Fine Grained Text Style Transfer using Diffusion

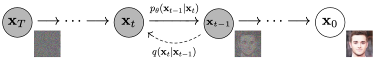
CS6420 - Topics in Deep Learning

Anudeep Rao Perala (CS21BTECH11043)
Asli Nitej Reddy Busireddy (CS21BTECH11011)

Contents

- Introduction to Diffusion
- Problem Statement
- Background of the Problem
- Diffusion in Seq2Seq setting
- State of the Art approach
- StylePTB Dataset
- Future Extensions
- Conclusion

Introduction to Diffusion



- A diffusion model typically contains forward and reverse processes. Given a data point sampled from a real-world data distribution x_0, the forward process gradually corrupts x into a standard Gaussian noise $x \sim N(0, I)$.
- For each forward step t [1, 2, ..., T], the noise addition is controlled by $q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right)$, with $\beta_t \in (0, 1)$.
- Once the forward process is completed, the reverse denoising process tries to gradually reconstruct the original data x_0 via sampling from x_T by learning a diffusion model f_θ.

Math in Diffusion

- We want to minimize the negative log likelihood of $p_\theta(x_0)$, for that following the regular mathematical steps involved in diffusion we have the result of:

$$-\log p_\theta(x_0) \leq \mathbb{E}_{q(x_{1:T}|x_0)}\left[-\log \frac{p(x_{0:T})}{q(x_{1:T}|x_0)}\right] \tag{1}$$

The RHS is called **ELBO**. We try to minimze ELBO for minimizing -log $p_\theta(x_0)$.

Math in Diffusion contd.

- Upon expanding ELBO we get the final expression:

$$= -\underbrace{\mathbb{E}_{q(x_1|x_0)}\left[\log p_\theta(x_0|x_1)\right]}_{L_0} + \underbrace{D_{KL}(q(x_T|x_0)\|p(x_T))}_{L_T}$$
$$+ \underbrace{\sum_{t=2}^{T}\mathbb{E}_{q(x_t|x_0)}\left[D_{KL}(q(x_{t-1}|x_t, x_0)\|p_\theta(x_{t-1}|x_t))\right]}_{L_{t-1}} \tag{2}$$

- L_0**:** This can be interpreted as the reconstruction term.
- L_T**:** This term has no optimization as it has no parameters and for large T the final distribution is Gaussian which makes this term zero.
- L_t**:** Tries to make the distribution at x_t consistent, from both forward and backward processes. We try to minimize this.

Minimizing L_{t-1}

- We try to:

$$\arg\min_{\theta} D_{KL}(q(x_{t-1}|x_t, x_0) \| p_\theta(x_{t-1}|x_t)) \tag{3}$$

- We set the variances of the two Gaussian's to match exactly, upon performing the mathematical steps we end at the result where, optimizing the KL Divergence term reduces to minimizing the difference between the means of the two distributions.

$$\arg\min_{\theta} \frac{1}{2\sigma_q^2(t)} \|\mu_\theta(x_t, t) - \mu_q(x_t, x_0)\|_2^2 \tag{4}$$

Further simplification of Minimizing L_{t-1}

- Upon simplifying the equation:

$$q(x_{t-1}|x_t, x_0) = \frac{q(x_t|x_{t-1}, x_0)q(x_{t-1}|x_0)}{q(x_t|x_0)} \tag{5}$$

- We end up at a relation:

$$q(x_{t-1}|x_t, x_0) \propto \mathcal{N}\left(x_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)x_0}{1-\bar{\alpha}_t}}_{\mu_q(x_t,x_0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\Sigma_q(t)}\right) \tag{6}$$

- So from the above equation we get the value of $\mu_q(x_t, x_0)$.

Further simplification of Minimizing L_{t-1} contd.

