DMLP: A Library for Learning Text-Diffusion Model

Yunhao Li
yul080@ucsd.edu

Jieqi Liu
jil146@ucsd.edu

Zhiting Hu
zhh019@ucsd.edu

Abstract

Diffusion models achieved remarkable results in the image generation, while left insufficiently studied in natural language processing (NLP). To fully utilize the strength of diffusion model in NLP and accelerate the research cycle, we introduced Diffusion Model Learning Package (DMLP), a versatile Python library designed for the training, evaluation, and development of text diffusion models. This library focuses on an architecture synthesizing variational autoencoders (VAEs) with diffusion process, which enables text representation, reconstruction, and generation in one model. DMLP comes equipped with pre-defined functions and classes, offering users a package to implement, experiment, and compare customized text diffusion models. By providing a comprehensive toolkit for researchers and practitioners, DMLP aims to catalyze advancements in text generation and representation learning.

Website: https://yunhaoli12138.github.io/DMLP/
Code: https://github.com/YunhaoLi12138/DMLP.git

1 Introduction

Diffusion models are generative models originally designed for image generation, interpolation, and reconstruction (Sohl-Dickstein et al. 2015). Nowadays, combining with variational autoencoder (VAE) (Kingma and Welling 2019), diffusion models are extended to natural language processing (NLP) (Li et al. 2022), and text-diffusion has become a novel challenging field to explore. In this capstone project, we will introduce Diffusion Model Learning Package (DMLP), a versatile Python library designed for the training, evaluation, and development of text diffusion models.
DMLP is based on the Joint Autoencoding Diffusion (JEDI) proposed in "GENERATION, RECONSTRUCTION, REPRESENTATION ALL-IN-ONE: A JOINT AUTOENCODING DIFFUSION MODEL" (Liu et al. 2023b). JEDI was an architecture designed for generation, reconstruction, and representation in image, text, and gene fields, and DMLP focuses on the text-diffusion component - generation and reconstruction of sentences.
DMLP modularizes the JEDI model and allows a flexible combination of different VAE models and diffusion process. Both a predefined JEDI structure and abstract models are provided for customized training tasks. Moreover, DMLP contains a complete training, evaluation, and saving pipeline that reduces implementation workload. The significance of DMLP lies in its versatility, fostering innovation and exploration in the evolving field of text-diffusion. Its comprehensive functions further streamlines the development process, making DMLP a resource for advancing research and applications in NLP.

1.1 Literature Review

VAEs commonly face a challenge trading off state-of-the-art generation and accurate reconstruction due to its reconstruction loss and regularization loss (Bowman et al. 2016; Chen et al. 2016). Denoising Diffusion Probabilistic Models (DDPM) (Ho, Jain and Abbeel 2020) are successful in continuous data and can produce high-quality images. With a U-Net (Ronneberger, Fischer and Brox 2015) backbone, they simplified the training objective by proposing an innovative connection between diffusion models and the denoising score matching. However, compared with images, texts are discrete data that does not follow the conventional framework of diffusion models. Diffusion-LM (Li et al. 2022) addressed the obstacle preventing diffusion model succeeding in natural language processing by mapping discrete texts to a continuous space through an embedding process. In addition, it used additional classifiers to control the text generation process. The classifiers allowed the model to generate high-quality outputs for various input prompts. Diffusion-LM, nevertheless, demands high computational resources to finish the training for both the diffusion model and the classifiers. Upon this research, DIFFUSEQ (Gong et al. 2023) controlled the intermediate latent variables by unchanged prompts. Noise will not be added to the embedded prompt so that it can guide the model for accurate generation. This partial noising technique lowers the training requirements, but it still failed to generate meaningful representations in the latent space.
JEDI Liu et al. (2023b) addressed the issues discussed above and achieved text generation,

reconstruction, and representation by synthesizing the advantages of VAEs and diffusion models. It applied BERT (Devlin et al. 2019) as encoder and GPT-2 (Razavi, van den Oord and Vinyals 2019) as decoder. The VAE structure will learn parameters simultaneously with the diffusion model, which controls the VAE latent space, so JEDI can generate high-level sentences while retaining its reconstruction capability.
Based on JEDI, we proposed the library DMLP that facilitates deep learning researchers to solve real-world tasks with the most up-to-date model easily. It will be a comprehensive toolbox that ease the barrier to understand and apply JEDI and text-diffusion in general.

