적응형 전략 탐색 기반의 원샷 프루닝 기법 평가

이현수[1], 문용혁[2*]
[1]성신여자대학교 컴퓨터공학과 학부생
[2]성신여자대학교 컴퓨터공학과 교수
stella9921@gmail.com, yhmoon@sungshin.ac.kr

Adaptive Strategy Exploration for One-Shot Pruning: An Empirical Evaluation

Hyunsu Lee[1], Yong-Hyuk Moon[1*]
[1]Dept. of Computer Engineering, Sungshin Women's University

요 약

딥러닝 모델 효율화를 위한 원샷 프루닝은 점진적 프루닝에 비해 효율적이지만, 최적의 전략을 사전에 결정하기 어렵다는 한계가 있다. 본 논문에서는 이를 해결하기 위해, 실제 프루닝에 앞서 필터 민감도 분석과 성능 피드백을 통해 레이어별 최적의 프루닝 전략을 탐색하는 기법을 제안한다. 이렇게 탐색된 최종 전략을 적용해 모델을 단 한 번에 프루닝함으로써, CIFAR-100 데이터셋과 ResNet 모델 기반 실험에서 높은 압축률을 유지하면서도 안정적인 성능을 보장함을 입증하였다.

1. 서론

최근 딥러닝 모델 경량화를 위해 프루닝(Pruning) [1]이 널리 활용되고 있으나, 기존 방식들은 연산 비용이 크거나 성능 저하의 위험이 존재한다. 특히 원샷 프루닝(One-Shot Pruning)은 속도 면에서 유리하지만, 레이어별 제거 비율을 사전에 최적화하기 어렵다는 한계를 가진다. 본 연구에서는 이러한 한계를 극복하기 위해, 레이어별 압축대비 성능 변화를 기반으로 프루닝 전략을 학습하는 사전 단계를 제안한다. 이 적응형 전략 탐색 기법은 레이어별 민감도를 동적으로 탐색하고, 최적의 스코어 함수를 도출함으로써 원샷 프루닝 기반 모델 경량화 성능 향상에 기여할 수 있다.

2. 적응형 전략 탐색 기반 원샷 프루닝 기법

본 원샷 프루닝 기법은 최적 전략을 탐색하는 사전 준비 단계 1-2와 실제 프루닝을 대상 모델에 적용하는 단계 3으로 구성된다.

- (단계 1) 초기 민감도 획득: 각 레이어를 특정 압축률(%)에 따라 개별적으로 프루닝한 후, 정확도 하락을 측정하여 초기 민감도를 산출한다.
- (단계 2) 적응형 전략 탐색: n-라운드의 가상 프루닝 시뮬레이션을 반복하며, 매 라운드의 성능 피드백을 바탕으로 민감도 점수를 동적으로 업데이트하여 최적화한다.
- (단계 3) 최종 원샷 프루닝: 최종 탐색된 레이어별 민감도를 최적의 경량화 전략으로 사용하여 원본 모델을 단 한번 프루닝하고, 재학습시켜 최종 경량 모델을 도출한다.

3. 적응형 전략 탐색을 위한 평가 함수 설계

원본 모델의 각 레이어가 특정 압축 상황에서 성능(정확도)에 미치는 영향을 민감도 s_i로 정의하고, 압축률 변화에 따른 민감도 변화 특성을 세 가지 평가(Score) 함수로 설계한다. 본 평가 함수에 따라 레이어별 적응형 윗샷 프루닝이 수행된다.

1) 제안 방법 1(민감도 정규화): 레이어의 크기(파라미터 개수)와 둔감도를 곱하여 스코어 결정 방식

$$pr_i = C_{target} \cdot f_i \cdot \frac{w_i}{\sum_j f_j w_j},\ w_i = \frac{1/(s_i+\epsilon)}{\sum_j 1/(s_i+\epsilon)}$$

- pr_i : 레이어 i에 적용될 최종 프루닝 비율
- C_{target}: 모델에 대한 최대 목표 압축률 (필터 기준)
- f_i : 레이어 i의 파라미터 수, 즉 레이어의 크기
- w_i : 레이어 i의 정규화된 둔감도 (민감도의 역수)
- $\sum_j f_j w_j$: 모든 레이어의 '크기 × 둔감도' 의 총합

2) 제안 방법 2(민감도 증폭): 민감도에 증폭 계수($p > 1$)를 적용하여 민감도 차이를 극대화하는 '선

* 교신저자

Score Function	Parameters	Compression Ratio(%)	Validation Acc.(%)	Test Acc(%)
Regularization	-	37.61%	72.80%	70.77%
Amplification (p)	p=1.5	37.60%	74.34%	**71.56%**
	p=2.0	37.60%	72.62%	69.98%
	p=2.5	**37.61%**	**72.80%**	70.77%
Weighted Sum (β)	β=0.2	21.22%	**73.84%**	70.30%
	β=0.5	22.76%	71.90%	**71.40%**
	β=0.8	**33.76%**	72.04%	71.14%

Table 1. Ablation study on the proposed score functions for ResNet-18 with 80% target pruning ratio.