- We can now similarly formulate $\mu_\theta(x_t, t)$ by setting it to the following form:

$$\mu_\theta(x_t, t) = \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})x_t + \sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)\hat{x}_\theta(x_t, t)}{1 - \bar{\alpha}_t} \quad (7)$$

- Where $\hat{x}_\theta(x_t, t)$ is parameterized by a neural network that predicts x_0 from noisy image x_t and time index t.
- After substituting the above values of $\mu_q(x_t, x_0)$ and $\mu_\theta(x_t, t)$ in eq(4) the optimization problem simplifies to:

$$\arg\min_\theta \frac{1}{2\sigma_q^2(t)} \frac{\bar{\alpha}_{t-1}(1 - \alpha_t)^2}{(1 - \bar{\alpha}_t)^2} \|\hat{x}_\theta(x_t, t) - x_0\|_2^2 \quad (8)$$

Problem Statement

- We target the Seq2Seq text generation task for fine grained control over text-to-text modification.
- Given a m-length source sequence $w^x = w_1^x, ..., w_m^x$ and style tokens $s = s_1, , ..., s_k$ we aim to learn a diffusion model that can produce a n-length target sequence $w_y = w_1^y, ..., w_n^y$ conditioning on the source sequence and the style tokens.

Background of the Problem

- **Text style transfer** aims to controllably generate text with targeted stylistic changes while maintaining core meaning from the source sentence.
- **Fine grained TST** aims to have control on fine-grained/ low level stylistic changes while maintaining the core meaning.
- We have 4 categories of fine-grained style constructs:
  - Lexical transfer
  - Syntax transfer
  - Semantic transfer
  - Thematic transfer

Style constructs

- **Lexical Transfer**: Involves word level changes focusing on vocabulary and word meanings. operations such as replacing words with their synonyms or antonyms.

  ex: The cat is very quick $\xrightarrow{\text{synonym change}}$ The cat is very slow

- **Syntax Transfer**: Modifies grammatical structures without altering the content. Involves transforming sentence elements like tense, voice, or proposition positions.

  ex: She writes a letter $\xrightarrow{\text{past tense}}$ She wrote a letter

Style constructs contd.

- **Semantic Transfer**: These changes affect the meaning of the sentence which includes removing or adding information. These changes are beyond just word or syntax-level modifications.
  ex: The dog is barking loudly at the stranger $\xrightarrow{\text{Info removal}}$ The dog is barking at the stranger (adj. loudly is removed)
- **Thematic Transfer**: Adjusts the emphasis within a sentence to highlight different parts to shift the perspective or importance.
  ex: The comparable year-earlier number was 56 million a spokesman said $\xrightarrow{\text{attention}}$ A spokesman said the year-earlier number of 56 million was comparable (56 million $\rightarrow$ comparable)

TST

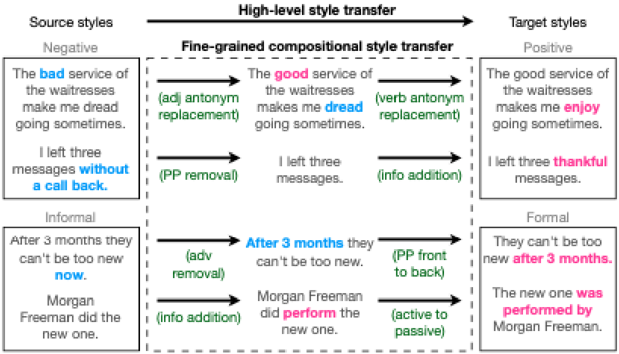
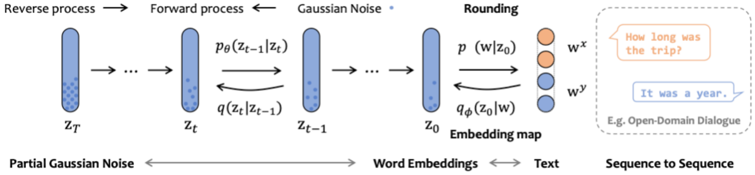


Figure: Fine grained style transfer to achieve high-level style transfer

Diffusion in Seq2Seq setting - DiffuSeq

- Now we look into the setting of diffusion where the target sequence is on a source sequence.