2 Methods

2.1 Background Information

2.1.1 DMLP Library Overview

DMLP achieved the integration of generation, reconstruction, and representation based on the JEDI model(Liu et al. 2023b). It constructed the parameterized encoder and decoder to realize a learnable prior, which addressed the gap between the predicted posterior and the prior. This method governs the training of the VAE latent space so that the trade-off between generation and reconstruction is resolved. The detailed structure of models used in DMLP library is depicted in Figure.1.

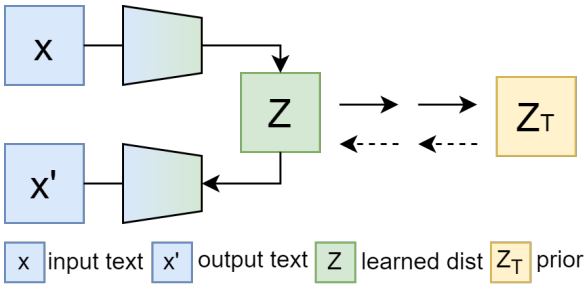


Figure 1: DMLP Model Visualization: JEDI architecture in DMLP Library(Liu et al. 2023b)

2.1.2 Variational Autoencoder

The Markov chain diffusion process deals with continuous data. In order to apply it to discrete texts, DMLP maps texts to a continuous space and convert numbers back to texts by utilizing a VAE architecture based on LatentOps model (Liu et al. 2023a). Traditionally, the VAE standard Gaussian prior is difficult to achieve (Bowman et al. 2016; Kingma et al. 2017). The discrepancy between the prior and the posterior may lead to unfaithful reconstruction. DMLP addressed this drawback by introducing a learnable prior modeled by the diffusion process in the latent space. The VAE architecture will be trained jointly with the diffusion process. As a result, although the diffusion model will not directly participate in

the reconstruction task, it controls and improves the encoded latent representation during the training process.
The encoder will create a latent representation of the input corpus, and based on the representation, the distribution of the texts can be modeled by mean and variance. By sampling from this distribution, the decoder can reconstruct the input texts.

2.1.3 Diffusion Process

Diffusion models are probabilistic models consisted of a forward process adding noise to the input data and a denoising reverse process (Sohl-Dickstein et al. 2015). They are famous for exceling data generation in computer vision fields but are insufficiently explored in the natural language processing. DMLP utilized this architecture to integrate text generation and reconstruction on one model.

Denoising Diffusion Probabilistic Models (DDPMs) (Ho, Jain and Abbeel 2020) provide rigorous formulations for diffusion models.
In the forward process, Gaussian noise will be added to the input data gradually until the information pattern contained in data is destroyed. This operation is described by

$$q(z_{1:T}|z_0) := \prod_{t=1}^{T} q(z_t|z_{t-1}), \quad q(z_t|z_{t-1}) := \mathcal{N}(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t \mathbf{I}), \tag{1}$$

where T is the number of diffusion steps, β_t represents the defined variance similar to hyperparameters, $z_1, \ldots, z_T$ represents latent vectors approaching a vector from a standard Gaussian distribution as T approaches infinity, z_0 is the input and $q(z_{1:T}|z_0)$ is the posterior distribution that will be learned (Ho, Jain and Abbeel 2020).
The reverse process denoises z_T back to z_0 by

$$p_\theta(z_{0:T}) := p(z_T) \prod_{t=1}^{T} p_\theta(z_{t-1}|z_t), \quad p_\theta(z_{t-1}|z_t) := \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t), \Sigma_\theta(z_t, t)). \tag{2}$$

The reverse process is a learned Gaussian transition from a standard Gaussian distribution vector (Ho, Jain and Abbeel 2020).

