Educating Text Autoencoders:
Latent Representation Guidance via Denoising

Tianxiao Shen  Jonas Mueller  Regina Barzilay  Tommi Jaakkola
Text Autoencoders

Represent sentences as vectors in a latent space
Text Autoencoders

Represent sentences as vectors in a latent space
Manipulate sentences via modifying their latent representation

This talk is awesome

This talk was awesome

This talk is awful
Latent Space Geometry

Which mapping between sentences and latent vectors will be learned?

Sentence Space

This talk is awesome
This talk was awesome

It’s sunny today
It was sunny yesterday

Latent Space

?
Latent Space Geometry

Fortuitous geometry that captures sentence semantics is unlikely to arise.
Latent Space Geometry

Fortuitous geometry that captures sentence semantics is unlikely to arise. Minimal latent space manipulations can yield random, unpredictable changes in the resulting text.
Adversarial Autoencoder (AAE)

encoder $E$, decoder $G$, discriminator $D$
sample $z \sim p(z), x \sim p_G(x|z)$ to generate new data

This talk is awesome

min

$\max_{E,G} \min_{D} \mathcal{L}_{\text{rec}} - \lambda \mathcal{L}_{\text{adv}}$

[Makhzani et al., 2015]
Our Model: Denoising AAE (DAAE)

Introduce a perturbation process \( C \) that maps \( x \) to nearby \( \tilde{x} \) (e.g., randomly drop each word with probability \( p \)), and ask the model to reconstruct \( x \) from \( \tilde{x} \) [Vincent et al., 2008]

\[
\min_{E,G} \max_D \mathcal{L}_{rec} - \lambda \mathcal{L}_{adv}
\]
Toy Experiment

\( \mathcal{X} = \{0, 1\}^{50}, \ Z = \mathbb{R}^2 \) Data stem from 5 clusters, with 100 sequences sampled from each

AAE Latent Space

similar sequences \( \rightarrow \) distant representations

DAAE Latent Space

similar sequences \( \rightarrow \) similar representations
Toy Experiment

\[ \mathcal{X} = \{0, 1\}^{50}, \quad \mathcal{Z} = \mathbb{R}^2 \]  
Data stem from 5 clusters, with 100 sequences sampled from each.

Theoretically analyze which type of \( x - z \) mappings will be learned by AAE and DAAE under global optimality.

**AAE Latent Space**

**DAAE Latent Space**

- similar sequences \( \rightarrow \) distant representations
- similar sequences \( \rightarrow \) similar representations
AAE Can Learn a Random Mapping Between $X$ and $Z$

**Theorem 1.** With high-capacity encoder/decoder networks, any assignment between \( \{x_1, \cdots, x_n\} \) and \( \{z_1, \cdots, z_n\} \) can achieve the same optimal value under the AAE objective.
DAAE Learns to Map Similar $X$ to Close $Z$

**Theorem 2.** In a simple scenario with only four examples, the optimal value under the DAAE objective is achieved when close pairs of $x$ are mapped to close pairs of $z$. 
Theorem 3 (sketch). Suppose $x_1, \ldots, x_n$ are divided into $n/K$ clusters of equal size $K$. Let the perturbation process $C$ be uniform within clusters. The DAAE objective is “best achieved” when examples in the same cluster are mapped to the latent space in a manner that is well-separated from encodings of other clusters.
Experiments

Compare DAAE with:

• AAE [Makhzani et al., 2015]
• Latent-noising AAE (LAAE) [Rubenstein et al., 2018]
• $\beta$-VAE [Higgins et al., 2017]
• ARAE [Zhao et al., 2018]

Evaluate:

• Neighborhood Preservation
• Generation-Reconstruction Trade-Off
• Style Transfer
• Sentence Interpolation

Datasets:

• Yelp reviews
• Yahoo answers
Neighborhood Preservation

<table>
<thead>
<tr>
<th>Source</th>
<th>my waitress katie was fantastic, attentive and personable.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAE</td>
<td>my cashier did not smile, barely said hello.</td>
</tr>
<tr>
<td></td>
<td>the service is fantastic, the food is great.</td>
</tr>
<tr>
<td></td>
<td>the employees are extremely nice and helpful.</td>
</tr>
<tr>
<td>DAAE</td>
<td>the manager, linda, was very very attentive and personable.</td>
</tr>
<tr>
<td></td>
<td>stylist brenda was very friendly, attentive and professional.</td>
</tr>
<tr>
<td></td>
<td>the manager was also super nice and personable.</td>
</tr>
</tbody>
</table>