Forward Process with Partial Noising

- We have an embedding function EMB(w). given a pair of source sequence w^x and target sequence w^y DiffuSeq tries to learn the unified feature space of $\text{EMB}(\mathbf{w}^{x \oplus y}) \in \mathbb{R}^{(m+n) \times d}$. $(|w^x| = m, |w^y| = n)$
- We model this sequence of embedding to a new markov transition parametrized $q_\phi(z_0 | \mathbf{w}^{x \oplus y}) = \mathcal{N}(\text{EMB}(\mathbf{w}^{x \oplus y}), \beta_0 \mathbf{I})$.
- We now define $z_t = x_t \oplus y_t$, where $x_t \in w^x$ and $y_t \in w^y$. In each forward step $q(\mathbf{z}_t | \mathbf{z}_{t-1})$ we inject noise only into y_t, unlike conventional diffusion models. This modification is called **Partial Noising**.

Reverse Process

- Our goal is to recover original z_0 by denoising the z_t. By the learning process $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t), \sigma_\theta(\mathbf{z}_t, t))$.
- μ_θ and σ_θ are parameters for the predicted mean.
- We model $f_\theta(z_t, t)$ as our NN, we use transformer architecture to model f_θ, as this models the semantic relation between x_t and y_t.
- The model tries to learn the diffusion model, embedding parameters jointly.
- The variational lower bound(L_{vlb}) is formulated as:

$$\mathcal{L}_{\text{vlb}}(\mathbf{w}) = \mathbb{E}_{q_\phi(z_0|\mathbf{w})}\left[\mathcal{L}_{\text{vlb}}(z_0) + \log q_\phi(z_0|\mathbf{w}) - \log p(\mathbf{w}|z_0)\right] \tag{9}$$

Where $L_{\text{vlb}}(z_0)$ corresponds to the standard variational lower bound in diffusion.

Reverse Process contd.

- On modifying eq(9) more we get:

$$\mathcal{L}_{\text{vlb}}(\mathbf{w}) = \mathbb{E}_{q_\phi(z_{0:T}|\mathbf{w})}\Bigg[\underbrace{\log \frac{q(z_T|z_0)}{p_\theta(z_T)}}_{L_T} + \sum_{t=2}^{T} \underbrace{\log \frac{q(z_{t-1}|z_0, z_t)}{p_\theta(z_{t-1}|z_t)}}_{L_{t-1}} + \underbrace{\log \frac{q_\phi(z_0|\mathbf{w})}{p_\theta(z_0|z_1)}}_{L_0} - \underbrace{\log p(\mathbf{w}|z_0)}_{L_{\text{round}}}\Bigg] \tag{10}$$

Reverse Process contd.

- After doing all the approximations like how we do in standard diffusion models we get:

$$\mathcal{L}_{\text{vlb}}(\mathbf{w}) = \min_{\theta} \Bigg[\|\mu(z_T)\|^2 + \sum_{t=2}^{T} \|z_0 - f_\theta(z_t, t)\|^2 + \|\text{EMB}(\mathbf{w}^{x \oplus y}) - f_\theta(z_1, 1)\|^2 + \mathcal{R}(\|\mathbf{z}_0\|^2) \Bigg] \tag{11}$$

- During training the model estimates the z_0 via $f_\theta(z_t, t)$.
- The term $\mathcal{R}(\|\mathbf{z}_0\|^2)$ is introduced to learn regularize the embedding learning.