In DMLP, the diffusion process is trained jointly with the encoder and decoder. Instead of implementing full diffusion T for all z_0, DMLP followed the training algorithm proposed by DDPMs (Ho, Jain and Abbeel 2020). A random t is selected from Uniform({1, …, T}), marking the end of the forward process. Gradient descent steps are taken to compare the actual noise added and the predicted noise. Subtracting the noise can recover the original z_0. The simplified diffusion process and the noise calculation is describe by

$$\nabla_\theta ||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)||^2, \tag{3}$$

where ϵ_θ is an approximation of ϵ, $\bar{\alpha}_t$ represents a joint process of adding noise t times with $\alpha_t = 1 - \beta_t$ (Ho, Jain and Abbeel 2020).

The diffusion process targets the text generation task. A sample of z_T will be generated from a standard Gaussian distribution, and the reverse process will denoise z_T to z_0 by

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}}(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(z_t, t)) + \sigma_t z, \tag{4}$$

where z is sampled from a standard Gaussian distribution and σ_t comes from $\sigma_t \mathbf{I} = \Sigma_\theta(z_t, t)$.

2.2 DMLP Modules

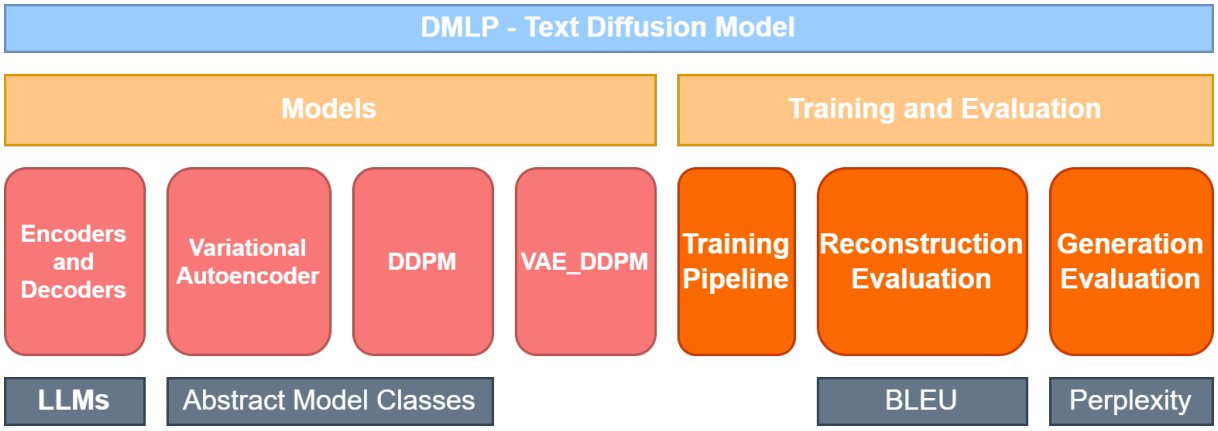


Figure 2: DMLP Module Overview

2.2.1 Models

The Models module of DMLP comprises three key components: abstract VAE, abstract Diffusion, and abstract VAE_Diffusion. These elements form the fundamental structure of the JEDI model, which serves as a reference for the design of DMLP. The abstract models offer users the flexibility to implement their architectures based on their preferences. Multiple combinations of encoders and decoders, along with corresponding tokenizers, are accessible through Hugging Face. Users can simply download the models and set them as instances of our classes.

Additionally, we have incorporated JEDI versions of the VAE and DDPM models within DMLP. This feature enables swift adaptation to new tasks, providing users with pre-configured models that can be readily employed for their specific applications. This approach streamlines the process of leveraging state-of-the-art models while offering the flexibility needed for customization.

