

# 전이 학습을 위한 점진적 동결 해제 파인튜닝 기법에 대한 연구

정진우, 차예찬, 김장호\*

국민대학교 인공지능학부

imaboybut@kookmin.ac.kr, 3004yechan@kookmin.ac.kr, \*jangho.kim@kookmin.ac.kr

## A Study on Gradual Unfreeze Fine-Tuning for Transfer Learning

Jin Woo Chung, Ye Chan Cha, Jang Ho Kim\*

Department of Artificial Intelligence, Kookmin Univ.

### 요약

본 논문에서는 전이 학습의 연산·메모리 비용을 줄이기 위해 Gradual Unfreeze Fine-Tuning(GUFT) 프레임워크를 제안한다. GUFT는 파라미터의 그래디언트 절댓값을 기반으로 중요 파라미터를 예측하고, 중요 파라미터만 점진적으로 학습하여 모델의 안정성과 표현력을 확보한다. VTAB-1K 데이터에서 ViT-B/16 모델 전이 학습 실험에서 전체 모델 파라미터의 0.34%만 사용하여 풀 파인튜닝(Full - Fine Tuning) 대비 +7.6%의 성능을 확인하고, 전체 19개의 데이터셋 중 10개의 데이터셋에서 가장 좋은 성능을 보였다.

### 1. 서론

기존 연구에서는 풀 파인튜닝(Full-Fine Tuning)을 통해 효과적인 전이 학습 효과를 낸다. 하지만 풀 파인튜닝은 대규모 네트워크에서는 매우 높은 연산량을 요구한다. 이를 완화하기 위해 최근에는 핵심적인 가중치만 선택적으로 업데이트 하는 파인튜닝 기법들이 제안되었다. 그러나 기존 기법은 단일 단계로만 선택된 파라미터만 해제(Unfreeze)하여 재학습한다는 점에서, 동결 해제 순서나 단계별 업데이트 전략의 융통성이 부족하다.

본 논문에서는 이러한 한계를 극복하기 위해 Gradual Unfreeze Fine-Tuning(GUFT)라는 새로운 전이 학습 프레임워크를 제안한다. GUFT는 다음과 같은 특징을 가진다. 1. 중요도 기반 파라미터 선별. 2. 점진적 동결 해제 및 재학습. 3. 풀 파인튜닝 대비 성능 향상. 이로써 GUFT를 통해 성능은 향상하고 기존 전이 학습 방식의 연산량 문제를 해결하며 경량화된 환경에서 최적의 전이 학습 성능을 보장한다.

### 2. 관련 연구

#### 2.1 풀 파인튜닝(Full-Fine Tuning)

기존 많은 연구에서 사전 학습된 모델에서 새로운 과제에 효율적으로 전이하기 위해 모든 파라미터를 재학습하는 풀 파인튜닝 기법을 사용한다. 성능은 뛰어나지만 연산 메모리 등 제약이 많다. 대규모 모델에서 현실적 제약이 있다.

#### 2.2 파라미터 효율적 파인튜닝 기법

LoRA[1]에서는 기존 가중치를 고정하고 소수의 새로운 파라미터(1% 미만)만 학습하여 학습 비용을 절감한다. 그러나 추론 시 추가 모듈 호출이 필요하고 모델 가중치를 직접 업데이트할 수 없다. GUFT 방법에서는 LoRA[1]와 비슷하게 혹은 더 적은 파라미터를 학습하여 더 좋은 성능을 낸다.

#### 2.3 중요도 기반 파라미터 선별 방법

기존 연구들 Lottery Ticket Hypothesis[2] 연구에서는 파라미터의 절댓값 크기를 기준으로 중요 파라미터를 판별한다. 하지만 최근 연구에서는 초기화 스케일에 민감하고 실제 손실에 얼마나 기여하는지 알 수 없기 때문에 다른 방법을 제시한다. SNIP(Single-Shot Network Pruning Based on Connection Sensitivity)[3]와 SalUN(Gradient-Based Weight Saliency)[4]에서는 각 파라미터의 그래디언트의 절댓값 크기를 기준으로 중요 파라미터를 판별한다.

$$w_i \leftarrow w_i - \eta \nabla_{w_i} L(f(x; w), y) \quad (1)$$

그 이유는 수식(1)에서 학습률  $\eta$ 와 그래디언트의 크기  $|\nabla_{w_i} L|$ 의 곱만큼 가중치  $w_i$ 가 이동한다. 즉  $|\nabla_{w_i} L|$ 가 클수록 이동량이 커져 학습 단계에서 영향력이 높다는 뜻이다. 따라서  $|\nabla_{w_i} L|$ 가 클수록 중요 파라미터라고 가정한다. 본 연구는 이들 기법의 강점을 살리면서 추가적인 방법을 제안한다.