Inference

- Given the condition $\mathrm{EMB}(w^x)$, we randomly sample $y_T \sim N(0, I)$ and concatenate y_T with $\mathrm{EMB}(w^x)$ to obtain z_T. We now repeat the reverse process until we arrive at z_0 by calculating z_0^{temp}.
- We sample $\mathbf{z}_{t-1}$ from $q(\mathbf{z}_{t-1} \mid f_\theta(\mathbf{z}_t, t), \mathbf{z}_t)$, which is fed as input to the next diffusion step.
- The equation for obtaining $\mathbf{z}_{t-1}$:

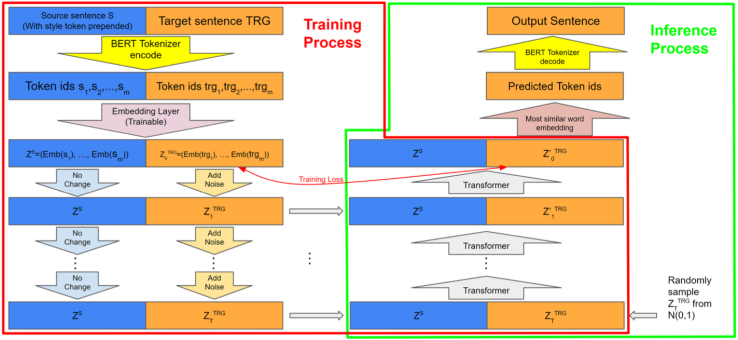
$$\mathbf{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} f_\theta(\mathbf{z}_t, t) + \sqrt{1 - \bar{\alpha}_{t-1}}\epsilon, \text{where } \bar{\alpha}_t \quad = \prod_{i=0}^{t} (1 - \beta_i)$$

- At each sampling step anchoring function is executed towards the obtained z_{t-1} which does:
  - Rounds the obtained z_{t-1} back to word embedding space.
  - Replaces the part of recovered z_{t-1} that belongs to w_x with the original x_0.

State of the Art approach

- Adopted DiffuSeq for performing fine-grained text style transfer.
- We first define a set of special style tokens, one for each possible individual fine- grained transfer. If we wish to perform one or more transfer on the source sentence, we will prepend the corresponding special token(s) to the beginning of the source sentence to form the condition S.

Architecture



- Used BERT tokenizer for converting text into tokens.
- Then we include a token embedding layer to encode both the source(prepended with style tokens) and target.

Training

- Both the diffusion transformer and the token embeddings are initialized randomly and jointly optimized.
- Z^S are source embeddings and Z_0^{TRG} are target embeddings.
- We then apply the partial noise in forward process until $t \sim U(1, T)$, after which we get Z_t^{TRG}. We then concatenate Z^S and Z_t^{TRG} input that to the diffusion transformer.
- We then follow the loss in DiffuSeq for minimization.

Inference

- We randomly initialize ${Z^*_T}^{TRG} \sim N(0,1)$, and encode the condition (source sentence and style tokens) into Z^S.
- Then we concatenate them and use the transformer to predict a temporary ${Z^*_{0_{temp}}}^{TRG}$, then we add $\mathbf{T-1}$ steps of noise to obtain ${Z^*_{T-1}}^{TRG}$. Now for each embedding in finally obtained Z_0^{TRG}, we find the closest embedding in our token embedding layer by cosine distance, and decode the embedding to that token.
- Then we combine the tokens to form the output sentence in natural language.

Dataset

- We are working with StylePTB dataset with:
  - We have paired sentences under 21 fine-grained stylistic changes.
  - We even have compositions of multiple transfers for more complex modeling.
- We classify transfers as Easy, Medium, Hard by calculating the token level Hamming distance between original and transferred sentences.

Future Work

- And we want to see how the work of fine-grained style transfer is done in the domain of vision for further ideas.
- Yet to be explored.

'

References

- Understanding Diffusion Models: A Unified Perspective
- DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models
- Fine-grained Text Style Transfer with Diffusion-Based Language Models
- StylePTB: A Compositional Benchmark for Fine-grained Controllable Text Style Transfer