Logits 기반 외부 보정을 통한 손상된 심층 신경망의 기능적 복구

황선준, 하지혁*
연세대학교

sunjun7559012@yonsei.ac.kr, *jihyeok0502@yonsei.ac.kr

Functional Recovery of Corrupted Deep Neural Networks via Logits-Based External Correction

Sun Jun Hwang, Ji Hyeok Ha*
Yonsei Univ.

요 약

안전-필수 응용 분야에 배포된 심층신경망은 메모리 오류, 비트 플립 등 소프트웨어 오류로 인한 파라미터 손상에 취약하다. 본 논문에서는 원본 파라미터를 수정하지 않고 추론 시점에서 손상된 모델 출력을 보정하는 기능적 복구(Functional Recovery) 기법을 제안한다. 지식 증류를 통해 학습된 외부 델타 매핑 네트워크(EDMN)를 활용하여 손상된 logits 을 정상 모델 출력 방향으로 변환하며, 잔차 학습을 통해 보정항 Δ(z)를 학습한다. MNIST 실험 결과, 가우시안 노이즈(σ=0.1)에서 + 6.29%p, 영점화 결함(level=0.9)에서 + 4.68%p 의 정확도 복구를 달성하였다. 그러나 특정 결함 수준에서 복구가 오히려 성능을 저하시키는 과보정 문제도 확인되어, 적응적 선택 메커니즘의 필요성을 제시한다.

Ⅰ. 서 론

심층신경망(Deep Neural Networks, DNNs)이 자율주행 차량, 의료 진단 등 안전-필수 응용 분야에 배포됨에 따라, 하드웨어 및 소프트웨어 결함 상황에서의 신뢰성 보장이 중요해지고 있다[1,2]. 최근 연구들은 고에너지 입자 충돌, 메모리 손상, 저장 매체 열화 등으로 인한 소프트 오류가 DNN 추론 정확도를 크게 저하시킬 수 있음을 보여주었다[3].

삼중 모듈 중복(TMR)이나 오류 정정 코드(ECC)와 같은 전통적인 결함 내성 기법들은 하드웨어 자원, 전력 소모 측면에서 상당한 오버헤드를 발생시킨다[4]. 더욱이 파라미터 손상 발생 시 모델 재학습은 계산 비용이 높고 엣지 배포 환경에서는 비실용적이다.

본 논문에서는 파라미터 복원 없이 추론 시점에서 손상된 DNN 출력을 보정하는 기능적 복구(Functional Recovery) 기법을 제안한다. 핵심 통찰은 모델 파라미터가 손상되더라도 출력 Logits 이 복구를 가능하게 하는 충분한 구조적 정보를 유지한다는 것이다. Hinton 등[5]의 지식 증류와 He 등[6]의 잔차 학습을 결합하여 손상된 Logts 을 정상 분포로 유도하는 보정 네트워크를 학습한다.

Ⅱ. 본론

2.1 제안 방법

외부 델타 매칭 네트워크(EDMN)는 손상된 Logits z_{cor} 를 입력으로 받아 보정항을 출력하는 경량 MLP(10→256→256→10)이다. 잔차 학습 원리에 따라 $z_{rec} = z_{cor} + \Delta(z_{cor})$로 정의하며, 마지막 레이어를 0 으로 초기화하여 학습 초기 항등 함수 동작을 보장한다.

학습 목적 함수는 지식 증류와 분류 손실을 결합한다.

$$L = \alpha KL\left(softmax\left(\frac{z_{rec}}{T}\right), softmax\left(\frac{z_{clean}}{T}\right)\right) \cdot T^2 + \gamma \cdot CE(z_{rec}, y) + \beta||\Delta||^2$$

KL 발산 항은 정상 모델의 소프트 확률 분포 매칭을 유도하고, 교차 엔트로피 항은 올바른 분류를 보장하며, L2 정규화는 과도한 보정을 방지한다. $\alpha = \gamma = 1.0, T = 2.0, \beta = 1e-4$를 사용하고, 정상 샘플을 20% 포함하여

표 1. 영점화 결함 복구 결과 (FC2 레이어)

수준(ρ)	손상 정확도	복구 정확도	ΔAcc
0.0	98.97%	98.91%	-0.06%p
0.2	98.16%	98.32%	+ 0.16%p
0.5	94.91%	95.53%	+ 0.62%p
0.8	63.51%	59.35%	-4.16%p
0.9	49.28%	53.96%	+ 4.68%p

표 2. 가우시안 노이즈 결함 복구 결과 (전체 레이어)

수준(σ_W)	손상 정확도	복구 정확도	ΔAcc
0.00	98.97%	98.94%	-0.03%p
0.05	96.60%	97.36%	+0.67%p
0.10	50.83%	57.12%	+6.29%p
0.15	20.24%	17.71%	-2.53%p
0.20	13.60%	14.23%	+0.63%p

항등 동작을 강화한다. 사용된 MLP 의 전체 파라미터 수는 약 7.1 만(0.07M) 수준으로 추론 시 오버헤드가 매우작다.