The modular design facilitates both customization and quick hands-on projects, allowing users to seamlessly integrate various encoder and decoder combinations to suit their specific requirements.

2.2.2 Training

DMLP provides a complete training pipeline that allows users set up the training process by calling a single function. Users can fine-tune hyperparameters through function arguments. Specific keywords are listed in the DLMP documentation. The loss function by default is set as belowing:

$$\mathscr{L}(\lambda,\phi,\theta) = \mathbb{E}_q[\underbrace{\mathrm{KL}(q(z_T|z_0)||p(z_T)) + \sum_{t=2}^{T}\mathrm{KL}(q(z_{t-1}|z_t,z_0)||p_\theta(z_{t-1}|z_t))}_{\mathscr{L}_{\text{DDPM}}} \tag{5}$$

$$+\underbrace{\mathrm{KL}(q_\lambda(z_0|x)||p_\theta(z_0|z_1))}_{\mathscr{L}_{\text{align}}} + \underbrace{\log p_\phi(x|z_0)}_{\mathscr{L}_{\text{rec}}}], \tag{6}$$

where $\mathscr{L}_{\text{align}}$ and $\mathscr{L}_{\text{rec}}$ are new extensions unique to JEDI (Liu et al. 2023b).
$\mathscr{L}_{\text{align}}$ compares the latent representation generated by the encoder and the diffusion model. That is, it calculates the KL distance between $q_\lambda(z_0|x)$ and $p_\theta(z_0|z_1)$. $\mathscr{L}_{\text{rec}}$ measures the reconstruction loss of VAE.
While the customization of loss functions is permitted through the redefinition of loss functions, it is essential to ensure compatibility with the existing pipeline. Self-defined loss functions must adhere to a similar structure in terms of return values, aligning with the established conventions within the framework.

2.2.3 Evaluation

The evaluation module comprises two primary components: reconstruction and generation. In the reconstruction phase, the trained model's performance is assessed in reconstructing input sentences using the BLEU score. A higher BLEU score means a more accurate reconstruction result. It is worth noting that the BLEU score relies on exact matches. Therefore, in scenarios where the downstream task involves paraphrasing, it is advisable to enhance the evaluation metrics by incorporating external resources like MAUVE (Pillutla et al. 2021). MAUVE considers the semantic meaning of sentences, providing a more comprehensive assessment about fluency beyond strict exact matching.
The evaluation of generation mainly relies on perplexity. A low perplexity indicates that the model is more confident in its prediction, that implying a better generation. Moreover, the information of generation is reflected by the sentence length and the normalized mean value of latent z. The average BLEU score of one sentence referring to others is also calculated. The lower, the better, as this means sentences vary from each other.

2.2.4 Parallel Processing

DMLP allows training over multiple GPUs. The parallel processing is achieved by torch distribution. Therefore DMLP is able to handle large datasets. The drawback is the training speed is not as fast as expected. This is a direction for future improvement.

Table 1: Reconstruction Task Output Examples

Input Sentences	Reconstructed Sentences
ever since joes has changed hands it 's just gotten worse and worse .	since then it 's been horrible and they have never been better .
i 've never had a worse experience than this !	i 've never had such a bad experience !
when i arrived , no one was at the desk .	when i arrived , no one was there .
she did an amazing job on my color and my cut !	she did an amazing job on my hair and color !
will definitely go back and recommend to friends .	will definitely go back and recommend to friends .

2.2.5 Compatibility

The experiments were done on RTX 3090 and RTX A6000. Since pre-trained large language models are used for encoder and decoder, the fine-tuning of LLMs requires enough memory to load BERT and GPT. It is recommended to check the available GPU space before moving forward.

