Sequence-to-sequence Text Generation with Coupled Diffusion Process

Boming Yang[1], **Aosong Feng**[2], **Zihui Li**[1]
[1]University of Tokyo, [2]Yale University

Abstract

In recent years, diffusion models have exhibited notable advancements in domains such as image and audio processing, emerging as a prominent trend in generative models. However, the discrete signals of textual language in natural language processing (NLP) necessitate further exploration of the application of diffusion models. This study aims to examine the effects and outcomes of the denoising process in diffusion models for text generation tasks involving sequence-to-sequence (Seq2Seq). It investigates the influence of target denoising and full denoising on the performance of paraphrase tasks, while also analyzing the fluctuation patterns of evaluation metrics throughout the training process. Our findings indicate that, in contrast to expectations, full denoising resulted in a decrease in performance for paraphrase tasks. This decline was observed in both the coherence and fluency of generated text, suggesting that full denoising may be less suited for complex sequence-to-sequence text generation in the current model framework.

1 Introduction

The success of the diffusion model in high-resolution image generation presents a promising avenue for advancing sequence-to-sequence (Seq2Seq) generation within the natural language domain. Drawing inspiration from its success in visual tasks, the diffusion model can be adapted to enhance the coherence and contextuality of generated sequences in natural language processing by sequentially denosing the sequence into the desired target distribution, either in the embedded continuous space (Gong et al., 2022; Li et al., 2022) or discrete space (Hoogeboom et al., 2021; Austin et al., 2021).

To pave the way for adopting diffusion models in NLP, a key issue to solve is conditional generation, which encompasses a broad range of Seq2Seq applications, including machine translation and QA. Unlike conventional text-to-image generation (Rombach et al., 2022), where the text condition must be encoded by a modality-aligned encoder (e.g., CLIP (Radford et al., 2021)) and takes effect through cross-attention, text-to-text generation sidesteps the complexity of modality alignment. This allows for direct modeling by coupling scores derived from source and target distributions. DiffusionLM (Li et al., 2022) achieves conditional controls using classifier guidance, introducing an additional classifier during target sequence generation. DiffuSeq (Gong et al., 2022) models the target distribution using a diffusion process and models the target score by concatenating noisy target samples with clean source samples as model inputs. However, these methods heavily rely on the conditional modeling power of the language model in use and only leverage the iterative refinement capability of the diffusion process in target domain generation. Additionally, the distribution mismatch between noised target samples and clean source samples increases the burden on the score prediction model to handle inputs with different distributions.

In this work, we consider the diffusion process on both the source and target domain and explore the possibility of coupling both source and target diffusion processes by modeling the joint distribution. We apply the diffusion forward process to both the source and target domains, thereby decreasing the gap between the two by adding white noise until they converge to the same Gaussian distribution. The noised source and target samples at the intermediate steps are then used as the score prediction model input. The predicted score of the joint distribution can then be adopted for joint sampling by running both reverse diffusion processes or conditional sampling by running source forward diffusion and target reverse diffusion. This adaptation holds substantial potential for improving the performance of various language-oriented tasks,

such as QA, machine translation, and dialogue generation, ushering in the possibility of more effective design of coupled diffusion systems for conditional and joint distribution modeling. However, contrary to expectations, our experiments revealed that the full denoising method led to a deterioration in performance across several tasks, particularly in machine translation and dialogue generation, indicating challenges in the current approach and the need for further refinement in this method.

2 Method

We have developed an advanced method to demonstrate the impact of the bi-denoising and diffusion processes on the sequence-to-sequence framework.

2.1 Dual-Noising in the Forward Process

In alignment with DiffusionLM (Li et al., 2022), we utilize an embedding layer to convert the discrete text, w, into a continuous space. Notably, in our case, the paired representations (x, y) are learned concurrently, effectively applying the process to both x and y simultaneously. Following this, a unified embedding is generated from the concatenation of x and y. Upon completion of this process, we are poised to incorporate the standard diffusion forward process into discrete textual input by extending a new Markov transition $q_\phi(z_0|w^{x\oplus y}) = \mathcal{N}(EMB(w^{x\oplus y}), \beta_0 I)$.

