초심층 잔차 네트워크의 그래디언트 흐름 분석과 격자 기반 분화-병합 패러다임

박병준, 한동석

경북대학교 대학원 전자전기공학부

qudwns7171@knu.ac.kr, dshan@knu.ac.kr

Gradient Flow Analysis of Ultra-Deep Residual Networks and a Grid-Based Bifurcation-Merge Paradigm

Byeong-Jun Park, Dong Seog Han

School of Electronic and Electrical Engineering, Kyungpook National Univ.

요 약

초대규모 합성곱 신경망은 다양한 분야에서 성능을 크게 향상시켰으나, 잔차 연결 기반 구조는 깊이가 증가할수록 그래디언트 불안정성이 발생한다. 본 논문에서는 그래디언트 흐름의 관점에서 순차적 잔차 네트워크를 분석하고, 깊이에 따른 그래디언트 분산 감소와 장경로 지배 현상을 이론적으로 규명한다. 이를 극복하기 위해, 본 논문은 레이어와 블록을 일차원적으로 적층하는 대신 정사각 격자 기반 분화-병합 구조를 제안한다. 제안 방식은 입력을 다중 경로로 분화하고 병합 지점에서 평균 연산을 적용하여 균형 잡힌 그래디언트 전파를 유도한다. 실험 결과, 본 구조는 기존 순차형 네트워크 대비 더 안정적인 그래디언트 분포와 빠른 수렴 특성을 달성하며, 추가 파라미터 없이도 학습 효율과 정확도를 향상시킬 수 있음을 입증한다.

Ⅰ. 서 론

초대규모 합성곱 신경망(convolutional neural network, CNN)은 컴퓨터 비전, 자연어 처리, 멀티모달 학습 등 다양한 분야에서 성능을 크게 향상시켰다. 잔차 네트워크(ResNet [1])와 그 변형 구조는 항등 기반 지름길 연결을 통해 기울기 소실 문제를 완화하여 매우 깊은 모델의 학습을 가능하게 했으나, 수백~수천 레이어로 확장되는 초대규모 모델에서는 여전히 그래디언트 불안정성이 발생한다. 특히 순차적 잔차 구조에서는 깊이에 따라 그래디언트 분산이 누적되어 최적화 효율과 일반화 성능이 저하되는 한계가 보고되었다 [2].

이러한 문제를 완화하기 위해 Inception [3], ResNeXt [4]와 같은 다중 브랜치 구조가 제안되었으나, 이는 주로 특징 융합 효율에 초점을 두고 있어 그래디언트 전파 안정성을 직접적으로 해결하지는 못한다. 우리는 초대규모 네트워크에서 그래디언트가 분화(bifurcation)되고 병합(convergence)되는 지점이 불안정성이 핵심이라고 보고, 이를 구조적으로 재구성함으로써 내부 연산을 변경하지 않고도 그래디언트 안정성을 향상시킬 수 있다는 가설을 제시한다.

Ⅱ. 기울기 흐름 분석

딥러닝 네트워크를 방향성 비순환 그래프(directed acyclic graph, DAG)로 나타내면, 각 노드의 순전파와 역전파는 다음과 같이 정의된다.

$$\mathbf{x}_i = \mathcal{F}_i(\mathbf{z}_i), \quad \mathbf{z}_i = \sum_{j \in \mathcal{P}(i)} \mathbf{x}_j \cdot w_{j,i} \tag{1}$$

여기서 $F_i(\cdot)$는 해당 노드의 변환 연산(예: 합성곱 + 정규화 + 활성화 함수), $P(i)$는 노드 i의 부모 노드 집합, $w_{j,i}$는 연결 가중치(잔차 연결의 경우 보통 1)를 나타낸다.

역전파(backpropagation)에서, 손실 함수 L에 대한 노드 i 출력의 그래디언트는 다음과 같이 표현된다.

$$\mathbf{g}_i = \frac{\partial \mathcal{L}}{\partial \mathbf{x}_i} = \sum_{k \in \mathcal{C}(i)} \mathbf{g}_k \cdot \frac{\partial \mathbf{x}_k}{\partial \mathbf{x}_i} \tag{2}$$

여기서 $C(i)$는 노드 i의 자식 노드 집합이다.

