MambaDiff: Revolutionizing Seq2Seq Models with Diffusion Model and Mamba Architectures

Guannan He, Xiangyu Zhang, Xinyu Chang

Johns Hopkins University

1.Background and Contribution Highlights

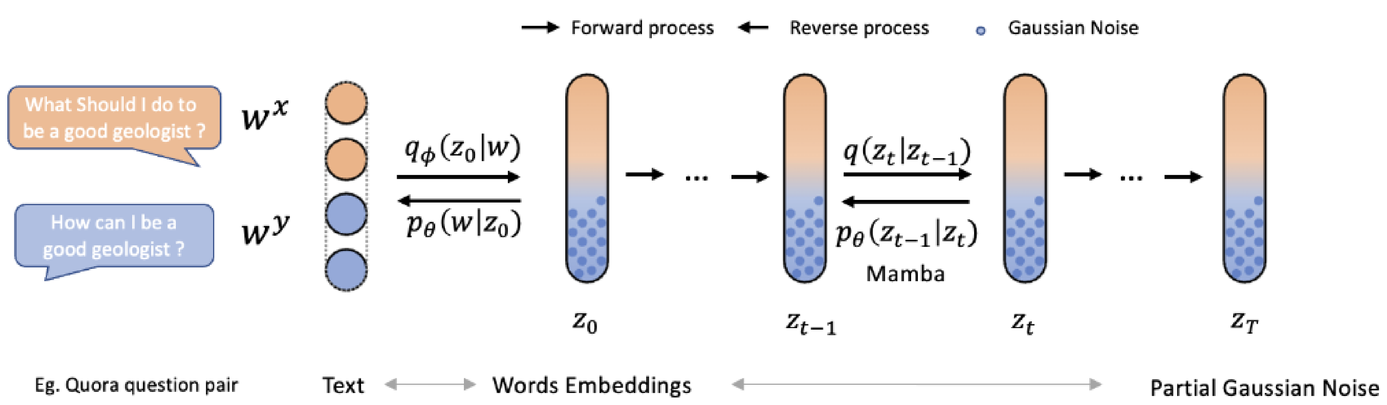
Motivated by two models:

- **DiffuSeq**: Diffusion model for natural language tasks.
  Advantage: generate coherent and contextually relevant text, classifier-free
- **Mamba**: selective state-space model + selective mechanism for optimizing input filtering.
  Advantage: has ability to handle longer contexts efficiently with a smaller memory footprint, speedy.

In our project, we take advantage of DiffuSeq and Mamba:

- DiffuSeq (Diffusion + Transformer) -> **MambaDiff (Diffusion + Mamba)**
- Greatly shorten the training time

2.DiffuSeq



DiffuSeq concatenates the space of x and y into z, the embedding space is jointly trained. In the forward process, we only impose partial noise on the space of y. The x signal stays in an un-noised state. In the reverse process, the neural network is optimized with the help of conditional signals x as **guidance**.

3.Mamba

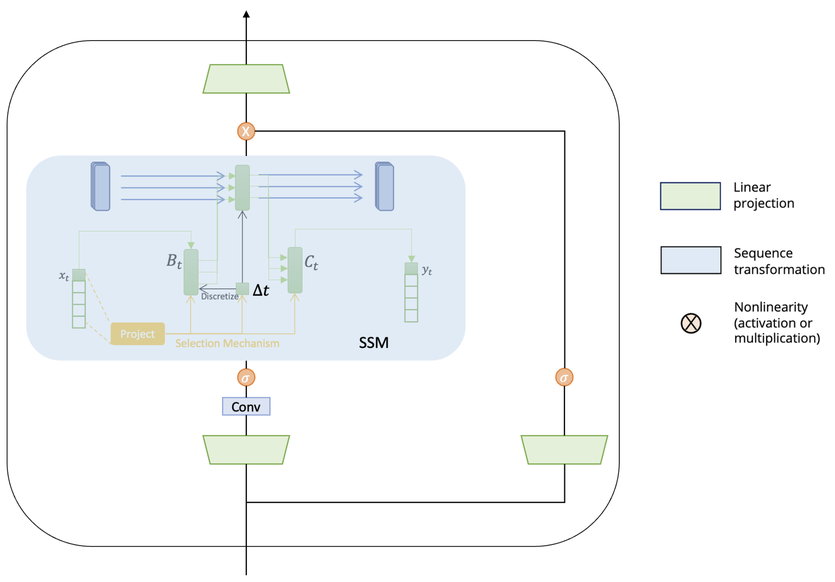


Fig 2. Mamba Architecture

Structured SSMs independently map each channel (e.g. $D = 5$) of an input x to output y through a higher dimensional latent state h (e.g. $N = 3$).

