GraNet 기반의 필터 프루닝을 적용한 경량 모델의 양자화 효과에 대한 연구

(A Study of Quantization Effect on a Lightweight Model with GraNet Filter Pruning)

설 광 수, 노 시 동, 정 기 석*
한양대학교
(Kwang-Soo Seol, Si-Dong Roh, Ki-Seok Chung)
(Hanyang Univ.)

Abstract : As convolutional neural networks get deeper and wider, model compression is being widely used to reduce the amount of computation and memory usage. Pruning, which includes structured pruning and unstructured pruning, is one of the widely-adopted model compression methods. The structured pruning can reduce the size of the network model by model thinning, but it may suffer from worse accuracy degradation than the unstructured method. In this study, we claim that if quantization is used in conjunction with the structured pruning, the data size can be reduced without significantly sacrificing the model's performance. We propose a lightweight model on which both the GraNet structured pruning and an 8-bit weight quantization are applied. We evaluate the performance of both static and dynamic quantization to quantize the pruned model. The experiment was conducted to perform image classification tasks using the ResNet18 model with pruning and quantization on CIFAR-100 datasets. Compared to the original model, we reduced the weight size of the model by 84.25%, 88%, and 96.25% with constraints of 2.5%, 5%, and 10% accuracy degradation using GraNet filter pruning and 8-bit quantization.

Keywords : Deep learning, Model compression, Pruning, Quantization

I. 서 론

모델 경량화는 신경망 모델의 정확도 손실을 최소화 하면서 모델의 크기를 줄이고, 학습 및 추론 시간을 단축시키는 기법이다. 딥러닝 분야가 발전하면서 모델의 정확도는 높아지고 있지만, 모델의 크기는 커지고, 학습 및 추론 시간은 길어지고 있다. 이로 인해 모델 경량화 기술의 중요도가 높아지고 있다. 대표적인 모델 경량화 기술로는 pruning과 quantization이 존재한다.

Pruning은 weight를 제거하는 단위에 따라서 unstructured pruning과 structured pruning으로 나눌 수 있다. 일반적으로 structured pruning을 사용해 경량화 한 모델이 unstructured pruning을 사용해 경량화 한 모델보다 정확도가 더 많이 감소한다. 따라서 structured pruning은 정확도의 큰 하락 없이는 모델의 크기를 크게 줄이지 못한다는 단점이 존재한다. 하지만, 본 연구에서는 이러한 단점은 pruning을 통해 경량화 된 모델에 quantization을 추가로 사용해서 완화할 수 있다는 것을 보여준다.

효과적인 pruning 기법을 위해서는 lottery ticket hypothesis (LTH) [1]가 제시된 이 후, dense neural network와 동일한 성능을 가지는 subnetwork를 효율적으로 찾는 연구들이 주로 진행되고 있다. 예를 들어 학습하는 도중 pruning을 진행하는 방식인 gradual pruning [2]이 있다. 최근에 gradual pruning에 의해 제거된 일부 연결을 재생성하는 개념을 도입한 gradual pruning with zero-cost neuroregeneration (GraNet) [3]이라는

*Corresponding Author (kchung@hanyang.ac.kr)
설광수, 노시동, 정기석: 한양대학교
※ 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원 (No.2021-0-00131, 제조검사장비 경량화를 위한 지능형 엣지컴퓨팅 반도체 개발)을 받아 수행된 연구임.

pruning 기법이 제시되었다. 해당 기법은 기존의 다른 pruning 기법 대비 훌륭한 성능을 보이나, 해당 pruning 기법과 함께 quantization이 적용되었을 때의 성능 결과는 아직 충분히 연구되지 않았다. 본 연구에서는 gradual structured pruning 및 neuroregeneration이 적용된 상황에서 quantization이 함께 적용된 모델을 제시하고, 이 모델의 성능을 평가 및 분석하였다.

Ⅱ. 관련 연구

1. Pruning

Pruning은 모델의 일부분을 제거하는 단위에 따라서 unstructured pruning과 structured pruning으로 나눌 수 있다. Unstructured pruning은 모델에 있는 각각의 weight를 일정한 기준에 따라 제거하는 기법이다. 반면에, structured pruning은 모델 내부의 filter와 같은 특정 구조 단위로 제거하는 기법이다.

