# 다중 LoRA 결합 개선 방법에 대한 연구

## Improved Approaches to Merge Multiple LoRAs


**요 약**

최근 이미지 생성 분야에 있어 DALL-E-2, Stable diffusion과 같은 Diffusion 기반 생성 모델 등이 뛰어난 성능을 보이며 이미지 생성에 대한 파장을 일으켰다. 하지만 이러한 모델들을 추가 학습하기 위해서는 많은 시간과 큰 연산 자원이 필요로 하며, 새로운 정보 학습이 어렵다는 단점을 안고 있다. 그러한 문제를 해결하기 위해서 Low-Rank Adaptation(LoRA)를 차용하지만, 한 번에 많은 LoRA를 적용 시 이미지 퀄리티 저하, 구성 요소가 모호해지는 문제가 존재한다. 본 논문에서는 해당 문제를 해결하기 위해 가중치 조작과 연산 방식을 변화하여 해결하고자 한다.


## 1. 서 론

DALL · E-2[1], Stable diffusion[2] 등의 여러 Diffusion[3] 기반 모델들이 놀라운 성능을 보여줬다. 그러나 기존 모델들은 큰 크기와 높은 학습 자원과 시간을 소모하는 단점이 있다. 이러한 문제를 해결하기 위해 차용한 LoRA[4]는 적은 자원으로도 원하는 결과물을 학습할 수 있는 점으로 인해 현재 많은 커뮤니티에서 사용되고 있다. 그러나 기존 다중 LoRA 결합 방법인 LoRA-Merge는 여러 LoRA를 동시에 적용하는 방식으로 작동하기에 이미지 품질[Image Quality or Image]과 LoRA 조합의 일관성[Composition Quality]을 해친다. 해결책으로 LoRA-Switch[5]와 LoRA-Composition[5]가 제안되었다. 본 논문에서는 관련 연구에서 제안된 LoRA-Switch[5], LoRA-composition[5]와 다른 접근법으로 기존의 LoRA-Merge의 문제를 해결하기 위한 연구들에 대한 내용을 다룬다.

## 2. 관련 연구

### 2.1 LoRA-Merge

$$W' = W + \sum_{i=1}^{N} w_i \times B_i A_i \tag{1}$$

식 (1)는 일반적으로 LoRA들을 결합하는 방법이다. 전체 LoRA 개수인 N까지의 Low Rank 행렬들인 $B_i$와 $A_i$들을 곱하여 행렬 $W_i$들을 만들어 각각 더하고 원본 모델의 $W$에 더한다.

### 2.2 LoRA-HUB[6]

$$w' = (\alpha_1 \times A_1 + ... + \alpha_N \times A_N)(\alpha_1 \times B_1 + ... + \alpha_N \times B_N) \tag{2}$$

식 (2)는 LM에 각각 다른 task에 맞는 LoRA들을 적용해 $\alpha_i$를 학습함으로써 모델이 다중 task에 적응함과 동시에 성능을 올리는 방법론이다. i번째 LoRA인 $(A_i)(B_i)$을 스케일링 해주는 $\alpha_i$를 학습한다. $\alpha_i$는 LoRA의 표현력을 스케일링한다.

### 2.3 LoRA-Composition

$$\tilde{\epsilon}(\mathbf{z}_t, c) = \frac{1}{N} \sum_{i=1}^{N} \alpha_i \times [\epsilon_{\theta'_i}(\mathbf{z}_t) + s \times (\epsilon_{\theta'_i}(\mathbf{z}_t, c) - \epsilon_{\theta'_i}(\mathbf{z}_t))] \tag{3}$$

식 (3)에서 $\alpha_i$는 LoRA에 적용되는 하이퍼파라미터다. $\epsilon_{\theta'_i}$은 LoRA i번째를 적용한 노이즈다. Classifier-free diffusion guidance[7]의 식을 인용하여 Condition $c$가 적용된 노이즈와 아닌 노이즈들을 이용하여 균등하고 퀄리티를 보존하며 결합한다.

### 2.4 LoRA-Switch

$$q = \lfloor ((t-1) mod(N\chi))/\chi \rfloor + 1 \tag{4}$$

식 (4)에서 $q$는 현재 time step에서 LoRA의 index다. $\chi$는 하이퍼파라미터이며 $N$은 전체 LoRA의 수이다. 식(4)는 LoRA를 $N\chi$마다 번갈아 가면서 적용해, 이미지 퀄리티를 보존하면서 균등하게 LoRA를 적용하는 방법이다.