택과 집중' 전략

$$w'_i = \frac{(1/(s_i+\epsilon))^p}{\sum_j (1/(s_i+\epsilon))^p}$$

3) 제안 방법 3(민감도 가중합): 크기와 민감도를 베타(β) 계수로 가중합하여 두 요소의 균형을 조절하는 전략

$$pr_i = C_{target} \cdot (\beta \cdot f_i + (1-\beta) \cdot w_i)$$

4. 성능 평가 실험

제안한 적응형 전략 탐색 기반의 원샷 프루닝 기법의 평가는 ResNet-18과 ResNet-34 [2]를 기반으로 CIFAR-100 데이터셋을 사용하여 수행하였다. 또한 적응형 전략 탐색(단계 2)에서는 10 라운드(5 epochs/round) 민감도 평가가 수행되고, 최종 프루닝(단계 3) 이후에는 20 epochs의 재학습을 설정하였다. 특히 레이어 별 최적 프루닝 비율 탐색을 위해 목표 압축률의 최대 상한을 80%로 설정하고, 세 가지 민감도 평가함수(정규화, 증폭, 가중합)가 원샷 프루닝 성능에 미치는 영향을 비교하였다.

4.1 스코어 함수별 성능 비교

Table 1의 결과에서 알 수 있듯이, 민감도 증폭 방식(p=1.5)은 중요한 레이어를 보호하고 둔감한 레이어를 과감히 제거함으로써 71.56%의 최고 정확도를 달성하였다. 반면에 민감도 가중합 방식은 레이어 크기와 민감도를 균형있게 (β=0.5) 설정할 때 가장 우수한 성능을 보여, 본 방식의 효과성을 뒷받침하였다. 한편, 목표 압축률(80%)은 필터 개수를 기준으로 하나, 실제 압축률(20-30%대)은 파라미터 기준으로 산정된 것이다. 이러한 차이는 모델 후반부 레이어에 위치한 소수의 필터가 전체 파라미터의 대부분을 차지하기 때문에 발생한 차이로 이해된다. 따라서 실제 파라미터 압축률이 목표치보다 낮게 나타난 것은, 제안한 평가 함수가 성능에 치명적인 대규모 필터를 의도적으로 보호했음을 보여주는 긍정적인 결과로 판단할 수 있다.

4.2 제안 방법의 일반화 성능 검증

Table 2의 결과에서 보듯, 더 깊은 ResNet-34 모델의 일반화 성능 검증에서는 민감도 가중합 방식(β=0.5)이 가장 안정적이고 높은 성능을 보였다. 이는 모델이 깊어질수록 레이어의 기능이 전문화되기 때문에, 민감도만을 기준으로 한 과감한 '선택과 집중' 전략보다는 크기와 민감도를 균형 있게 반영하는 방식이 핵심 기능을 보존하는 데 더 효과적임을 의미한다.

Score Function	Parameters	Compression Ratio(%)	Original Test Acc.(%)	Pruned Test Acc.(%)
Sensitivity Amplification	p = 1.5	**18.81%**	75.50%	69.65%
Weighted Sum	β = 0.5	14.17%	75.50%	**70.47%**

Table 2. Generalization performance on ResNet-34.

5. 결론 및 향후 연구

본 논문은 최적의 프루닝 전략을 사전 탐색한 뒤 단 한 번에 적용하는 새로운 원샷 프루닝 프레임워크를 제안하였다. 실험을 통해 모델 아키텍처에 따라 최적의 스코어 함수가 달라질 수 있음을 확인하였으며, 특히 베타 가중합 방식이 깊은 모델에서 더 우수한 일반화 성능을 보임을 입증하였다. 다만 제안 방법은 단계 1, 2에서 높은 사전 계산 비용이 요구된다는 한계를 지니며, 향후 연구에서는 이러한 탐색 비용을 줄일 수 있는 비용 효율적 전략 탐색 기법을 연구할 계획이다.

Acknowledgement. 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (RS-2025-24292968).

참고문헌

[1] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning Filters for Efficient ConvNets", International Conference on Learning Representations (ICLR), Toulon, France, 2017.

[2] He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 2016, pp. 770-778.