3 Results

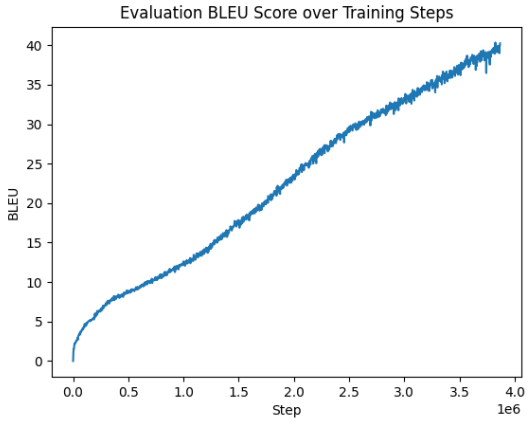
Two main experiments were performed, sentence reconstruction and setence generation. The training and evaluation are based on the Yelp Review dataset with about 179K negative and 268K positive sentences in the training dataset and 1K sentences in the evaluation dataset (Shen et al. 2017).

3.1 Reconstruction

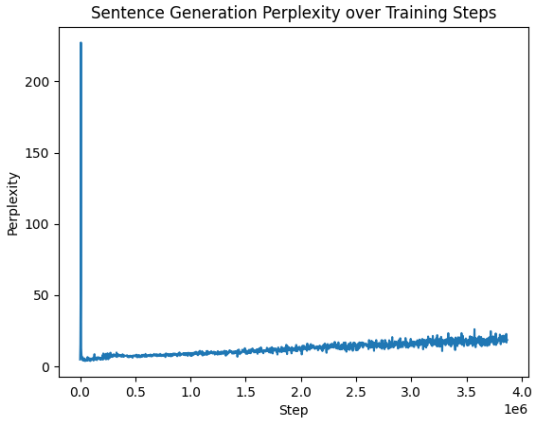
The quality of reconstruction was evaluated by corpus level BLEU score and MAUVE (Pillutla et al. 2021). On the evaluation dataset, BLEU score achieved approximately 40 by the time this report is done. Figure. 3(a) indicates the increasing trace of BLEU score over iterations, and it was still not converging. Therefore the final BLEU score will be higher. MAUVE enriches the evaluation dimension by measuring reconstruction outputs from the perspective of semantic meanings. The model achieved a MAUVE score of 0.767. Table 1 shows some output example. The result can be improved as the training process continues.

3.2 Generation

To train the DDPM model, MLPSkipNet was chosen to calculate the approximation of noise. Other possible choices are provided in DMLP as well. The generation was evaluated by perplexity. Figure. 3(b) shows the trend of perplexity over training. It is observed that the perplexity is decreasing first and slowing creeping upward. Although it ended at approximately 22, indicating a fairly close to the checkpoint result in the original JEDI paper (Liu et al. 2023b). VAEs has a trade-off between the quality of reconstruction and generation (Bowman et al. 2016), the upward trend perplexity exists concurrently with a larger



(a) Evaluation BLEU score over time



(b) Generation Perplexity over time

Figure 3: Reconstruction and Generation Evaluation

increase in BLEU score. Therefore it is likely this trade-off leads to slight increase in the perplexity figure. Table 2 shows some output examples.

Table 2: DMLP generation examples

Generation Example Ouptuts
the free drinks are only a few dollars more and you can even get free wifi .
the only thing that impressed was their customer service .
the sandwich here is very good and the price is reasonable .
no one was in the room for me , and i was sick .
now this is their shame on their branding .

4 Discussion

Based on the results, DMLP successfully eases the difficulty of applying the VAE-DDPM architecture while maintaining a high level performance. It can reduce the workload for research benchmark and task transfer on new dataset. However, it is noticed that DMLP requires enough GPU memory to load and fine-tune LLMs and DDPM. More optimization can be done on memory usage and parallel processing. The quality of reconstruction can be further improved if more time is given for training, suggested by the non-converging BLEU score figure. In addition, during experiments, it is observed that batch size has a significant effect on the loss converging speed. Surprisingly a small batch size helps the DDPM model converges faster.