## 3. 제안 방법

### 3.1 MERGEV2

일반적으로, 대부분의 식 (1)는 모델 내부에서 연산이 진행된다. MERGEV2는 식 (2)에서 영감을 얻어 모델 내부에서가 아닌 외부에서 Merge를 수행하여 속도와 품질을 개선하고자 한다.

$$B'A' = \frac{w'}{\beta * N} \tag{5}$$

$w'$는 식 (2)으로 구하며 Merge 된 새로운 LoRA 가중치다. 식 (5)에서 $w'$을 구할 때 $\alpha_i$는 식 (2)에서 학습되는 것과 다르게 하이퍼파라미터다. $w'$를 그대로 모델에 사용할 경우, Guidance 가중치가 과도하게 커져 이미지가 검은색으로 출력되는 단점이 있다. 이 문제를 해결하기 위해, 총 LoRA의 수 $N$로 정규화가 필요하다. 그러나 $N$으로 정규화할 경우, 각 LoRA의 가중치가 과도하게 감소하는 문제를 해결하기 위해 하이퍼파라미터인 $\beta$를 곱하여 보정해주었다.

### 3.2 S-Composition

LoRA-Switch는 LoRA를 교체하면서 적용하므로 계산 시간은 적으며 LoRA-Composition은 적은 Denoising steps에서도 성능이 좋은 경향이 보였다. 식 (3)의 주요 단점은 모든 LoRA를 적용할 때 발생하는 계산 시간의 증가이며 LoRA-Switch은 많은 LoRA

를 적용할 때 좋은 성능을 위해서 높은 Denoising steps가 요구되는 경향이 있다. 이는 일반적인 Merge보다 훨씬 더 많은 자원이 필요로 한다. 이러한 두 가지 방법의 단점을 보완하기 위해 S-Composition을 제안한다. 이 방법은 식 (3)와 식 (4)를 함께 적용하여 LoRA-Composition에서 발생하는 계산 비용을 LoRA-Switch로 줄이면서도 성능을 유지한다.

$$k = \max(N - x, 1) \tag{6}$$

$$q = \lfloor ((t-1)) mod( \binom{N}{k} \chi)) / \chi \rfloor + 1 \tag{7}$$

식 (4)와 다르게 q는 다음 time step에서 적용될 LoRA들이 있는 배열의 인덱스 선택하는데 사용된다. k는 총 LoRA 개수(N)에서 매 timestep마다 적용될 LoRA의 수다. 해당 값은 N에서 하이퍼파라미터 x를 뺀 값으로 구한다. x는 $1 \leq x \leq N$을 만족시켜야 한다. k개의 LoRA를 가진 리스트는 총 $\binom{N}{k}$만큼 만들어지며 매 time step마다 q번째 리스트 안에 있는 k개의 LoRA들을 식 (3)으로 계산한다.

### 3.3 SUB-O

많은 LoRA에서의 품질 저하는 한 번에 많은 LoRA를 식 (1)에 적용해 일어난다. 기존의 방법들은 해당 문제를 해결하는 과정에서 LoRA를 적용하지 않은 원본 모델의 분포와 노이즈를 배제한다. 이에 따라 LoRA를 변경하면서 적용하거나 개별적으로 값을 계산하여 적용하는 방법 즉 각각의 LoRA를 분리하여 실제로 적용되지 않은 원본 노이즈와의 차이를 계산하고, 이를 원본 노이즈에 적용하는 방법을 소개한다.

$$\tilde{\epsilon}(\mathbf{z}_\lambda, c) = \epsilon_\theta(\mathbf{z}_\lambda) + \gamma \cdot (\epsilon_\theta(\mathbf{z}_\lambda, c) - \epsilon_\theta(\mathbf{z}_\lambda)) \tag{8}$$

$$\tilde{\epsilon}(\mathbf{z}_\lambda, c)_i = \alpha_i \times [\epsilon_{\theta'_i}(\mathbf{z}_\lambda) + \gamma \cdot (\epsilon_{\theta'_i}(\mathbf{z}_\lambda, c) - \epsilon_{\theta'_i}(\mathbf{z}_\lambda))] \tag{9}$$