순차형 잔차 네트워크에서는 대부분의 경우 $|P(i)| = 1$이며, 스킵 연결이 있을 때만 2가 된다. 이 경우 그래디언트 경로 수가 제한되는데, 분화 구간에서는 $|P(i)| > 1$이 되어 그래디언트 신호가 여러 경로로 분리되고, 병합 구간에서는 $|P(i)| = 2$가 되어 서로 다른 경로의 그래디언트가 합쳐진다. 병합 연산은 다음과 같이 정의된다.

$$\mathbf{x}_m = \frac{\mathbf{x}_{p_1} + \mathbf{x}_{p_2}}{2} \tag{3}$$

$$\mathbf{g}_{p_1} = \frac{1}{2}\mathbf{g}_m, \quad \mathbf{g}_{p_2} = \frac{1}{2}\mathbf{g}_m \tag{4}$$

여기서 p_1, p_2는 병합 노드 m의 두 부모 노드이다.

이 평균 연산 규칙은 두 부모 노드의 기여도를 동일하게 유지하도록 하여, 병합 지점에서의 그래디언트 분포를 균형 있게 만든다. 이론적으로, 분화-병합 사이클을 반복적으로 적용하면 깊이에 따른 그래디언트 분산의 급격한 감소를 완화하고, 순차형 구조에서 나타나는 장경로(long-path) 지배 현상을 줄일 수 있다.

깊이 d에서의 그래디언트 분산을 $\sigma_g^2(d)$라고 하면, 순차형 모델에서는 경험적·이론적 분석 모두 다음과 같은 경향을 보인다.

$$\sigma_g^2(d) \propto \alpha^d \tag{5}$$

여기서 $|\alpha| < 1$이면 기울기 소실 구간이 된다. 반면, 제안하는 분화-병합 패러다임에서는 다음과 같이 변화한다.

$$\sigma_g^2(d) \approx \beta^d + \frac{1-\beta^d}{2} \tag{6}$$

여기서 d가 커질수록 순차 구조보다 그래디언트 감쇠 속도가 느려지고 깊이에 따라 보다 균일한 그래디언트 에너지를 유지한다.

Ⅲ. 격자 기반 분화-병합 패러다임

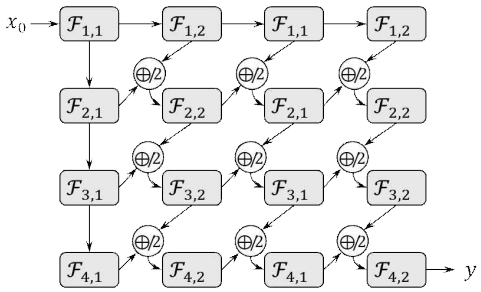

그림 1. 격자 구조의 분화-병합 패러다임

기존의 1차원 순차형 잔차 구조의 한계를 극복하기 위해, 본 논문에서는 레이어 또는 블록을 배치하는 데 있어 2차원 정사각 격자 토폴로지를 제안한다. 그림 1에 나타난 바와 같이, 네트워크는 $R \times R$ 격자로 구성되며 각 노드는 연산 블록 $F_{i,j}$에 해당한다. 데이터는 격자의 수평 및 수직 방향으로 흐르며, 이 과정에서 구조적 그래디언트 분화가 발생하고, 이후 그래디언트 병합이 이루어진다.

노드 연결 방식은 두 가지로 분류된다. 분화 노드는 하나의 입력 피처맵이 두 방향으로 복제되어 각각 다른 경로로 전달된다. 병합 노드는 두 개의 입력 피처맵이 다음과 같이 평균 연산을 거친 후 다음 블록으로 전달된다.

$$\mathbf{x}_{i,j} = \frac{\mathbf{x}_{i-1,j} + \mathbf{x}_{i,j-1}}{2} \tag{7}$$

이 평균 규칙은 두 입력의 기여도를 동일하게 유지하도록 하며, 역전파 과정에서는 다음과 같이 균형 잡힌 그래디언트 분포를 보장한다.

$$\mathbf{g}_{i-1,j} = \frac{1}{2}\mathbf{g}_{i,j}, \quad \mathbf{g}_{i,j-1} = \frac{1}{2}\mathbf{g}_{i,j} \tag{8}$$

그림 1의 격자는 하나의 구조적 모듈을 나타낸다. 이 모듈을 K번 반복하여 대규모 네트워크를 구성할 경우, 분화-병합 이벤트의 수는 격자 해상도 R의 증가에 따라 제곱 비율로 늘어난다. R이 커질수록 입력과 출력 사이의 평균 그래디언트 경로 길이가 다양해져, 특정 경로에 그래디언트가 집중되는 현상을 줄이고 다중 경로 기반의 그래디언트 안정성을 촉진한다.

Ⅳ. 실험

모든 실험은 Windows - Conda 환경에서 NVIDIA RTX 4090 GPU, CUDA 12.6, PyTorch 2.7을 기반으로 수행되었다. 재현성을 위해 랜덤 시드를 42로 고정하였다. 결과는 세 번의 독립 실행 평균으로 보고하였다.

