Diffusion 모델에서의 효율적인 파라미터 튜닝을 위한 기법 비교

김은지, 안재신, 정희철
경북대학교
duqtj00@knu.ac.kr, asj0420@knu.ac.kr, heechul@knu.ac.kr

Comparative Study of Efficient Parameter Tuning Techniques in Diffusion Models

EunJi Kim, Jaesin Ahn, Heechul Jung
Kyungpook National University

요 약

딥러닝 기반 생성 모델의 발전으로 생성 이미지의 품질과 다양성이 크게 향상되었다. 그중에서도 Diffusion 모델은 딥러닝 기술을 활용하여 주어진 텍스트 설명이나 이미지를 조건으로 받아들여 이를 시각적으로 표현하는 인공지능 기술로 주목받고 있다. 그러나 대규모 Diffusion 모델의 모든 파라미터를 학습하는 것은 메모리와 시간 측면에서 비효율적이다. 이를 해결하기 위해, 본 논문에서는 Diffusion 모델의 파라미터를 효율적으로 튜닝하기 위한 다양한 방법을 실험하고 비교 분석하였다. 특히, Adapter, LoRA, Prefix, BitFit과 같은 효율적인 파라미터 튜닝 기법을 diffusion 생성 모델에 적용하였습니다. 실험은 Food101 데이터셋에서 FID 점수로 비교하였으며, full fine-tuning 방식과 비교하여 성능이 향상된 것을 확인할 수 있었다.

Ⅰ. 서 론

딥러닝 기반 생성 모델 중 하나인 Diffusion 모델은 최근 이미지 생성 분야에서 큰 주목을 받고 있다. Diffusion 모델은 Transformer 기반 아키텍처를 활용하여 텍스트 입력을 시각적 표현으로 변환하는 역할을 한다. 이 모델은 forward process 단계에서 이미지를 점차적으로 노이즈가 포함된 상태로 변환한 후, reverse process 단계에서 주어진 텍스트 조건을 바탕으로 노이즈를 제거하여 최종적으로 텍스트와 일치하는 이미지를 복원한다. 특히, reverse process를 학습함으로써 노이즈에서 이미지를 복원하는 과정을 학습하게 된다. 이렇게 Diffusion 모델은 조건부 생성 과정에서 텍스트 정보를 효과적으로 활용하여 다양한 시각적 콘텐츠를 생성하는 능력을 지니고 있다. 그러나 Latent Diffusion Model(LDM) [1]과 같은 대규모 Diffusion 모델은 약 10억 개에 달하는 파라미터를 가지고 있어 전체 파라미터를 학습하는 것은 굉장히 비효율적이다. 따라서 이러한 모델의 일부 파라미터만 학습하거나 소수의 추가 파라미터만을 학습하도록 함으로써 모델을 보다 효율적으로 학습하기 위한 다양한 연구들이 진행되고 있다.

대표적으로 Adapter [2]는 Adapter 모듈을 삽입하여 기존 모델의 파라미터는 고정하고 Adapter의 파라미터만 학습함으로써 모델의 표현력을 유지하면서도 학습해야 할 파라미터 수를 줄였다. 또한, LoRA [3] (Low-Rank Adaptation)는 고차원 파라미터를 저차원 행렬로 분해하고, 이 저차원 행렬만을 학습해 효율성을 높였다. Prefix Tuning [4] 기법은 모델 입력 앞부분에 고정된 벡터를 추가하고, 이 prefix 벡터만 학습하여 전체 모델을 학습하지 않더라도 모델의 성능이 향상되었다. BitFit [5]은 모델의 bias를 제외한 파라미터는 고정하고 bias만을 학습하는 방법으로 효율적인 파라미터 튜닝을 수행하였다.

본 연구에서는 이러한 선행 연구들을 사전 학습된 Diffusion 모델에 적용했을 때 각 기술의 효율성 및 효과성을 비교 분석하였다. Diffusion 모델에 Adapter, LoRA, Prefix, BitFit과 같은 효율적인 파라미터 튜닝 기법을 적용했을 때 모델의 text-to-image 성능을 비교 평가하였으며, 이때 각 기법의 효율성을 평가하기 위해 학습 파라미터 수, 학습 시간 또한 비교 평가하였다. 실험 결과, ## 기법이 상대적으로 높은 성능을 보였으며, BitFit이 가장 적은 학습 파라미터 수를 요구하는 것으로 나타났다.