일반적으로 같은 양의 weight를 제거할 때, unstructured pruning이 structured pruning보다 정확도가 높지만, structured pruning의 경우 deploy 과정에서 미사용 filter를 모델 연산에서 제거하기에 용이하므로 경량화된 모델의 추론 속도는 structured pruning이 unstructured pruning보다 더 빠르다.

Lottery ticket hypothesis (LTH) [1]에 의하면, 무작위로 초기화된 dense neural network는 기존 network와 같은 학습 횟수로 유사한 정확도를 달성 가능한 pruned subnetwork를 포함한다. 이러한 subnetwork를 효과적으로 찾는 방법 중 하나로 gradual pruning with zero-cost neuroregeneration (GraNet) [3]이 연구되었다. GraNet은 gradual pruning에 의해 제거된 일부 연결을 재생성하는 개념을 접목시킨 방식이다. GraNet의 동작방식은 먼저 모델을 학습시키고, 학습시킨 모델의 weight를 사전에 정한 기준에 따라서 pruning한다. 그 후에 모델을 다시 pruning하고, 두 번째로 pruning한 weight의 개수만큼 weight를 재생성한다. 이 과정을 학습이 종료될 때까지 반복한다.

그림 1은 GraNet의 동작 예시이다. 작은 정사각형은 weight를 나타내고, 16개의 작은 정사각형으로 이루어진 정사각형은 weight matrix를 나타낸다. 회색 정사각형은 제거되지 않은 weight이고, 흰색 정사각형은 제거된 weight를 나타낸다.

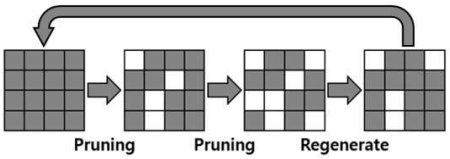

그림 1. GraNet 예시
Fig. 1. Example of GraNet

2. Quantization

Quantization [5]은 모델에 quantization을 적용하는 시점에 따라서 Post Training Quantization (PTQ)과 Quantization Aware Training (QAT)으로 나눌 수 있다. PTQ는 학습이 완료된 모델을 quantization하는 기법이고, QAT는 학습을 진행하며 quantization을 진행하는 기법이다.

PTQ는 activation을 위한 clipping range를 정하는 시점에 따라서 static quantization과 dynamic quantization으로 나눌 수 있다. Static quantization은 추론 전에 weight와 activation을 quantization하는 방식이다. Dynamic quantization은 추론 전에 weight를 quantization하고 추론 중에 activation을 quantization하는 방식이다. 일반적으로 추론 속도는 static quantization이 dynamic quantization보다 빠르지만, 정확도 손실은 static quantization이 dynamic quantization보다 더 크다.

Ⅲ. 본 론

이 논문에서 사용한 structured pruning은 filter pruning [4]이다. Filter pruning 은 모델의 weight를 filter 단위로 제거하는 방식으로 그림 2와 같이 동작한다. 그림 2는 16개의 weight를 가지는 weight matrix를 filter pruning을 사용해 50% 만큼 제거했을 경우의 예시를 나타낸다.

Filter pruning에 GraNet을 적용하게 되면 그림 3과 같이 동작하게 된다. Filter pruning을 목표하는 비율만큼 먼저 진행하고, filter pruning을 추가로 진행한 뒤, 두 번째 pruning에서 제거된 filter의 개수만큼 제거된 filter를 재생성한다. 실험에서 사용한 filter 제거 기준은 filter에 있는 weight들의 L_1 Norm을 사용했다. 실험에서 사용한 재생성 기준은 filter에 있는 weight들의 gradient의 L_1 Norm이다.

