Gaussian noise is added into $\mathbf{c}$ which is then fed into the decoder. The standard deviation of that noise is initialized to be 0.4, and then exponentially decayed to 0.
- The decoder is a four-layer MLP, `100-400-800-1000-784`
- The autoencoder is optimized by Adam, with learning rate `5e-04`.
- An MLP generator `32-64-100-150-100`, using batch normalization, and ReLU non-linearity.
- An MLP critic `100-100-60-20-1` with weight clipping $\epsilon = 0.05$. The critic is trained by 10 iterations within each GAN loop.
- Both components of GAN is optimized by Adam, with learning rate `5e-04` on the generator, and `5e-05` on the critic.
- Weighing factor $\lambda^{(1)} = 0.2$.

Text experiments

- The encoder is an one-layer LSTM with 300 hidden units.
- Gaussian noise into $\mathbf{c}$ before feeding it into the decoder. The standard deviation of that noise is initialized to be 0.2, and then exponentially decayed every 100 iterations by a factor of 0.995.
- The decoder is a one-layer LSTM with 300 hidden units.
- The decoding process at each time step takes the top layer LSTM hidden state and concatenates it with the hidden codes $\mathbf{c}$, before feeding them into the output (i.e. vocabulary projection) and the softmax layer.
- The word embedding is of size 300.
- We adopt a grad clipping on the encoder/decoder, with max `grad_norm` = 1.
- The encoder/decoder is optimized by vanilla SGD with learning rate 1.

- An MLP generator `100-300-300`, using batch normalization, and ReLU non-linearity.
- An MLP critic `300-300-1` with weight clipping $\epsilon = 0.01$. The critic is trained by 5 iterations within each GAN loop.
- Both components of GAN are optimized by Adam, with learning rate `5e-05` on the generator, and `1e-05` on the critic.
- We increment the number of GAN training loop[7] by 1 (it initially is set to 1) , respectively at the beginning of epoch #2, epoch #4 and epoch #6.

Semi-supervised experiments

Similar to the SNLI generation experiment setup, with the following changes:

- We employ larger network to GAN components: MLP generator `100-150-300-500` and MLP critic `500-500-150-80-20-1` with weight clipping factor $\epsilon = 0.02$. The critic is trained by 10 iterations within each GAN loop.

Yelp/Yahoo transfer

Similar to the SNLI setup, with the following changes

- The encoder and decoder size are both increased to 500 hidden units.
- The style adversarial classifier is an MLP with structure `300-200-100`, with learning rate 0.1 trained with SGD.
- We employ both larger generator and discriminator architectures in GAN: generator `200-400-800` with z dim being set to 64; discriminator `300-160-80-20`.
- Weighing factor for critic gradient $\lambda_a^{(1)} = 1$, $\lambda_b^{(1)} = 10$.
- No GAN loop scheduling is employed here.

G. Style Transfer Samples

In the following pages we show randomly sampled style transfers from the Yelp/Yahoo corpus.

[7]The GAN training loop refers to how many times we train GAN in each entire training loop (one training loop contains training autoencoder for one loop, and training GAN for one or several).