Ⅱ. 방 법

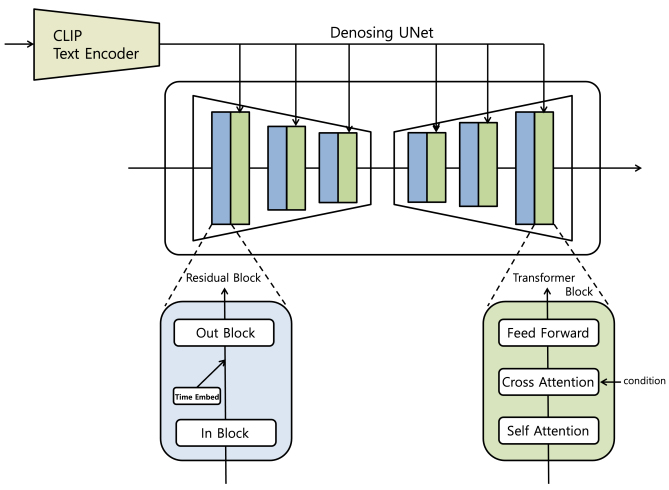

그림1. Diffusion 모델의 구조 [1, 8]

그림1은 LDM의 일부를 나타낸 것으로, 본 논문에서는 그 중 U-Net 모듈을 학습 대상으로 삼아 모델의 성능을 최적화하는 것에 중점을 두고 있다. 특히, U-Net의 Transformer Block에 학습 가능한 파라미터를 추가하여 노이즈 제거와 이미지 생성 성능을 극대화하는 방법을 제안하고, 이를 U-Net 전체를 학습하는 전통적인 full fine-tuning 방법과 비교하였다.

구체적으로, Adapter는 그림2와 같이 Transformer Block 내 Feed Forward 레이어 이후에 학습 가능한 파라미터를 추가하여 모델을 미세 조정하였다. 이 방법은 기존 구조에 크게 영향을 주지 않으면서도 성능 향상을 도모할 수 있다는 장점을 가지고 있다. LoRA와 prefix-tuning은 Transformer Block의 Cross Attention 부분에 파라미터를 추가하였다. UniPELT[6]와 같이 LoRA는 query와 key 부분에 학습 가능한 파라미터를 삽입하여 효율적인 조정을 하고 prefix-tuning은 key와 value의 앞부분에 prefix를 추가하여 학습을 진행하였다. 이를 통해, 기존의 Attention 메커니즘에 변형을 가하지 않고도 성능을 높일 수 있다.

마지막으로, BitFit은 앞서 설명한 기법들과 달리 새로운 파라미터를 추

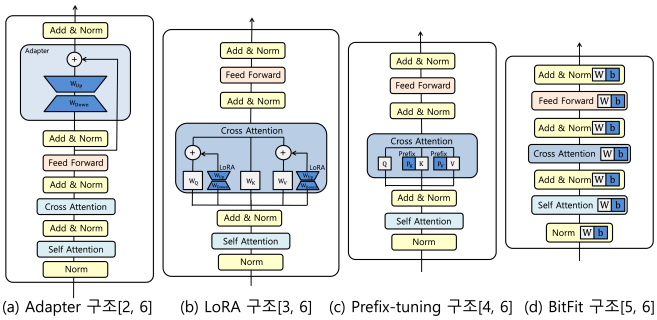

그림2. (a) Adapter 구조[2, 6] (b) LoRA 구조[3, 6] (c) Prefix-tuning 구조[4, 6] (d) BitFit 구조[5, 6]

가하지 않고, U-Net의 Transformer Block에서 기존의 bias만을 학습하여 모델을 미세 조정하였다. 이는 매우 경량화된 방식으로, 추가적인 메모리나 계산 자원 없이도 일정 수준의 성능 향상을 기대할 수 있는 방법이다. 이러한 다양한 기법들을 활용한 실험을 통해 본 논문에서는 LDM 기반 U-Net의 성능을 향상시키고, 다양한 데이터셋에 적합한 미세 조정 가능성을 탐구하였다.

Ⅲ. 실 험 결 과

Method	Food101
Full fine tuning	43.33
Adapter	47.41
LoRA	35.68
Prefix-tuning	32.57
BitFit	68.53

표1. 방법에 따른 FID 점수 비교

실험에 사용한 하드웨어 환경은 GPU RTX A6000으로, 총 10000스텝 학습을 하였으며, 학습률(learning rate)은 1e-5, 옵티마이저는 adamw(bf16)을 사용하였다. 데이터 셋은 Food101 [7]을 사용하여 학습을 진행하였다. 학습 모델은 stable diffusion v1.4를 사용하였으며 adapter, lora, prefix, bitfit을 적용하였다. 표1은 Food101의 FID점수의 결과를 나타낸다. FID는 생성된 이미지의 품질을 평가하는 데 사용되는 metric으로, 생성된 이미지와 실제 이미지 간의 분포 차이를 측정한다. 이 연구에서는 모델이 생성한 이미지의 다양성과 품질을 정량적으로 평가하기 위해 FID를 사용했다. FID 값이 낮을수록 생성된 이미지가 실제 이미지와 더 유사하다는 것을 의미하므로, 모델의 성능 향상을 평가하는 데 유용한 지표이다.