For each sentence’s 10-NN in terms of normalized edit distance, count how many of them lie among the k-NN in the latent space
Generation-Reconstruction Trade-Off
Generation-Reconstruction Trade-Off
Generation-Reconstruction Trade-Off
Generation-Reconstruction Trade-Off

![Graph showing the trade-off between generation and reconstruction](image)

- **β-VAE**
- **AAE**
- **LAAE**
- **DAAE**

Parameters:
- **β**: 0.01, 0.05, 0.1, 0.15
- **λ₁**: 0.01, 0.05, 0.2
- **p**: 0.1, 0.3

**Perplexity** vs **Reconstruction BLEU**
Unsupervised Text Style Transfer

**style 1**
- the pizza was pretty bland
- we got steak and drink
- I had to knock it down a star

**style 2**
- everything is pretty fresh
- food seems decent overall
- you get what you pay for

\[ \text{dec} ( \text{enc} ( \text{input} ) \pm v ) = ? \]

No style labels required during training!
Tense Transfer

• AAE has the highest BLEU but the lowest ACC → not change the source sentence
Tense Transfer

• DAAE achieves the highest ACC, the lowest PPL, relatively high BLEU
  ✓ proper tense ✓ high quality ✓ faithful to source
Tense Transfer

- DAAE achieves the highest ACC, the lowest PPL, relatively high BLEU
  - proper tense
  - high quality
  - faithful to source

<table>
<thead>
<tr>
<th>Input</th>
<th>the staff is rude and the dr. does not spend time with you.</th>
</tr>
</thead>
<tbody>
<tr>
<td>β-VAE</td>
<td>the staff was rude and the dr. did not spend time with your attitude.</td>
</tr>
<tr>
<td>AAE</td>
<td>the staff was rude and the dr. does not spend time with you.</td>
</tr>
<tr>
<td>LAAE</td>
<td>the staff was rude and the dr. is even for another of her entertained.</td>
</tr>
<tr>
<td>DAAE</td>
<td>the staff was rude and the dr. did not make time with you.</td>
</tr>
</tbody>
</table>
Sentiment Transfer

• As the scaling factor increases, the resulting sentences generated by DAAE get more and more positive/negative

<table>
<thead>
<tr>
<th>Input</th>
<th>the food is entirely tasteless and slimy.</th>
</tr>
</thead>
<tbody>
<tr>
<td>+v</td>
<td>the food is tremendous and fresh</td>
</tr>
<tr>
<td>+1.5v</td>
<td>the food is sensational and fresh</td>
</tr>
<tr>
<td>+2v</td>
<td>the food is gigantic</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input</th>
<th>the patrons all looked happy and relaxed</th>
</tr>
</thead>
<tbody>
<tr>
<td>-v</td>
<td>the patrons all helped us were happy and relaxed</td>
</tr>
<tr>
<td>-1.5v</td>
<td>the patrons that all seemed around and left very stressed</td>
</tr>
<tr>
<td>-2v</td>
<td>the patrons actually kept us all looked long and was annoyed</td>
</tr>
</tbody>
</table>
Sentiment Transfer

- As the scaling factor increases, the resulting sentences generated by DAAE get more and more positive/negative
Sentiment Transfer

• DAAE with $\pm 1.5v$ is comparable to previous models trained with sentiment labels [Shen et al., 2017]
Sentence Interpolation via Latent Space Traversal

\[ z_i = t \cdot z_1 + (1 - t) \cdot z_2 \]
<table>
<thead>
<tr>
<th>AAE</th>
<th>DAAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>it’s so much better than the other chinese food places in this area.</td>
<td>it’s so much better than the other chinese food places in this area.</td>
</tr>
<tr>
<td>it’s so much better than the other food places in this area.</td>
<td>it’s much better than the other chinese places in this area.</td>
</tr>
<tr>
<td>better, much better.</td>
<td>better than the other chinese places in this area.</td>
</tr>
<tr>
<td>better than other places.</td>
<td>better than the other places in charlotte.</td>
</tr>
<tr>
<td></td>
<td>better than other places.</td>
</tr>
</tbody>
</table>
Takeaways

- Minimizing $D(p_{data}(x)\|p_{model}(x))$ does NOT ensure X-structure is preserved in Z-space
- Denoising helps induce latent space organization
- DAAE best preserves sequence neighborhood, provides superior generation-reconstruction trade-off, and enables zero-shot style transfer

Moving Forward

- Better/task-specific text perturbations
- Additional properties of latent space geometry
- Finer control over text generation

https://github.com/shentianxiao/text-autoencoders

Thank you!