2.2 Dual-Denoising in the Reverse Process

The aim of the reverse process is to recover the original z_0 by denoising z_t: $p_\theta(z_{0:T} := p(z_T)\sum_{t=1}^{T} p_\theta(z_{t-1}|z_t)$. To achieve this, we start a series of reverse iterations where each iteration is designed to remove the noise added in the forward process for both x and y. This includes operating an inference network to generate a noise distribution at each time interval. By sequentially applying this network in reverse order, noise is incrementally removed, thereby guiding the combined sequence back to the initial state, z_0. The successful recovery of z_0 from z_t provides further insights into the functional dynamics of the sequence-to-sequence framework, in turn offering potential strategies to intensify the effectiveness of the dual-application denoising procedure.

3 Experiments

We conducted experiments on the Quota Question Pairs (QQP) dataset, which was collected from the Quora community question answering forum. The dataset consists of 147,000 positive pairs and is used for the Paraphrase Generation Task.

For evaluation, we selected four commonly used metrics: BLEU(Papineni et al., 2002), ROUGE(Lin, 2004), and BERTScore(Zhang et al., 2019). We utilized a Transformer model with 12 layers, a maximum sequence length of 128, an embedding dimension of d = 128, diffusion step T = 2000, and a square-root noise schedule. To mitigate issues with out-of-vocabulary generation, we employed Byte Pair Encoding (BPE)(Salimans et al., 2016) to construct the vocabulary.

The experiments were executed on NVIDIA A100 Tensor Core GPUs, employing four GPUs for training and a single GPU for sampling.

4 Results

As indicated in Tab 1, we conducted a training process consisting of 50,000 steps and stored checkpoints at every 10,000 steps for decoding purposes. Through a thorough analysis of the variations in evaluation metrics observed at each checkpoint on the test dataset, we investigated the influence of the denoising process employed in the diffusion model.

Initially, at 10,000 steps, all metrics were low as it was still in the early stage of training. By 20,000 steps, the length had stabilized, and the Dist-1 score had risen above 90. In the subsequent 30,000 steps, there was a significant improvement in the BLEU and ROUGE metrics. By 50,000 steps, the model had already learned the most accurate results and answer lengths effectively.

Case Study We randomly selected a subset of samples to exemplify the outcomes of the decoding process. In this context, the term "recover" denotes the output generated by the diffusion model, referred to as $\hat{y}$, while "reference" corresponds to the target sentences in the dataset, indicated as y. Utilizing these two elements, various metrics such as BLEU can be computed. Additionally, the term "source" pertains to the x value in the dataset. The provided examples clearly demonstrate the model's ability to acquire accurate grammatical representations of the queries and manifest remarkable word similarity. For instance, in Example 1, there is a close resemblance between "YouTube" and "channel."