5 Conclusion

Using DMLP, a model with high generation and reconstruction quality is constructed. Applying diffusion models in text processing demonstrated promising capabilities in generating human-like text. This success opens the door to various future applications of the VAE-DDPM model, especially when combined with other methodologies. DMLP provides researchers with the essential components needed for exploring and advancing the capabilities of text diffusion models. It eases the research process by releasing researchers from the burden of coding, allowing them to focus on the development of their methods.

References

Bowman, Samuel R., Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. "Generating Sentences from a Continuous Space." In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. Berlin, Germany Association for Computational Linguistics. [Link]

Chen, Xi, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. "InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets."

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding."

Gong, Shansan, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2023. "DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models."

Ho, Jonathan, Ajay Jain, and Pieter Abbeel. 2020. "Denoising Diffusion Probabilistic Models."

Kingma, Diederik P., Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. 2017. "Improving Variational Inference with Inverse Autoregressive Flow."

Kingma, Diederik P., and Max Welling. 2019. "An Introduction to Variational Autoencoders." *Foundations and Trends® in Machine Learning* 12(4), p. 307–392. [Link]

Li, Xiang Lisa, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. 2022. "Diffusion-LM Improves Controllable Text Generation."

Liu, Guangyi, Zeyu Feng, Yuan Gao, Zichao Yang, Xiaodan Liang, Junwei Bao, Xiaodong He, Shuguang Cui, Zhen Li, and Zhiting Hu. 2023a. "Composable Text Controls in Latent Space with ODEs."

Liu, Guangyi, Yu Wang, Zeyu Feng, Liping Tang, Yuan Gao, Zhen Li, Shuguang Cui, Eric Xing, Zichao Yang, and Zhiting Hu. 2023b. "Generation, Reconstruction, Representation All-in-One: A Joint Autoencoding Diffusion Model." In *Submitted to The Twelfth International Conference on Learning Representations*. [Link]

Pillutla, Krishna, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck,

Yejin Choi, and Zaid Harchaoui. 2021. “MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers.” In *NeurIPS*.

Razavi, Ali, Aaron van den Oord, and Oriol Vinyals. 2019. “Generating Diverse High-Fidelity Images with VQ-VAE-2.”

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. 2015. “U-net: Convolutional networks for biomedical image segmentation.” In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer

Shen, Tianxiao, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. “Style Transfer from Non-Parallel Text by Cross-Alignment.”

Sohl-Dickstein, Jascha, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. “Deep Unsupervised Learning using Nonequilibrium Thermodynamics.”

Appendices

A.1 DMLP Usage Pipeline

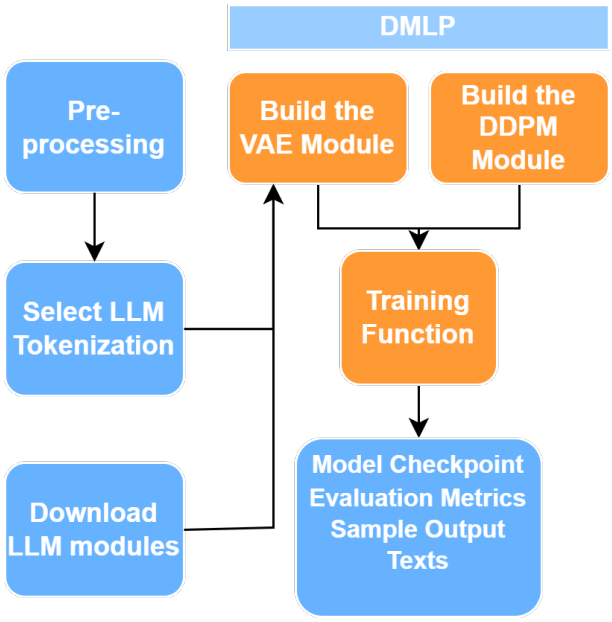


Figure A 1: DMLP Usage Pipeline Visualization