실험 결과, Prefix-tuning이 FID 32.57로 가장 좋은 성능을 보였으며, 이는 해당 기법이 모델의 성능을 가장 효율적으로 향상시켰다는 것을 의미한다. 반면, BitFit은 FID 68.53로 가장 낮은 성능을 기록했으며, 이는 다른 기법들에 비해 이미지 품질에서 큰 차이가 발생했음을 나타낸다. LoRA와 Adapter는 각각 35.68, 47.41로 비교적 우수한 성능을 보였으며, 특히 LoRA는 Full Fine-tuning(FID 43.33)보다도 더 적은 파라미터를 사용하면서도 성능이 더 좋았다.

이러한 결과는, 파라미터 효율적 미세 조정 방법들이 적은 파라미터로도 모델의 성능을 크게 저하시키지 않으면서 특정 기법, 특히 Prefix-tuning은 실제 Fine-tuning보다도 더 좋은 성능을 낼 수 있다는 가능성을 시사한다.

Method	# Param	$Time_T$
Full fine tuning	860M(100%)	100%
Adapter	1.23M(0.143%)	56%
LoRA	0.23M(0.029%)	42%
Prefix-tuning	0.25M(0.030%)	44%
BitFit	0.17M(0.020%)	39%

표2. 파라미터 수와 학습 시간의 상대적 비교

표 2는 각 방법의 파라미터 수와 학습 시간을 비교한 결과를 보여준다. 이 표에서 확인할 수 있듯이, 전체 모델을 완전히 미세 조정하는 것에 비해 학습 시간이 약 절반으로 감소하였다. 이러한 파라미터 효율적인 미세 조정 기법은 특정 데이터셋에 맞춰 성능을 향상시킬 수 있는 잠재력을 지니고 있으며, 향후 다양한 연구 및 실제 응용에서도 매우 유용하게 활용될 것으로 기대된다.

Ⅳ. 결 론

본 연구에서는 Parameter-Efficient Fine-Tuning 기법을 Diffusion 모델에 적용하여, 전체 파라미터를 학습하지 않으면서도 성능 저하 없이 모델을 효과적으로 학습할 수 있음을 확인하였다. 특히, Adapter, LoRA, Prefix, BitFit 기법이 Full Fine-Tuning에 비해 메모리와 계산 효율성을 크게 개선하면서도 우수한 성능을 유지함을 입증하였다. 이 결과는 대규모 모델의 효율적 학습을 위한 새로운 가능성을 제시하며, 향후 다양한 생성 모델 및 응용 분야에서 더욱 효율적인 학습 방법론 개발에 기여할 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

본 연구는 2021년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(2021R1A6A1A03043144)이며, 부분적으로 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업의 지원을 받아 수행되었음 (IITP-2024-2020-0-01808).

참 고 문 헌

[1] R. Rombach, A. Blattmann, D. Lorenz, Esser, and B. Ommer, "High resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* 2022, pp. 10684-10695.

[2] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp,"in *International conference on machine learning*. PMLR, 2019, pp. 2790-2799.

[3] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "Lora: Low-rank adaptation of large languague models," *arXiv preprint arXiv:2106.09685,* 2021.

[4] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," *arXiv preprint arXiv:2101.00190,* 2021.

[5] E. B. Zaken, S. Ravfogel, and Y. Goldberg, "BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models," *arXiv preprint arXiv:2106.10199,* 2021.

[6] Y. Mao, L. Mathias, R. Hou, A. Almahairi, H. Ma, J. Han, W.-t. Yih, and M. Khabsa, "Unipelt: A unified framework for parameter-efficient language model tuning," *arXiv preprint arXiv:2110.07577,* 2021.

[7] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101-mining discriminative components with random forests," in *Computer vision-ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13.* Springer, 2014, pp. 446-461

[8] C. xiang, F. Bao, C. Li, H. Su, and J. Zhu, "A closer look at parameter-efficient tuning in diffusion models," *arXiv preprint arXiv:2303.18181,* 2023.