먼저, $\epsilon_\theta$는 LoRA가 적용되지 않은 원본 노이즈를 나타낸다. $\epsilon_{\theta'_i}$은 N개의 LoRA 중에서 i번째가 적용된 노이즈로 간주한다. 각각의 노이즈에 대해 LoRA에 적용된 하이퍼파라미터 값인 $\alpha_i$를 곱하고 최종 노이즈를 구한다. 해당 노이즈들을 기반으로 식 (10)을 계산한다.

$$\tilde{\epsilon}(\mathbf{z}_\lambda, c) = \eta * \tilde{\epsilon}(\mathbf{z}_\lambda, c) + (1 - \eta) \sum_{i=1}^{N} \tilde{\epsilon}(\mathbf{z}_\lambda, c) - \tilde{\epsilon}(\mathbf{z}_\lambda, c)_i \tag{10}$$

식 (10)에서 $\eta$는 하이퍼파라미터로서, 원본 노이즈와 LoRA들의 합으로 만들어진 노이즈들 사이에서 중요도를 설정한다. 이 값이 높을수록 원본 노이즈가 작을 때 LoRA들이 적용된 노이즈가 최종 노이즈에 높은 가중치로 반영된다. 따라서 작은 $\eta$ 값은 LoRA들의 영향력을 강조하고, 높은 $\eta$ 값은 원본 노이즈를 더 강하게 가져간다.

## 4. 실험 조건

NVIDIA Corporation GM200 [GeForce GTX TITAN X] 12GB GPU를 사용하여, huggingface에 존재하는 2D체 모델 richyrichmix-v2와 현실체 모델 epiCPhotoGas를 활용했다. 자원의 제약으로 인해 LoRA 2개와 LoRA 3개만을 적용했다. 2D체의 denoise step은 100이며 cfg scale은 9이고, 현실체는 150,7이다. SUB-O의 $\eta$는 0.2, S-Composition의 $\chi$는 1, MERGEV2의 $\alpha$는 0.75이다. 각 방법론에서 2D체는 20개씩(LoRA 2 10개, LoRA 3 10개), 현실체는 20(LoRA 2 10개, LoRA 3 10개)개씩으로 총 240개의 이미지를 random sampling을 통해 선택했다. 이전 연구(Multi-lora composition for image generation[5])의 평가 방식을 차용하면서도 멀티모델인 GPT-4에 더해서 Gemini를 추가로 채택하여 이미지 평가를 수행했다. 이미지 평가는 이미지를 2개씩 비교하는 방식을 사용했다. 이전 연구에서처럼, LoRA-Merge 방식의 대안으로 제안된 MERGEV2와 SUB-O로 생성된 이미지들은 LoRA-Merge로 생성된 이미지들과 비교하였다. 또한, LoRA-Composition와 LoRA-Switch의 대안으로 제시된 S-Composition은 LoRA-Composition와 LoRA-Switch로 생성된 이미지들과 비교하였다. 사람 평가는 소수의 인원으로 블라인드 평가를 했기 때문에 노이즈가 있을 수 있다.

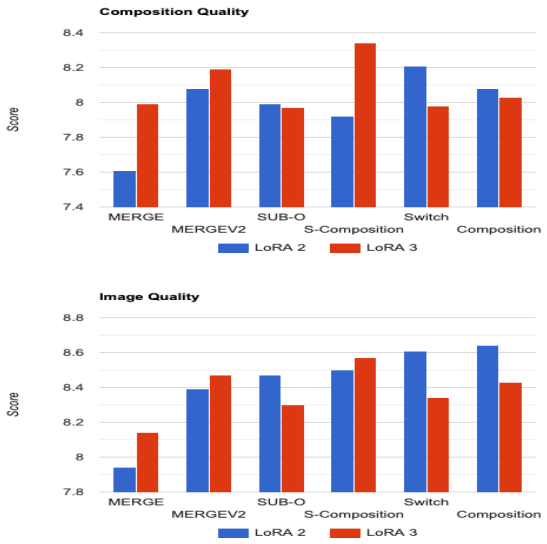


그림 1: Gemini 평가지표

## 5. 실험결과 및 분석

Mergev2는 기존 LoRA-Merge 방식보다 유의미한 성능 향상을 보였다. SUB-O는 LoRA-Merge보다 높은 이미지 퀄리티를 보여주었다. S-Composition은 LoRA-Switch와 LoRA-Composition과 사람 평가와 GPT4에서 유사한 성능을 보여주었고, Gemini에서는 높은 성능을 보였다.

