저계수행렬 근사 및 CP 분해 기법을 이용한 CNN 압축

문현철, 문기화, 김재곤
한국항공대학교
{hcmoon, rlghk1461}@kau.kr, jgkim@kau.ac.kr

Compression of CNN Using Low-Rank Approximation and CP Decomposition Methods

Hyeon-Cheol Moon, Gi-Hwa Moon, and Jae-Gon Kim
Korea Aerospace University

요 약

최근 CNN(Convolutional Neural Network)은 영상 분류, 객체 인식 등 다양한 비전 분야에서 우수한 성능을 보여주고 있으나, CNN 모델의 계산량 및 메모리가 매우 커짐에 따라 모바일 또는 IoT(Internet of Things) 장치와 같은 저전력 환경에 적용되기에는 제한이 따른다. 따라서, CNN 모델의 임무 성능을 유지하면서 네트워크 모델을 압축하는 기법들이 연구되고 있다. 본 논문에서는 행렬 분해 기술인 저계수행렬 근사(Low-rank approximation)와 CP(Canonical Polyadic) 분해 기법을 결합하여 CNN 모델을 압축하는 기법을 제안한다. 제안하는 기법은 계층의 유형에 상관없이 하나의 행렬분해 기법만을 적용하는 기존의 기법과 달리 압축 성능을 높이기 위하여 CNN 의 계층 타입에 따라 두 가지 분해 기법을 선택적으로 적용한다. 제안기법의 성능검증을 위하여 영상 분류 CNN 모델인 VGG-16, ResNet50, 그리고 MobileNetV2 모델 압축에 적용하였고, 모델의 계층 유형에 따라 두 가지의 분해 기법을 선택적으로 적용함으로써 저계수행렬 근사 기법만 적용한 경우 보다 1.5 ~ 12.1 배의 동일한 압축율에서 분류 성능이 향상됨을 확인하였다.

1. 서론

CNN(Convolutional Neural Network)을 기반으로 하는 인공신경망은 최근 컴퓨터 비전, 영상 인식 및 화질 개선 등 다양한 응용에서 뛰어난 성능을 보이고 있다. 그러나, 성능의 향상을 위한 계층의 깊이 및 학습할 가중치(weight) 수의 증가로 모델의 크기 및 추론 과정에서 접근해야 할 특징 맵(feature map)을 저장하기 위한 메모리가 크게 증가하게 되었다. 이에 따라 연산 속도나 메모리가 제한된 모바일 및 IoT 기기에 인공신경망을 추론하기에는 제한이 따른다. 따라서, 기존의 학습된 네트워크 모델의 성능을 최대한 유지하면서 모델의 크기를 줄이는 인공신경망 압축 기법들이 연구되고 있다[1].

CNN 모델을 압축하는 기법은 각 모델에 포함된 가중치의 수를 줄이는 데에 중점을 두고 있다. 대표적인 기법으로는 모델의 성능에 큰 영향을 주지 않은 가중치와 노드의 연결을 끊는 가지치기(pruning) 기법 [2]과 가중치 행렬을 2 개 이상의 행렬로 분해하여 가중치의 수를 줄이는 행렬 분해 기법 등이 있다[3].

본 논문에서는 행렬 분해 기술인 저계수행렬 근사(LR: Low-Rank approximation)와 CP(Canonical Polyadic) 분해 기법을 결합하여 CNN 모델을 압축하는 기법을 제안한다.

2. 행렬 분해 기법

행렬 분해 기법은 2 차원 이상의 CNN 각 계층의 가중치 행렬을 2 개 이상의 행렬로 분해함으로써 가중치의 수 및 연산량을 줄이는 방법이다. 대표적인 행렬 분해 기법은 2 차원 행렬을 SVD(Singular Value Decomposition) 기법을 이용하여 2 개의 행렬로 분해하는 저계수행렬 근사 기법과 3 차원 이상의 행렬을 다수의 계수(rank)-1 텐서(tensor)의 선형결합 형태로 분해하는 CP 분해 기법 등이 있다.

저계수행렬 근사 기법은 2 차원 행렬을 SVD 분해 기반으로 2 개의 2 차원 행렬로 분해하는 것으로 식 (1)과 같이 표현된다.

$$W_i = U_i V_i^T \tag{1}$$

여기서 W_i, U_i, V_i^T는 각각 $M \times N$, $M \times R$, $R \times N$ 크기를 가지며, W_i 는 각 CNN 모델의 i번째 층의 가중치 행렬을 의미하며, U_i, V_i^T 는 i 번째 층의 가중치 행렬로부터 분해되는 행렬을 의미한다.