Yelp Sentiment Transfer

	Positive to Negative		Negative to Positive
Original	great indoor mall .	Original	hell no !
ARAE	no smoking mall .	ARAE	hell great !
Cross-AE	terrible outdoor urine .	Cross-AE	incredible pork !
Original	great blooming onion .	Original	highly disappointed !
ARAE	no receipt onion .	ARAE	highly recommended !
Cross-AE	terrible of pie .	Cross-AE	highly clean !
Original	i really enjoyed getting my nails done by peter .	Original	bad products .
ARAE	i really needed getting my nails done by now .	ARAE	good products .
Cross-AE	i really really told my nails done with these things .	Cross-AE	good prices .
Original	definitely a great choice for sushi in las vegas !	Original	i was so very disappointed today at lunch .
ARAE	definitely a _num_ star rating for _num_ sushi in las vegas .	ARAE	i highly recommend this place today .
Cross-AE	not a great choice for breakfast in las vegas vegas !	Cross-AE	i was so very pleased to this .
Original	the best piece of meat i have ever had !	Original	i have n't received any response to anything .
ARAE	the worst piece of meat i have ever been to !	ARAE	i have n't received any problems to please .
Cross-AE	the worst part of that i have ever had had !	Cross-AE	i have always the desert vet .
Original	really good food , super casual and really friendly .	Original	all the fixes were minor and the bill ?
ARAE	really bad food , really generally really low and decent food .	ARAE	all the barbers were entertaining and the bill did n't disappoint .
Cross-AE	really good food , super horrible and not the price .	Cross-AE	all the flavors were especially and one !
Original	it has a great atmosphere , with wonderful service .	Original	small , smokey , dark and rude management .
ARAE	it has no taste , with a complete jerk .	ARAE	small , intimate , and cozy friendly staff .
Cross-AE	it has a great horrible food and run out service .	Cross-AE	great , , , chips and wine .
Original	their menu is extensive , even have italian food .	Original	the restaurant did n't meet our standard though .
ARAE	their menu is limited , even if i have an option .	ARAE	the restaurant did n't disappoint our expectations though .
Cross-AE	their menu is decent , i have gotten italian food .	Cross-AE	the restaurant is always happy and knowledge .
Original	everyone who works there is incredibly friendly as well .	Original	you could not see the stage at all !
ARAE	everyone who works there is incredibly rude as well .	ARAE	you could see the difference at the counter !
Cross-AE	everyone who works there is extremely clean and as well .	Cross-AE	you could definitely get the fuss !
Original	there are a couple decent places to drink and eat in here as well .	Original	room is void of all personality , no pictures or any sort of decorations .
ARAE	there are a couple slices of options and _num_ wings in the place .	ARAE	room is eclectic , lots of flavor and all of the best .
Cross-AE	there are a few night places to eat the car here are a crowd .	Cross-AE	it 's a nice that amazing , that one 's some of flavor .
Original	if you 're in the mood to be adventurous , this is your place !	Original	waited in line to see how long a wait would be for three people .
ARAE	if you 're in the mood to be disappointed , this is not the place .	ARAE	waited in line for a long wait and totally worth it .
Cross-AE	if you 're in the drive to the work , this is my place !	Cross-AE	another great job to see and a lot going to be from dinner .
Original	we came on the recommendation of a bell boy and the food was amazing .	Original	the people who ordered off the menu did n't seem to do much better .
Cross-AE	we came on the recommendation and the food was a joke .	ARAE	the people who work there are super friendly and the menu is good .
Cross-AE	we went on the car of the time and the chicken was awful .	Cross-AE	the place , one of the office is always worth you do a business .
Original	service is good but not quick , just enjoy the wine and your company .	Original	they told us in the beginning to make sure they do n't eat anything .
ARAE	service is good but not quick , but the service is horrible .	ARAE	they told us in the mood to make sure they do great food .
Cross-AE	service is good , and horrible , is the same and worst time ever .	Cross-AE	they 're us in the next for us as you do n't eat .
Original	the steak was really juicy with my side of salsa to balance the flavor .	Original	the person who was teaching me how to control my horse was pretty rude .
ARAE	the steak was really bland with the sauce and mashed potatoes .	ARAE	the person who was able to give me a pretty good price .
Cross-AE	the fish was so much , the most of sauce had got the flavor .	Cross-AE	the owner 's was gorgeous when i had a table and was friendly .
Original	other than that one hell hole of a star bucks they 're all great !	Original	he was cleaning the table next to us with gloves on and a rag .
ARAE	other than that one star rating the toilet they 're not allowed .	ARAE	he was prompt and patient with us and the staff is awesome .
Cross-AE	a wonder our one came in a _num_ months , you 're so better !	Cross-AE	he was like the only thing to get some with with my hair .

Figure 8: Full sheet of sentiment transfer result on the Yelp corpus.