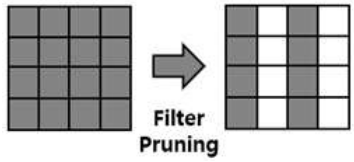

그림 2. Filter pruning 예시
Fig. 2. Example of filter pruning

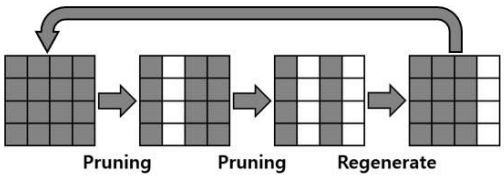

그림 3. GraNet filter pruning 예시
Fig. 3. Example of GraNet filter pruning

1. 실험 방식

첫 번째 실험 (III-2)은 GraNet을 사용한 pruning 한 모델들과 GraNet filter pruning을 사용해 pruning 한 모델들의 이미지 분류 문제에 대한 정확도와 연산량을 측정하는 방식으로 진행되었다. 연산량은 모델의 추론 과정에서 수행되는 곱셈 수를 통해 측정하였다.

두 번째 실험 (III-3)은 pruning 전 모델, pruning 적용 후 모델들과 각각의 모델들에 대해 quantization을 적용한 후 생성된 모델들의 이미지 분류 성능을 측정하는 방식으로 진행되었다.

두 실험에서 사용한 모델은 ResNet-18이고, 사용한 dataset은 CIFAR-100이다. Pruning에 사용한 pruning rate는 21% ~ 91% 이다.

실험에서 사용한 Quantization은 static quantization과 dynamic quantization을 사용하였다. Static quantization과 dynamic quantization 모두 32bit floating point로 표현된 모델을 8bit integer로 변환하는 방식을 사용하였다. 따라서 quantization을 통해 생성된 모델 가중치의 크기는 quantization 이전 모델의 1/4이다.

2. GraNet과 GraNet filter pruning

GraNet과 GraNet filter pruning의 성능을 비교한 실험 결과는 표 1과 같다. Pruning rate 21% ~ 91% 에 대해서 GraNet을 사용했을 때의 정확도가 GraNet filter pruning을 사용했을 때의 정확도보다 최대 7.69% 더 높음을 확인하였다. 표 1의 GraNet

표 1. CIFAR-100에서의 top-1 정확도와 연산량 (단위: MAC Operations)
Table 1. Top-1 accuracy and computational cost (unit: MAC Operations) in CIFAR-100

pruning rate	GraNet		GraNet filter pruning	
	정확도	연산량	정확도	연산량
0 %	74.84 %	742.64M	74.84 %	742.64M
21 %	73.23 %	742.64M	73.11 %	447.09M
37 %	73.39 %	742.64M	72.40 %	294.51M
52 %	73.26 %	742.64M	70.98 %	180.72M
65 %	73.13 %	742.64M	69.87 %	105.16M
76 %	72.88 %	742.64M	67.80 %	49.89M
85 %	72.72 %	742.64M	65.68 %	24.33M
91 %	72.01 %	742.64M	64.32 %	9.35M

의 경우, unstructured pruning 방식으로, 불규칙한 무작위 위치의 가중치에 대해서 pruning이 진행된다. 따라서 특별한 하드웨어의 지원 없이는 연산량을 줄이는 것이 불가능하다. 이에 따라 연산량은 pruning rate 변화에 상관없이 고정된다. 반면에 GraNet filter pruning의 경우, model thinning을 통해 연산량을 실질적으로 줄이는 것이 가능하다. 이에 따라, pruning rate가 최대 91%일 때, 연산량이 최대 98.74% 감소하는 것을 확인하였다.

3. GraNet filter pruning with quantization

GraNet filter pruning은 pruning rate 증가에 따라 정확도가 크게 감소한다. 이때 적절한 pruning rate와 함께 quantization을 사용한다면 정확도 감소를 최소화 하면서도 가중치의 크기를 줄일 수 있다. 실험 결과는 표 2와 같다. 실험에 사용한 모든 pruning rate에 대해서 no quantization을 기준으로 static quantization을 비교했을 때, 정확도 감소량의 최댓값은 0.72%이고, dynamic quantization과 비교했을 때의 정확도 감소량의 최댓값은 0.82%이다. 또한 static quantization과 dynamic quantization 모두 pruning rate가 커질수록 quantization으로 인한 정확도 감소량은 커지는 경향을 보인다. 그러나 8bit quantization을 적용한 경우, 가중치의 크기는 기존의 32bit 대비 1/4로 감소하므로 이를 고려하여 가중치 크기에 따른 정확도를 비교할 필요가 있다.