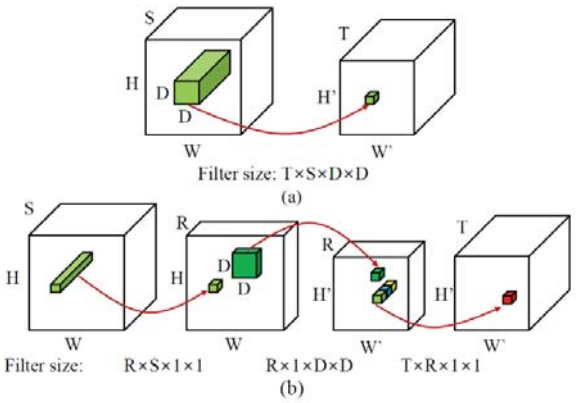
CP 는 3 차원 이상의 고차원 텐서들을 rank-1 텐서들의 선형결합으로 분해한다. 예를 들어, 식 (2)와 같이 3 차원 텐서 X를 R 개로 이루어진 3 개의 rank-1 의 텐서로 분해할 수 있다. 여기서 R 은 계수 값이고, 해당 기법을 적용하기 위해 입력되는 하이퍼 파라미터(hyper parameter)이다. 따라서, R 값에 따라 압축율이 상대적으로 결정된다.

$$\mathrm{X} \equiv \sum_{r=1}^{R} a_R \otimes b_R \otimes c_R \tag{2}$$

본 논문에서는 4 차원 기반의 컨볼루션(convolution) 가중치 텐서들을 CP 분해 기법을 적용하여 압축한다[4]. 2D 기반의 컨볼루션 층의 가중치 값들은 4D 형태(2D 필터 크기, 입력채널 수, 출력채널 수)로 구성되어 있다. 따라서, 컨볼루션 가중치 텐서 K에 대한 CP 분해는 다음 식 (3)과 같으며, 본 논문에서는 3 개의 계층으로 분해하는 기법을 제시한다.

$$K_{t,s,j,i} \equiv \sum_{r=1}^{R} U_{r,s}^{(1)} U_{r,j,i}^{(2)} U_{t,r}^{(3)} \tag{3}$$

여기서 $U_{r,s}^{(1)}$, $U_{r,j,i}^{(2)}$, $U_{t,r}^{(3)}$ 는 각각 $R \times S$, $R \times D \times D$, $R \times T$ 크기를 가진다. 이 때, S 와 T 는 각각 입력과 출력의 채널 수, $D \times D$ 는 컨볼루션 필터 크기를 의미한다. 따라서, $U_{r,s}^{(1)}$, $U_{t,r}^{(3)}$ 은 필터 크기가 1 인 Pointwise(PW) 컨볼루션 층의 형태로 구성되며, $U_{r,j,i}^{(2)}$ 는 필터 크기가 $D \times D$ 인 Depthwise(DW) 컨볼루션 층으로 구성된다. 그림 1 은 식 (3)을 그림으로 나타낸 것이다. 따라서, 결국 하나의 컨볼루션 층을 2 개의 Pointwise, 1 개의 Depthwise 컨볼루션 층으로 분해하게 된다.



(a) 컨볼루션 층 (b) CP 분해된 컨볼루션 층
그림 1. CP 분해 기법[4]

3. 실험결과

표 1 은 압축을 하지 않은 원본 모델의 성능으로 Top-5 Accuracy 는 모델로부터 예측된 클래스 중 상위 5 개에 정답이 있을 경우의 정확도를 의미한다. 성능을 측정하기 위해 사용된 데이터셋은 ILSVRC(ImageNet Large Scale Visual Recognition Competition) 2012 의 검증(validation) 데이터셋 50,000 개에 영상에 대한 분류 성능이다[5]. 표 2 는 본 논문에서의 각 실험 별 조건을 나타낸 것이다. 기존의 기법들은 계층의 유형에 상관없이 단일 분해 기법들을 적용하였다. 따라서, 기존의 기법 대비 성능 향상을 위해 각 실험에서는 계층의 유형에 따라 적용하는 행렬 분해 기법들을 다르게 적용하였다. 완전 연결(FC: Fully-Connected) 층의 경우 모든 실험조건에서 저계수행렬 근사(LR) 기법을 적용하였고, 컨볼루션 층에는 2 개의 기법을 모두 선택적으로 적용하였다. 또한, 해당 실험에서 단순히 행렬 분해 기법만을 적용하면 성능의 손실이 발생하는 관계로 별도의 재학습(Re-training)을 적용하였으며, 재학습을 위해 ILSVRC 2012 의 학습 데이터 셋을 사용하였다.