Yahoo Topic Transfer on Questions

	from Science		from Music		from Politics
Original	what is an event horizon with regards to black holes ?	Original	do you know a website that you can find people who want to join bands ?	Original	republicans : would you vote for a cheney / satan ticket in 2008 ?
Music	what is your favorite sitcom with adam sandler ?	Science	do you know a website that can help me with science ?	Science	guys : how would you solve this question ?
Politics	what is an event with black people ?	Politics	do you think that you can find a person who is in prison ?	Music	guys : would you rather be a good movie ?
Original	what did john paul jones do in the american revolution ?	Original	do people who quote entire poems or song lyrics ever actually get chosen best answer ?	Original	if i move to the usa do i lose my pension in canada ?
Music	what did john lennon do in the new york family ?	Science	do you think that scientists learn about human anatomy and physiology of life ?	Science	if i move the <unk> in the air i have to do my math homework ?
Politics	what did john mccain do in the next election ?	Politics	do people who knows anything about the recent issue of <unk> leadership ?	Music	if i move to the music do you think i feel better ?
Original	can anybody suggest a good topic for a statistical survey ?	Original	from big brother , what is the girls name who had <unk> in her apt ?	Original	what is your reflection on what will be our organizations in the future ?
Music	can anybody suggest a good site for a techno ?	Science	in big bang what is the <unk> of <unk> , what is the difference between <unk> and <unk> ?	Science	what is your opinion on what will be the future in our future ?
Politics	can anybody suggest a good topic for a student visa ?	Politics	is big brother in the <unk> what do you think of her ?	Music	what is your favorite music videos on the may i find ?
Original	can a kidney infection effect a woman 's <unk> cycle ?	Original	where is the tickets for the filming of the suite life of zack and cody ?	Original	wouldn 't it be fun if we the people veto or passed bills ?
Music	can anyone give me a good film <unk> ?	Science	where is the best place of the blood stream for the production of the cell ?	Science	isnt it possible to be cloned if we put the moon or it ?
Politics	can a landlord officer have a <unk> <unk> ?	Politics	where is the best place of the navy and the senate of the union ?	Music	isnt it possible or if we 're getting married ?
Original	where does the term " sweating <unk> " come from ?	Original	the <unk> singers was a band in 1963 who had a hit called <unk> man ?	Original	can anyone tell me how i could go about interviewing north vietnamese soldiers ?
Music	where does the term " <unk> " come from ?	Science	the <unk> river in a <unk> was created by a <unk> who was born in the last century ?	Science	can anyone tell me how i could find how to build a robot ?
Politics	where does the term " <unk> " come from ?	Politics	the <unk> are <unk> in a <unk> who was shot an <unk> ?	Music	can anyone tell me how i could find out about my parents ?
Original	what other <unk> sources are there than burning fossil fuels .	Original	what is the first metal band in the early 60 's ..... ? ? ? ?	Original	if the us did not exist would the world be a better place ?
Music	what other <unk> are / who are the greatest guitarist currently on tv today ?	Science	what is the first country in the universe ?	Science	if the world did not exist , would it be possible ?
Politics	what other <unk> are there for veterans who lives ?	Politics	who is the first president in the usa ?	Music	if you could not have a thing who would it be ?

Figure 9: Full sheet of Yahoo topic transfer on titles.