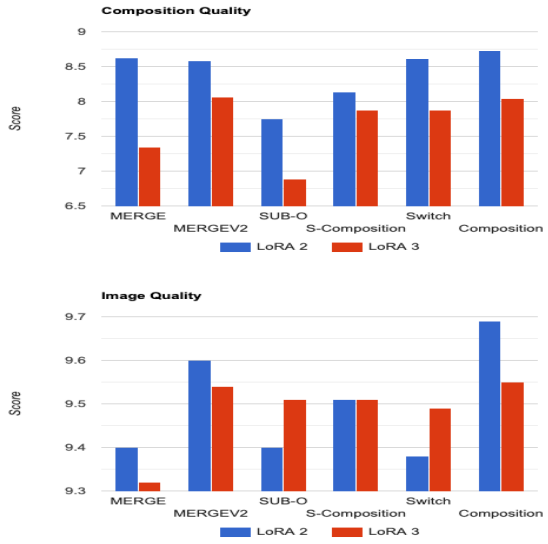


그림 2: GPT4 평가지표

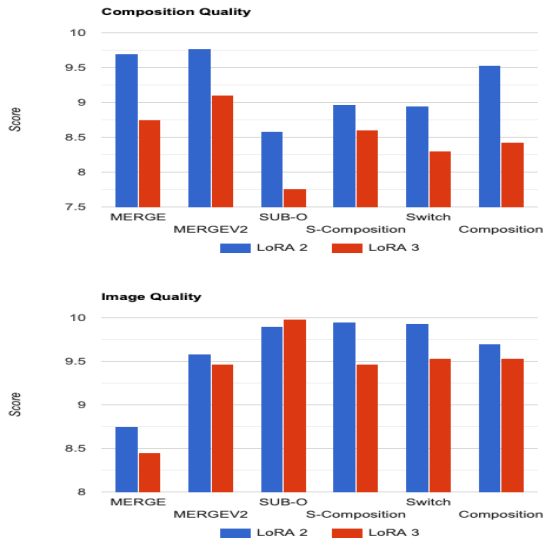


그림 3: 사람 평가지표

## 6. 결 론

표 1: 인간 평가와 멀티 모델 사이의 Pearson 상관 계수

| **Metrics** | LoRA-2 Pearson | LoRA-3 Pearson |
|---|---|---|
| GPT4 Composition | 0.202 | 0.415 |
| GPT4 Image Quality | -0.069 | 0.114 |
| Gemini Composition | 0.124 | 0.214 |
| Gemini Image Quality | 0.066 | 0.131 |

이 논문에서는 다중 LoRA의 핸들링에 대한 새로운 방법론들을 제안했다. MERGEV2, SUB-O, S-Composition 등의 방법들은 기존의 문제점을 완벽하게 해결하는 것보다는 새로운 방법을 제시하는 데 더 의의를 두었다. 특히 SUB-O와 같이 기존에 사용되지 않았던 방법들이 성능이 나온다는 점을 밝히면서 미래에 더욱 다양한 방법론에 영감이 되기를 기대한다.

표1에서 사람과 멀티모델 간의 평가의 상관관계를 Pearson 상관계수를 사용하여 수치화했다. 이를 통해 GPT-4와 Gemini의 평가들이 얼마나 사람과 일치하는지, 성능의 비교 우위를 파악하는 게 목표다. GPT-4가 Composition에 대해 Gemini보다 더 나은 상관관계를 보였지만, Image Quality 측면에서는 비슷하거나 상대적으로 낮은 상관관계를 보였다.

사람의 평가가 절대적으로 옳다고 가정하기는 어렵다. 사람의 평가는 주관적이며 평가 기준에 따라 편향될 수 있다. 멀티모델의 평가가 사람과 완벽하게 일치하지 않는다고 해도 모델의 평가를 참고할 수 없다는 것은 아니다.

## 7. 감사의 글

이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2021-II212068, 인공지능 혁신 허브 연구 개발)

## 참고 문헌

[1] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. 2022.

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. 2020.

[4] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[5] Ming Zhong, Yelong Shen, Shuohang Wang, Yadong Lu, Yizhu Jiao, Siru Ouyang, Donghan Yu, Jiawei Han, and Weizhu Chen. Multi-lora composition for image generation, 2024.

[6] Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition, 2024.

[7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.