Table 1. 원본 모델 성능 (ILSVRC 2012)

Model	Model Size (MB)	Top-5 Accuracy (%)
VGG-16	527.0	90.05
ResNet50	98.2	91.93
MobileNetV2	13.8	90.06

Table 2. 실험 조건

	Layer Type		
	2D Conv.	PW & DW Conv.	FC
Baseline (LR)	LR	LR	LR
Test 1	CP	CP	
Test 2	CP	LR	

표 3 은 행렬 분해 기법을 적용한 경우의 압축 성능 실험결과이다. 각 실험별 압축율은 Baseline 으로 가정한 LR 기법을 적용한 압축된 모델의 Top-5 분류 성능이 원본 모델 대비 3% 미만의 손실일 때를 기준으로 하였다. 행렬 분해 기법이 임무 성능 손실 없이 주어진 모델을 1.5~12.1 배 압축함을 확인하였으며, 특히 VGG-16 에서는 기존의 가지치기 기법[1] 대비 약 1.3 배 더 압축이 됨을 보여준다. 또한, 저계수행렬 근사 기법만을 적용한 Baseline 기법보다 계층의 유형에 따라 LR 과 CP 기법을 선택적으로 적용한 제안 기법이 VGG-16, ResNet50 에서 영상 분류 성능 향상이 있음을 확인하였다.

그림 2 는 ResNet50 에서의 각 실험 별 성능 압축율에 따른 성능를 나타낸 것이다. 컨볼루션 층에서 오직 CP 를 적용한 Test 1 이 Baseline 기법보다 전체 압축율 구간에서 분류 성능이 감소되었지만, 컨볼루션 층의 유형에 따라 선택적으로 분해 기법을 적용한 Test 2 에서는 낮은 압축율 범위(0.35 ~ 0.45) 에서는 기존 Baseline 기법보다 분류 성능의 향상이 있음을 확인하였다.

Table 3. 실험결과

Model	Test	Top-5 Accuracy (%)	Compression ratio (x)
VGG-16	**Baseline**	87.92	12.1
	Test 1	88.11	
	Hans [1]	88.05	9.0
ResNet50	**Baseline**	89.44	2.86
	Test 1	88.79	
	Test 2	89.87	
MobileNetV2	**Baseline**	88.20	1.54
	Test 1	88.07	

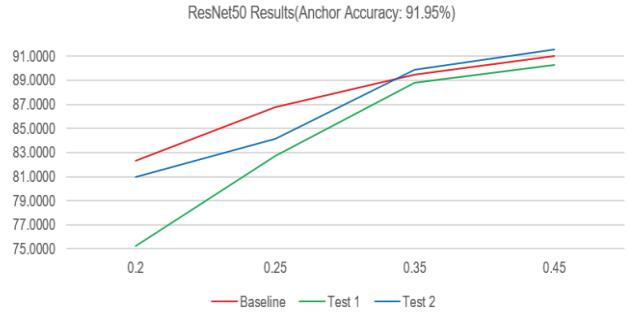


그림 2. ResNet50 에서의 압축률에 따른 분류 성능

4. 결론

본 논문에서는 행렬 분해 기술인 저계수행렬 근사(Low-rank approximation)와 CP(Canonical Polyadic) 분해 기법을 이용하여 학습된 CNN 모델을 압축하는 기법을 제안하였다. 영상 분류의 대표적 CNN 모델인 VGG-16, ResNet50, 그리고 MobileNetV2 에 대하여 2 가지의 행렬 분해 기술을 선택적으로 적용한 제안 기법의 압축 성능을 검증하였고, 실험결과 LR 와 CP 분해 기법을 컨볼루션 계층의 유형에 따라 선택적으로 적용하는 제안 기법이 단일 행렬 분해 기법을 적용한 경우보다 압축 성능 향상이 있음을 확인하였다.

Acknowledgement

이 논문은 2020 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2020R1F1A1068106).

참 고 문 헌

[1] S., Han, et al, “Deep Compression: Compressing Deep Neural Networks with pruning, trained quantization and Huffman coding,” Computer Vision and Patter Recognition, In Proc. ICLR 2016, May 2016.

[2] C. Aytekin, F. Cricri, T. Wang, E. Aksu, “Response to the Call for Proposals on Neural Network Compression: Training Highly Compressible Neural Networks,” ISO/IEC JTC1/SC29/WG11, m47379, Mar. 2019.

[3] M. Jaderberg, A. Vedaldi, and A. Zisserman, “Speeding up Convolutional Neural Networks with Low Rank Expansions,” In Proc. CVPR, Jun. 2014.

[4] V. Lebedev, Y. Ganin, M. Rakhuba, I. Oseledets, and V. Lemptisky, “Sppeding-up Convolutional Neural Networks Using Fine-tuned CP-Decomposition,” In Proc. CVPR, Jun. 2015.

[5] [Available at Online] http://www.imagenet.org/challenges/LSVRC/2012/