DiffusionBERT: Improving Generative Masked Language Models with Diffusion Models

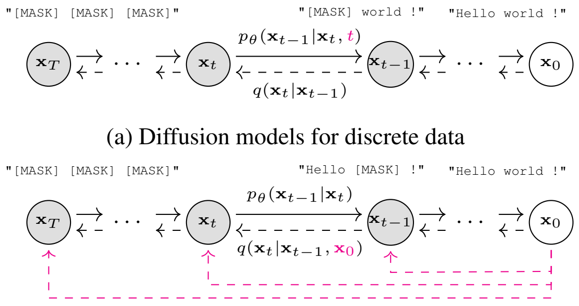
Zhengfu He* Tianxiang Sun* Qiong Tang Kuanning Wang
Xuanjing Huang Xipeng Qiu†
School of Computer Science, Fudan University
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
{zfhe19,txsun19,wangkn20,xjhuang,xpqiu}@fudan.edu.cn
qtang22@m.fudan.edu.cn

Abstract

We present DiffusionBERT, a new generative masked language model based on discrete diffusion models. Diffusion models and many pre-trained language models have a shared training objective, i.e., *denoising*, making it possible to combine the two powerful models and enjoy the best of both worlds. On the one hand, diffusion models offer a promising training strategy that helps improve the generation quality. On the other hand, pre-trained denoising language models (e.g., BERT) can be used as a good initialization that accelerates convergence. We explore training BERT to learn the reverse process of a discrete diffusion process with an absorbing state and elucidate several designs to improve it. First, we propose a new noise schedule for the forward diffusion process that controls the degree of noise added at each step based on the information of each token. Second, we investigate several designs of incorporating the time step into BERT. Experiments on unconditional text generation demonstrate that DiffusionBERT achieves significant improvement over existing diffusion models for text (e.g., D3PM and Diffusion-LM) and previous generative masked language models in terms of perplexity and BLEU score. Promising results in conditional generation tasks show that DiffusionBERT can generate texts of comparable quality and more diverse than a series of established baselines.

1 Introduction

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021) have recently emerged as a new class of state-of-the-art generative models, achieving high-quality synthesis results on image data (Ramesh et al., 2022; Rombach et al., 2022; Saharia et al., 2022). Though these models captured widespread attention from not only the research community but also the public, applying diffusion models to text data is still challenging and under-explored due to the discrete nature of the text. A few prior works that explored using diffusion models on text data can be divided into two lines. The first is to extend diffusion models to discrete state spaces (Hoogeboom et al., 2021; Austin et al., 2021). The second is to perform the diffusion process and its reverse process in the continuous domain and bridge the continuous and the discrete domain through embedding and rounding (Li et al., 2022; Gong et al., 2022). However, none of these works leveraged pre-trained language models (PLMs, Devlin et al. (2019); Lewis et al. (2020); Raffel et al. (2020); Brown et al. (2020); Qiu et al. (2020)), which are an unmissable treasure in the NLP community.



(a) Diffusion models for discrete data

(b) Non-Markovian DiffusionBERT

Figure 1: In contrast to conventional discrete diffusion models, DiffusionBERT uses BERT as its backbone to perform text generation. The main differences are highlighted in color: (1) DiffusionBERT performs decoding without knowing the current time step while canonical diffusion models are conditioned on time step. (2) The diffusion process of DiffusionBERT is non-Markovian in that it generates noise samples $\mathbf{x}_t$ conditioning not only on $\mathbf{x}_{t-1}$ but also on $\mathbf{x}_0$. Such a non-Markov process is due to our proposed noise schedule.