### 3. Gradual Unfreeze Fine-Tuning 제안 방법

#### 3.1 그래디언트의 절댓값 크기 기반 중요도 산정

일정 에포크마다 전체 데이터셋  $D$ 를 역전파 후 가중치별  $|\nabla_{w_i} L|$ 을 누적값  $G_i$ 를 계산한다.

$$G_i = \sum_{(x,y) \in D} |\nabla_{w_i} L(f(x; w), y)| \quad (2)$$

수식(2)에서  $G_i$ 이 클수록 해당 가중치는 손실에 영향이 크다고 판단한다.

#### 3.2 점진적 파라미터 해제(Gradual Unfreeze)

GUFT에서는 학습되는 파라미터 집합을 한 번에 해제하는 대신,  $|\nabla_{w_i} L|$ 의 크기 기반 마스크를 정해진 비율에 맞춰 재계산하며 학습되는 파라미터의 비율을 점진적으로 늘려 모델의 안정성과 성능 향상을 유도한다.

### 3.3 Gradual Unfreeze Fine-Tuning 알고리즘 요약

알고리즘 Gradual Unfreeze Fine-Tuning	
입력 :	
- 재계산 에포크 리스트 $E_{list}$	
- 새로 계산한 파라미터에 대한 그래디언트 마스크 $M_{new}$	
- 이전 파라미터에 대한 그래디언트 마스크 $M_{prev}$	
- 각 재계산 시점마다 계산되는 비율 리스트 $R_{list}$	
1. 모든 파라미터에 대해 $M_{prev}, M_{new}$ 을 0으로 초기화.	
2. 각 $E$ 에 대해 다음을 수행한다.	
3. 만약 $E$ 가 $E_{list}$ 내의 $i$ 번 항목과 같다면:	
4. $M_{new}$ 을 0으로 초기화.	
5. 해당 시점의 목표 비율 $r \leftarrow R_{list}[i]$	
6. 모든 파라미터의 그래디언트 절댓값을 누적할 $G$ 를 0으로 초기화	
7. 전체 학습 데이터를 배치 단위로 순회하며:	
8. - 순전파(forward)를 수행하여 손실(loss)을 계산하고	
9. - 역전파(backward)를 수행하여 그래디언트를 구한다.	
10. - 각 파라미터의 그래디언트 절댓값을 $G$ 에 누적한다.	
11. 누적된 $G$ 를 평탄화(flatten)하고, 상위 $r\%$ 파라미터 인덱스를 선택	
12. 선택된 인덱스에 1, 나머지는 0을 부여한 $M_{new}$ 초기화	
13. $M_{prev} \leftarrow M_{prev}$ or $M_{new}$ (기존 마스크와 누적)	
14. $M_{prev}$ 을 기준으로 마스크가 1인 파라미터만 업데이트	

## 4. 실험

### 4.1 실험 환경

GUFT 실험에서는 효율적 파인튜닝 기법의 범용성과 데이터 효율성을 동시에 검증하기 위해 Visual Task Adaptation Benchmark(VTAB-1K)을 채택하였다. VTAB-1K는 자연 이미지 기반[표 1]·특수[표 2]·구조적[표 2, 3] 시각 데이터의 세 영역에 걸쳐 총 19개 데이터셋을 포함하고, 각 데이터셋당 학습 샘플이 1,000장으로 제한되어 있어 적은 데이터 환경에서의 일반화 성능과 학습 안정성을 평가할 수 있다. 또한 기존 연구 BayesTune[5]에서 사용한 표준 벤치마크이므로, 동일 조건에서 성능, 연산량을 직접 비교할 수 있다. 실험 모델로는 ImageNet-22K에서 사전 학습된 ViT-B/16을 사용하여 기존 연구와 동일한 모델로 실험을 진행하였다.