2.2 실험 설정

MNIST 와 표준 CNN(합성곱 2 층 + 완전 연결 2 층, 정상 정확도 98.97%)을 사용한다. 결함 모델로 (1) 영점화: fc2 레이어 가중치를 $\rho \in \{0.0, 0.2, 0.5, 0.8, 0.9\}$ 비율로 0 설정(Stuck-at-0 시뮬레이션), (2) 가우시안: 전체 레이어에 $\sigma \in \{0.0, 0.05, 0.1, 0.15, 0.2\}$ 노이즈 추가(점진적 열화 모델링)를 적용한다. EDMN 은 15 에폭, Adam(lr=1e-3), 배치 크기 128 로 학습한다.

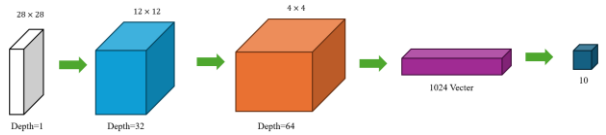

그림 1. 모델 구조도

2.3 실험 결과

표 1 은 FC2 레이어에 대한 영점화 결함 결과를, 표 2 는 전체 레이어에 대한 가우시안 노이즈 결함 결과를 보여준다.

2.4 결과 분석

복구 가능 영역의 존재: 실험 결과는 기능적 복구가 유의미한 이점을 제공하는 복구 가능 영역이 존재함을 입증한다. 가우시안 $\sigma = 0.1$ 에서 +6.29%p 의 정확도 향상(50.83%→57.12%), 영점화 0.9 에서 +4.68%p 향상(49.28%→53.96%)을 달성하였다. 이는 손상된 Logits 이 효과적인 보정을 가능하게 하는 충분한 구조적 정보를 유지함을 보여준다.

과보정 문제: 그러나 복구가 항상 유익하지 않음을 확인하였다. 영점화 0.8 에서 -4.16%p(63.51%→59.35%), 가우시안 $\sigma = 0.15$ 에서 -2.53%p(20.24%→17.71%)의 성능 저하가 발생하였다. 이는 단일 보정기 $\Delta(\cdot)$ 를 모든 결함 강도에 동일하게 적용할 경우, 결함 강도에 따라 최적 보정 방향이 달라져 일부 구간에서 over-correction 이 발생할 수 있음을 의미한다.

Ⅲ. 결론

본 논문에서는 파라미터 수정 없이 출력 Logits 을 보정하는 기능적 복구 기법을 제안하였다. 지식 증류와 잔차 학습을 결합한 EDMN 을 통해 중간 수준 손상에서 최대 +6.29%p 의 정확도 복구를 달성하였으며, 하드웨어 수정이나 재학습 없이 추론 시점에서 적용 가능하다는 장점이 있다.

그러나 특정 결함 수준에서 과보정 문제가 발생함을 확인하였으며, 이는 적응적 선택 메커니즘의 필요성을 제시한다. 향후 연구로는 손상 심각도 추정 기반 적응적 보정 강도 조절, 대규모 모델 및 데이터셋으로의 확장, 복구 가능 영역에 대한 이론적 분석이 있다.

참 고 문 헌

[1] C. Torres-Huitzil and B. Girau, “Fault and Error Tolerance in Neural Networks: A Review,” IEEE Access, Vol. 5, pp.17322-17341, 2017

[2] Y. Ibrahim et al., “Soft errors in DNN accelerators: A comprehensive review,” Microelectronics Reliability, vol. 115, p. 113969, 2020.

[3] G. Li et al., “Understanding Error Propagation in Deep Learning Neural Network Accelerators and Applications,” SC’17, 2017.

[4] M. Raji et al., “An ECC-based Fault Tolerance Approach for DNNs,” arXiv:2508.12347, 2025.

[5] G. Hinton, O. Vinyals, and J. Dean, Distilling the Knowledge in a Neural Network," arXiv:1503.02531, 2015.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," CVPR, 2016.