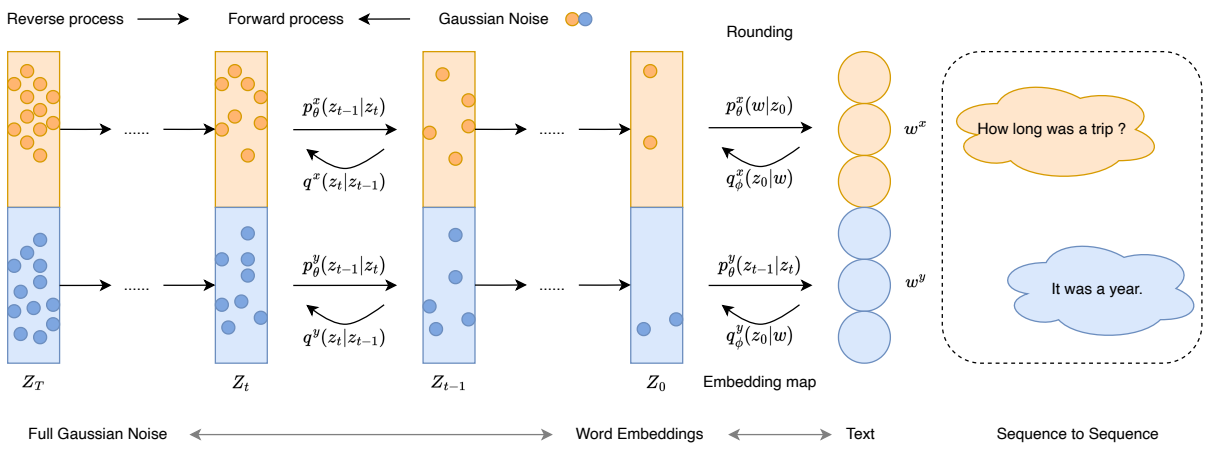

Figure 1: The diffusion process framework

Model	BLEU	ROUGE	BERTScore	Dist-1	len
DiffuSeq	0.1788	0.5272	0.7931	0.9749	10.98
Full-Noised(10000 interval)	4.1062e-05	0.000272	0.2728	0.5663	5.4712
Full-Noised(20000 interval)	0.01713	0.09945	0.4623	0.92441	12.2084
Full-Noised(30000 interval)	0.03477	0.17849	0.5464	0.90277	11.476
Full-Noised(40000 interval)	0.03942	0.19037	0.5665	0.8955	11.1572
Full-Noised(50000 interval)	0.04059	0.19196	0.5708	0.8929	11.1524

Table 1: Main Result

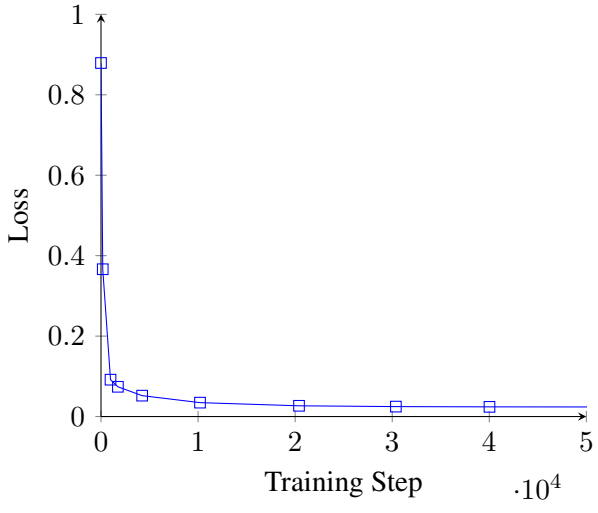

Figure 2: Loss of The Training Stage.

```
{recover: "[CLS] how do i earn money
    with my startup channel? [SEP]",
    reference: "[CLS] how can i see all
    my youtube comments? [SEP]",
    source: "[CLS] how do i read and
    find my youtube comments? [SEP] [
    SEP]"}

{recover: "[SEP] [CLS] why is my
    question marked as needing
    improvement when it is perfectly
    clear and well written? [SEP]",
    reference: "[CLS] why do people ask
    quora questions which can be
    answered easily by google? [SEP]",
    source: "[CLS] why are so many
    quora users posting questions that
    are readily answered on google? [
    SEP] [SEP]"}

{recover: "[CLS] as a good? video in
    chemistry, is it better to prepare
    for ups 12, e ors? [SEP]", reference
    : "[CLS] how do i prepare for civil
    service? [SEP]", source: "[CLS]
    how do we prepare for upsc? [SEP] [
    SEP]"}
```

Figure 3: Decoding Case.

5 Conclusion

In this study, we explored the use of diffusion processes in both source and target domains for natural language processing tasks. By applying the diffusion forward process to both domains and introducing white noise, we successfully decreased the gap between them, leading them to converge to a similar Gaussian distribution. We then utilized the noised samples from intermediate steps as inputs for a score prediction model. This model's predicted scores enable joint sampling through reverse diffusion processes or conditional sampling by combining forward diffusion in the source with reverse diffusion in the target. Our approach shows significant promise in enhancing performance in language-oriented tasks like QA, machine translation, and dialogue generation. It opens new avenues for the design of coupled diffusion systems for more effective conditional and joint distribution modeling. However, our experimental findings revealed that the full denoising approach, contrary to expectations, led to reduced performance, particularly in complex tasks like machine translation and dialogue generation. This suggests that while the conceptual framework is promising, the current implementation of full denoising requires further refinement to effectively handle the nuances of these language-oriented tasks.

References

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. 2022. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. 2021. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34:12454–12465.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. 2022. Diffusion-lm improves controllable text generation. *ArXiv*, abs/2205.14217.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *ArXiv*, abs/1606.03498.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.