그림 4는 GraNet filter pruning과 quantization을 사용해 얻은 모델의 가중치 크기를 megabyte(MB) 단위로 환산 시 가중치의 상대적 크기에 따른 정확도를 나타낸 그림이다. 모델의 가

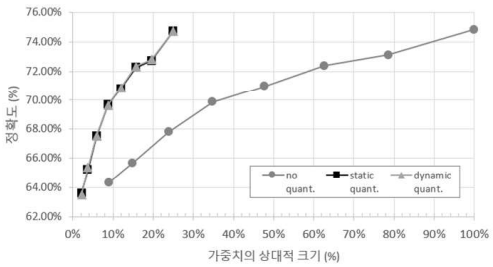

그림 4. CIFAR-100 에서의 양자화 방법과 가중치의 상대적 크기에 따른 top-1 정확도 비교
Fig. 4. Top-1 accuracy comparison by quantization method and relative weight size in CIFAR-100

표 2. CIFAR-100에서의 top-1 정확도
Table 2. Top-1 accuracy in CIFAR-100

pruning rate	no quantization	static quantization	dynamic quantization
0 %	74.84 %	74.75 %	74.74 %
21 %	73.11 %	72.72 %	72.81 %
37 %	72.40 %	72.26 %	72.35 %
52 %	70.98 %	70.81 %	70.87 %
65 %	69.87 %	69.66 %	69.64 %
76 %	67.80 %	67.51 %	67.54 %
85 %	65.68 %	65.21 %	65.34 %
91 %	64.32 %	63.60 %	63.50 %

중치의 크기가 같을 때, GraNet filter pruning만을 사용해 얻은 모델보다 GraNet filter pruning과 8-bit quantization을 동시에 사용해 얻은 모델이 정확도가 더 높은 것을 확인하였다. 이때, 표 2와 그림 4를 기반으로 정확도 하락 대비 압축 효율을 계산하기 위해서 pruning 및 quantization을 진행하지 않은 기존의 모델 정확도 (74.84%) 대비 각각 -2.5%, -5%, -10%의 정확도 하락 제약 조건을 두는 경우, 제거 가능한 최대 가중치 비율를 분석한 결과는 다음과 같다. 각 제한 사항을 만족하는 pruning rate는 각각 최대 37%, 52%, 85% 이며, 각각 8bit quantiztion을 통해 1/4의 가중치 크기가 추가로 줄어드므로, 제거 가능한 최대 가중치 비율은 각각 84.25%, 88.0%, 96.25%를 얻을 수 있었다.

Ⅳ. 결 론

Structured pruning은 모델의 크기를 줄임과 동시에 연산량을 감소시킬 수 있다는 장점이 있으나, pruning rate가 증가함에 따라 성능이 크게 감소하므로 모델의 크기를 줄이는 데 한계가 존재한다. 따라서 structured pruning과 quantization을 동시에 적용하여 단점을 완화할 수 있다. 본 논문에서는 먼저 pruning 기법 중 훌륭한 성능을 보이는 GraNet 기반의 structured pruning을 ResNet 모델에 적용하고, 동시에 8-bit quantization을 적용하여 모델의 정확도 손실을 최소화하면서 가중치 크기를 더욱 줄일 수 있음을 확인하였다.

References

[1] J. Frankle, and M. Carbin. “The lottery ticket hypothesis: Finding sparse, trainable neural networks”. In International Conference on Learning Representations, 2019.

[2] M. H. Zhu, and S. Gupta. “To prune, or not to prune: Exploring the efficacy of pruning for model compression”. arXiv preprint arXiv:1710.01878, 2018.

[3] S. Liu, T. Chen, X. Chen, Z. Atashgahi, L. Yin, H. Kou, L. Shen, M. Pechenizkiy, Z. Wang, and D. C. Mocanu. “Sparse training via boosting pruning plasticity with neuroregeneration”. Advances in Neural Information Processing Systems, 34, 2021a.

[4] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf. “Pruning filters for efficient convnets”. International Conference on Learning Representations, 2016.

[5] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer. 2021. “A Survey of Quantization Methods for Efficient Neural Network Inference”. arXiv preprint arXiv:2103.13630, 2021.