Yahoo Topic Transfer on Answers

	from Science		from Music		from Politics
Original	take 1ml of hcl (concentrated) and dilute it to 50ml .	Original	all three are fabulous artists , with just incredible talent ! !	Original	4 years of an idiot in office + electing the idiot again = ?
Music	take em to you and shout it to me	Science	all three are genetically bonded with water , but just as many substances , are capable of producing a special case .	Science	4 years of an idiot in the office of science ?
Politics	take bribes to islam and it will be punished .	Politics	all three are competing with the government , just as far as i can .	Music	4) <unk> in an idiot , the idiot is the best of the two points ever !
Original	oils do not do this , they do not " set " .	Original	she , too , wondered about the underwear outside the clothes .	Original	send me $ 100 and i 'll send you a copy - honest .
Music	cucumbers do not do this , they do not " do " .	Science	she , too , i know , the clothes outside the clothes .	Science	send me an email and i 'll send you a copy .
Politics	corporations do not do this , but they do not .	Politics	she , too , i think that the cops are the only thing about the outside of the u.s. .	Music	send me $ 100 and i 'll send you a copy .
Original	the average high temps in jan and feb are about 48 deg .	Original	i like rammstein and i don 't speak or understand german .	Original	wills can be <unk> , or typed and signed without needing an attorney .
Music	the average high school in seattle and is about 15 minutes .	Science	i like googling and i don 't understand or speak .	Science	euler can be <unk> , and without any type of operations , or <unk> .
Politics	the average high infantry division is in afghanistan and alaska .	Politics	i like mccain and i don 't care about it .	Music	madonna can be <unk> , and signed without opening or <unk> .
Original	the light from you lamps would move away from you at light speed	Original	mark is great , but the guest hosts were cool too !	Original	hungary : 20 january 1945 , (formerly a member of the axis)
Music	the light from you tube would move away from you	Science	mark is great , but the water will be too busy for the same reason .	Science	nh3 : 20 january , 78 (a)
Politics	the light from you could go away from your state	Politics	mark twain , but the great lakes , the united states of america is too busy .	Music	1966 - 20 january 1961 (a) 1983 song
Original	van <unk> , on the other hand , had some serious issues ...	Original	they all offer terrific information about the cast and characters , ...	Original	bulgaria : 8 september 1944 , (formerly a member of the axis)
Music	van <unk> on the other hand , had some serious issues .	Science	they all offer insight about the characteristics of the earth , and are composed of many stars .	Science	moreover , 8 $\hat{3}$ + (x + 7) (x $\hat{2}$) = (a $\hat{2}$)
Politics	van <unk> , on the other hand , had some serious issues .	Politics	they all offer legitimate information about the invasion of iraq and the u.s. , and all aspects of history .	Music	harrison : 8 september 1961 (a) (1995)
Original	just multiply the numerator of one fraction by that of the other .	Original	but there are so many more i can 't think of !	Original	anyone who doesnt have a billion dollars for all the publicity cant win .
Music	just multiply the fraction of the other one that 's just like it .	Science	but there are so many more of the number of questions .	Science	anyone who doesnt have a decent chance is the same for all the other .
Politics	just multiply the same fraction of other countries .	Politics	but there are so many more of the can i think of today .	Music	anyone who doesnt have a lot of the show for the publicity .
Original	civil engineering is still an umbrella field comprised of many related specialties .	Original	i love zach he is sooo sweet in his own way !	Original	the theory is that cats don 't take to being tied up but thats <unk> .
Music	civil rights is still an art union .	Science	the answer is he 's definitely in his own way !	Science	the theory is that cats don 't grow up to <unk> .
Politics	civil law is still an issue .	Politics	i love letting he is sooo smart in his own way !	Music	the theory is that dumb but don 't play <unk> to <unk> .
Original	h2o2 (hydrogen peroxide) naturally decomposes to form o2 and water .	Original	remember the industry is very shady so keep your eyes open !	Original	the fear they are trying to instill in the common man is based on what ?
Music	jackie and brad pitt both great albums and they are my fav .	Science	remember the amount of water is so very important .	Science	the fear they are trying to find the common ancestor in the world .
Politics	kennedy and blair hate america to invade them .	Politics	remember the amount of time the politicians are open your mind .	Music	the fear they are trying to find out what is wrong in the song .
Original	the quieter it gets , the more white noise you can here .	Original	but can you fake it , for just one more show ?	Original	think about how much planning and people would have to be involved in what happened .
Music	the fray it gets , the more you can hear .	Science	but can you fake it , just for more than one ?	Science	think about how much time would you have to do .
Politics	the gop gets it , the more you can here .	Politics	but can you fake it for more than one ?	Music	think about how much money and what would be <unk> about in the world ?
Original	h2co3 (carbonic acid) naturally decomposes to form water and co2 .	Original	i am going to introduce you to the internet movie database .	Original	this restricts the availability of cash to them and other countries too start banning them .
Music	phoebe and jack , he 's gorgeous and she loves to get him !	Science	i am going to investigate the internet to google .	Science	this reduces the intake of the other molecules to produce them and thus are too large .
Politics	nixon (captured) he lied and voted for bush to cause his country .	Politics	i am going to skip the internet to get you checked .	Music	this is the cheapest package of them too .

Figure 10: Full sheet of Yahoo topic transfer on answers.