This work, to our knowledge, is the first attempt to combine diffusion models with PLMs. Such a combination is built upon a shared training ob-

[*] Equal contribution.
[†] Corresponding author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4521–4534
July 9-14, 2023 ©2023 Association for Computational Linguistics

Multitask Pre-training of Modular Prompt for Chinese Few-Shot Learning

Tianxiang Sun[*] **Zhengfu He**[*] **Qin Zhu** **Xipeng Qiu**[†] **Xuanjing Huang**
School of Computer Science, Fudan University
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
{txsun19,zfhe19,xpqiu,xjhuang}@fudan.edu.cn zhuq22@m.fudan.edu.cn

Abstract

Prompt tuning is a parameter-efficient approach to adapting pre-trained language models to downstream tasks. Although prompt tuning has been shown to match the performance of full model tuning when training data is sufficient, it tends to struggle in few-shot learning settings. In this paper, we present **M**ulti-task **P**re-trained **M**odular **P**rompt (**MP^2**) to boost prompt tuning for few-shot learning. MP^2 is a set of combinable prompts pre-trained on 38 Chinese tasks. On downstream tasks, the pre-trained prompts are selectively activated and combined, leading to strong compositional generalization to unseen tasks. To bridge the gap between pre-training and fine-tuning, we formulate upstream and downstream tasks into a unified machine reading comprehension task. Extensive experiments under two learning paradigms, i.e., gradient descent and black-box tuning, show that MP^2 significantly outperforms prompt tuning, full model tuning, and prior prompt pre-training methods in few-shot settings. In addition, we demonstrate that MP^2 can achieve surprisingly fast and strong adaptation to downstream tasks by merely learning 8 parameters to combine the pre-trained modular prompts.

1 Introduction

Pre-trained models (PTMs) (Devlin et al., 2019; Lewis et al., 2020; Raffel et al., 2020; Qiu et al., 2020) with prompt-based learning have achieved remarkable progress in few-shot learning. A major reason behind their success is the closed gap between upstream pre-training and downstream fine-tuning (Liu et al., 2021a; Sun et al., 2022b). Since the downstream tasks are reformulated into a unified (masked) language modeling ((M)LM for short) task, one can reuse the pre-trained (M)LM head instead of training a randomly initialized classification head to solve tasks with limited data. However, prompt-based learning (e.g., PET (Schick and Schütze, 2021) and LM-BFF (Gao et al., 2021)) usually fine-tunes all the parameters of the PTM for each downstream task, which can be computationally expensive and deployment-inefficient, especially for large PTMs such as GPT-3 (Brown et al., 2020).

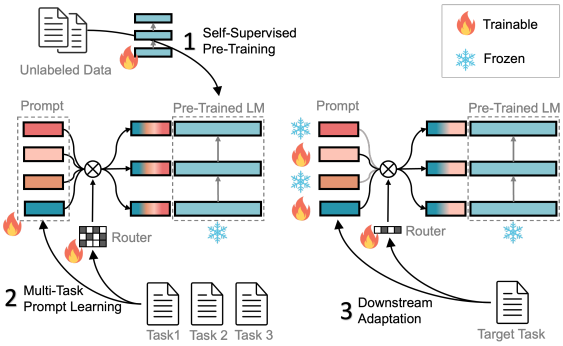


Figure 1: MP^2 achieves fast adaptation to downstream tasks through three steps: (1) Self-supervised pre-training on large-scale unlabeled data. (2) Pre-training modular prompts and the corresponding router with multi-task learning. (3) A subset of prompts is activated and tuned for adaptation to downstream tasks.

Recently, much effort has been devoted to parameter-efficient prompt tuning (Li and Liang, 2021; Lester et al., 2021; Liu et al., 2021c; Sun et al., 2022c), which only learns a small number of soft prompt parameters while keeping the main body of the PTM untouched. In contrast to full model tuning, prompt tuning can get specialized models for specific tasks by simply attaching task-specific prompts, and therefore is highly efficient for serving different tasks. Though it has been demonstrated that prompt tuning can match the performance of full model tuning when training data is sufficient (Lester et al., 2021), the soft prompt cannot be well trained from scratch in few-shot learning settings (Gu et al., 2021) because the randomly initialized soft prompt introduces a new gap between pre-training and fine-tuning.

[*] Equal contribution.
[†] Corresponding author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 11156–11172
July 9-14, 2023 ©2023 Association for Computational Linguistics