(Δ, A, B, C) parameters are constant across time (**Time-invariant**). After **discretization**, the parameters become **time-variant**, allowing more efficient computations.

4.Selected Results for Paraphrase Task

Seq2Seq contains many tasks, we select **Paraphrasing** to test the performance of our model. The **Paraphrasing** task involves crafting an alternative expression in the same language that conveys the same meaning. **Quora Question Pairs** (QQP) dataset consists of over 144k question pairs.

Methods	BLEU	R-L	Score	dist-1	Len
Paraphrase					
GRU-attention	0.1894	0.5129	0.7763	0.9423	8.31
GPT2-base FT	0.1980	0.5212	0.8246	0.9798	9.67
DiffuSeq (90 million para.) - 10k steps	0.0004	0.0022	0.2711	0.3056	62.08
DiffuSeq (90 million para.) - 50k steps	0.1921	0.5405	0.8042	0.9717	11.11
MambaDiff (7 million para.) - 10k steps	0.0069	0.0414	0.2937	0.9760	93.03
MambaDiff (7 million para.) - 60k steps	0.0774	0.3119	0.5755	0.9165	10.84
MambaDiff (7 million para.) - 70k steps	0.0857	0.3329	0.5944	0.9154	10.88
MambaDiff (7 million para.) - 80k steps	0.0895	0.3439	0.6032	0.9165	10.98

Table 1: Results on QQP dataset

BLEU: (0-1) measures the similarity of machine-translated text vs. a set of high-quality reference translations.
R-L: Rouge-L, good around 0.5, measures the longest common subsequence between the candidate text and the reference text.
Score: Bertscore, calculates the similarity between a machine translation output and a reference translation using sentence representation.
Dist-1: text diversity
Len: the average length of each sentence generated

5. Dataset Overview

Using Table 2 and Table 3 as references, the original sentence (source) will be compared with the paraphrase reference (reference) by evaluating the model's output (recover).

Original Sentence	How can I be a good geologist?
Paraphrase Referenct	What should I do to be a great geologist?

Table 2: Example of QQP

recover	what is the purpose of life?
reference	what's are the meaning of life?
source	what the meaning of this all life?

Table 3. Generation Output

6. Selected Plots for Paraphrase Task

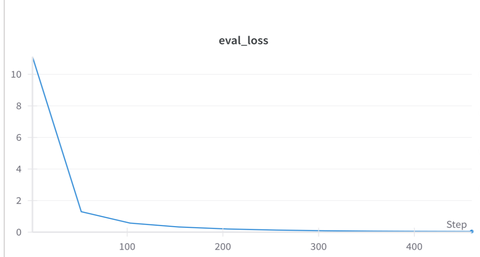


Fig3. Loss on Evaluation Dataset

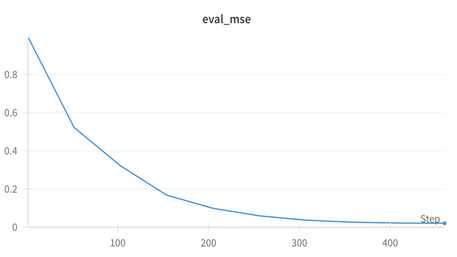


Fig4. Mean Squared Error on Evaluation Dataset

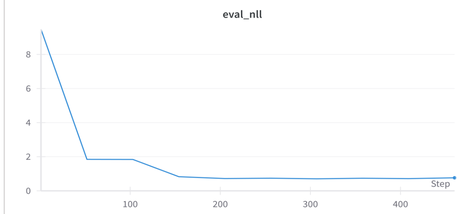


Fig5. Negative Log Likelihood on Evaluation Dataset

recover	what is the purpose of life?
reference	what's are the meaning of life?
source	what the meaning of this all life?

recover	how do i make money through youtube?
reference	how do i make money through youtube?
source	how can i earn money from youtube?

Fig6. Generation Output of MambaDiff

7. Conclusions & Future Works

Comparing the results of DiffuSeq (diffusion + Transformer) with our MambaDiff (diffusion + Mamba), it is evident that MambaDiff achieves competitive results while utilizing only one-tenth of the parameters. Additionally, MambaDiff demonstrates improved efficiency, with a runtime of 32 hours compared to DiffuSeq's 108 hours.
Comparing the results of two models at same step (10k step), we found that our model performs better than DiffuSeq at initial stage.

Future works:
The results of MambaDiff exhibit an increasing trend with additional steps and with the use of more parameters. This suggests that, given sufficient time, further runs will likely improve MambaDiff's performance, potentially surpassing DiffuSeq and other baseline models.

References

Gong, S., Li, M., Feng, J., Wu, Z., & Kong, L. (2022). Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933.*
Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752.*

Full References list is written in our report.