실험에는 CIFAR-10 [5]데이터셋을 사용하였다. 모델은 SGD(Nesterov 모멘텀 0.9, weight decay 0.0004)를 사용해 300 에폭 동안 학습되었으며, 초기 학습률 0.1은 학습 스케줄의 50%와 75%에서 10배씩 감소시켰다. 손실 함수는 교차 엔트로피를 적용하였고, 배치 크기는 128로 설정하였다. 데이터 증강은 랜덤 크롭과 수평 반전을 사용하였다.

역전파 과정에서 발생하는 그래디언트 소실 현상을 정량적으로 평가하기 위해, 모델의 모든 학습 가능한 파라미터에 대해 그래이언트 크기의 분포를 분석하였다. 각 파라미터 텐서에 대해, 검증 데이터를 기반으로 손실 함수의 그래디언트를 계산하고, 이를 1차원으로 평탄화한 후 전체 모델의 그래디언트 값을 하나의 집합으로 결합하였다. 이 때 그래디언트의 절댓값이 미리 정의된 임계값 ϵ보다 작으면 해당 그래디언트를 '소실된(vanished)' 것으로 간주한다. 이에 따라, 그래디언트 소실 비율은 다음과 같이 정의된다.

$$\text{Vanished Ratio} = \frac{\#\{|\nabla_\theta \mathcal{L}| < \epsilon\}}{\#\{\nabla_\theta \mathcal{L} \text{ elements}\}} \tag{9}$$

표 1에서 확인할 수 있듯이, 제안 구조는 파라미터 수와 FLOPs에서 기존 순차형 네트워크와 동일한 수준을 유지한다. 그러나 표 1의 정확도와 표 2의 분포 비교 결과, 모델의 규모가 커질수록 제안 구조는 순차형 네트워크에 비해 분류 정확도가 높으며, 그래디언트 값이 소실 범위(ϵ) 내에 치우치지 않고 더 균형적으로 분포한다. 이는 분화-병합 지점이 그래디언트의 분산을 완화하며, 학습 안정성을 개선하는 데 기여함을 시사한다.

표 1. 제안 구조와 순차형 구조의 모델 크기, 연산량 및 정확도 비교

#layers	#Params.	#FLOPs			Acc (%)		
		VGGs	ResNets	Ours	VGGs	ResNets	Ours
26	0.37M	55.70M	56.01M	56.01M	91.34	**93.07**	92.68
98	1.54M	228.32M	228.63M	228.63M	64.56	94.03	**94.43**
386	6.20M	918.81M	919.11M	919.11M	10.00	93.55	**94.93**
1538	24.87M	3.68G	3.68G	3.68G	10.00	94.21	**95.09**

표 2. 초심층 네트워크에 대한 기울기 분포 및 소실도 비교

Model	Total gradients	Standard deviation	Vanished gradients ratio (%)			
			$\epsilon < 10^{-4}$	$\epsilon < 10^{-5}$	$\epsilon < 10^{-6}$	$\epsilon < 10^{-7}$
ResNet-1538	$\approx 2.49 \times 10^9$	2.60×10^{-3}	64.64	39.55	22.68	14.72
Ours (1538 layers)	$\approx 2.49 \times 10^9$	1.52×10^{-3}	40.10	16.45	9.31	6.51

Ⅴ. 결론

본 논문에서는 초심층 CNN의 순차형 잔차 아키텍처를 기울기 흐름 관점에서 분석하고, 다중 브랜치의 위상적 장점에서 영감을 받은 격자 기반 분화-병합 구조를 제안하였다. 제안 방식은 추가 파라미터나 연산 수정 없이 기존 연산 단위를 재배치하는 것만으로 그래디언트 분포를 균형 있게 유지하며, 초대규모 환경에서 학습 안정성과 효율을 동시에 향상시킴을 실험을 통해 입증하였다. 향후 연구에서는 분화-병합 구조의 효과를 극대화할 수 있는 새로운 설계 방식을 고안하고, 수학적 분석과 더불어 더 큰 데이터셋과 다양한 아키텍처로 확장성을 검증할 예정이다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터(ITRC)의 지원을 받아 수행된 연구임 (IITP-2025-RS-2020-II201808).

참 고 문 헌

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[2] C. Zhang, S. Bengio, and Y. Singer, "Are all layers created equal?" *Journal of Machine Learning Research*, vol. 23, no. 67, pp. 1-28, 2022.

[3] Szegedy, Christian, *et al.*, "Going deeper with convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2025, pp. 1-9.

[4] S. Xie, R. Girchick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5987-5995

[5] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2007, accessed: 2025-08-15.