Method	#param	CIFAR100	Caltech101	DTD	Flower102	Pets	SVHN	Sun397
Full	85.8	68.9	87.7	64.3	87.2	86.9	87.4	38.8
Linear	0.04	64.4	85.0	63.2	97.0	86.3	36.6	51.0
LoRA	0.29	67.1	<b>91.4</b>	69.4	98.8	<b>90.4</b>	85.3	54.0
Ours	0.3	<b>72.8</b>	89.6	<b>70.9</b>	<b>99.2</b>	88.7	<b>88.5</b>	<b>57.0</b>

[표 1] VTAB-1K 자연 이미지 시각 분류 정확도(%)와 평균 학습 파라미터 수(M)

### 4.2 실험 결과 정리

GUFT 실험에서는 목표 마스크 비율을 전체 파라미터의 0.3%로 설정하고, 학습 과정에서 1, 6, 11 에포크마다 0.1%씩 점진적으로 누적 마스크를 적용하여 최종적으로 전체 파라미터의 약 0.34%가 학습되도록 선택되었다. 이렇게 학습된 모델은 학습 파라미터를 0.3M(전체의 0.34%)만 사용하여 풀 파인튜닝에 비해 약 (+7.6%)의 성능 향상과 LoRA[1]에 비해 약

Method	#param	Camelyon	EuroSAT	Resisc45	Retinopathy	Clever-Count	Clever-Dist	DMLab
Full	85.8	79.7	95.7	84.2	73.9	56.3	58.6	41.7
Linear	0.04	78.5	87.5	68.5	<b>74.0</b>	34.3	30.6	33.2
LoRA	0.29	84.9	95.3	<b>84.4</b>	73.6	<b>82.9</b>	<b>69.2</b>	<b>49.8</b>
Ours	0.3	<b>86.5</b>	<b>96.0</b>	<b>84.4</b>	73.0	80.0	67.4	48.1

[표 2] VTAB-1K 특수, 구조적 시각 분류 정확도(%)와 평균 학습 파라미터 수(M)

Method	#param	KITTI	dSpr-Loc	dSpr-Ori	sNORR-Azi	sNORR-Ele	ALL Arg	#Rank 1
Full	85.8	65.5	57.5	46.5	25.7	29.1	65.0	0
Linear	0.04	55.4	12.5	20.0	9.6	19.2	52.9	1
LoRA	0.29	78.5	75.7	<b>47.1</b>	<b>31.0</b>	<b>44.0</b>	72.5	9
Ours	0.3	<b>81.9</b>	<b>81.1</b>	45.5	27.7	41.0	<b>72.6</b>	<b>10</b>

[표 3] VTAB-1K 구조적 시각 분류 정확도(%)와 평균 학습 파라미터 수(M)와 VTAB-1K 전체 데이터셋에 대한 평균, Rank 1 히트

(+0.1%)의 성능을 향상을 보였다. 또한 전체 19개의 데이터셋중 10개에서 가장 높은 성능을 보였다.

## 5. 결론

본 연구에서는 Gradual Unfreeze Fine-Tuning(GUFT)를 통해 효율적인 전이 학습을 위해 새로운 프레임워크를 제안하여 기존 연구보다 높은 성능 향상을 보였다. 또한 학습 파라미터 비율  $R_{list}$ 과 재계산 주기  $E_{list}$ 를 하이퍼파라미터로 설정함으로써, 연산량을 조절하면서도 과제에 따라 성능을 더욱 향상할 수 있는 유연한 조정이 가능하다.

## ACKNOWLEDGMENT

이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구 결과임 (No.RS-2025-02219317, AI 스타펠로우십지원사업(국민대학교))

## 참고 문헌

- [1] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan AllenZhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.
- [2] J. Frankle and M. Carbin, “The lottery ticket hypothesis: Finding sparse, trainable neural networks,” in Proc. Int. Conf. Learn. Representations, 2019.
- [3] N. Lee, T. Ajanthan, and P. Torr, “SNIP: Single-shot network pruning based on connection sensitivity,” in Proc. Int. Conf. Learn. Representations, 2019.
- [4] Chongyu Fan, Jiancheng Liu, Yihua Zhang, Dennis Wei, Eric Wong, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. arXiv preprint arXiv:2310.12508, 2023.
- [5] Kim, M. and Hospedales, T. BayesTune: Bayesian Sparse Deep Model Fine-tuning. Advances in Neural Information Processing Systems, 2023.